ARTICLES
https://doi.org/10.1038/s41588-022-01049-4

nature genetics

Genome-wide analysis of *cis*-regulatory changes underlying metabolic adaptation of cavefish

Jaya Krishnan[1], Chris W. Seidel[1], Ning Zhang[1], Narendra Pratap Singh[1], Jake VanCampen[1,3], Robert Peuß[1,4], Shaolei Xiong[1], Alexander Kenzior[1], Hua Li[1], Joan W. Conaway[1] and Nicolas Rohner[1,2] ✉

***Cis*-regulatory changes are key drivers of adaptative evolution. However, their contribution to the metabolic adaptation of organisms is not well understood. Here, we used a unique vertebrate model, *Astyanax mexicanus*—different morphotypes of which survive in nutrient-rich surface and nutrient-deprived cave waters—to uncover gene regulatory networks underlying metabolic adaptation. We performed genome-wide epigenetic profiling in the liver tissues of *Astyanax* and found that many of the identified *cis*-regulatory elements (CREs) have genetically diverged and have differential chromatin features between surface and cave morphotypes, while retaining remarkably similar regulatory signatures between independently derived cave populations. One such CRE in the *hpdb* gene harbors a genomic deletion in cavefish that abolishes IRF2 repressor binding and derepresses enhancer activity in reporter assays. Selection of this mutation in multiple independent cave populations supports its importance in cave adaptation, and provides novel molecular insights into the evolutionary trade-off between loss of pigmentation and adaptation to food-deprived caves.**

CREs are major targets of evolution for shaping phenotypic diversity[1–3] and helping organisms adapt to various environmental niches[4,5]. The role of *cis*-regulatory changes in the evolution of metabolic adaptations is not well understood. The Mexican tetra, *Astyanax mexicanus*, with its two morphotypes—the river-dwelling surface fish and the cave-dwelling cavefish—provides an exceptional model system to study evolutionary changes at the genetic and genomic levels[6–8]. The surface fish live in nutrient-rich rivers, whereas the cavefish are well adapted to survive in dark and nutrient-deprived caves, which they colonized around 150,000 years ago[9,10]. Importantly, many of the cave populations have independently adapted to different cave environments, allowing researchers to study whether the same genes and regulatory networks were used in adaptation or whether evolution took different paths to arrive at similar phenotypes[11].

Missense mutations have been identified in key metabolic genes such as *insra* and *mc4r* that help cavefish to survive in low-nutrient conditions[12,13]. However, the contribution of regulatory changes in the complex metabolic phenotypes of cavefish has not been investigated. Here, we generated a high-resolution, genome-wide map of candidate CREs in *Astyanax mexicanus*. We focused our study on the liver because of its central role in glucose and fat metabolism[14]. Analysis of the regulatory profiles of two independent cave populations, Pachón and Tinaja, revealed remarkable similarity, indicating signs of repeated evolution at an epigenetic level. Furthermore, we show that several CREs display different epigenetic states between surface fish and cavefish, which correlate with their ability to modulate varying levels of reporter gene expression. Analysis of one of these CREs in *hpdb* (4-hydroxyphenylpyruvate dioxygenase b), a gene involved in the tyrosine metabolism pathway, showed that the deletion of a binding site for the interferon regulatory factor 2 (IRF2) repressor protein in cavefish is sufficient to drive increased gene expression. This suggests that the mutation may have an adaptive role, with a trade-off between the conversion of tyrosine to melanin and tricarboxylic acid (TCA) cycle intermediates. Our study not only demonstrates the value of using independently evolved populations to identify noncoding genomic loci relevant to cave adaptation, but also provides many candidates for future studies on metabolic adaptation.

Results

Genome-wide annotation of CREs. We analyzed accessible chromatin, histone modifications and gene expression to generate a genome-wide epigenetic profile from the liver tissue of surface fish and two independently derived cave morphotypes, Pachón and Tinaja (Fig. 1a and Supplementary Fig. 1a)[9,15–18]. Using the surface fish genome as a reference[19], we identified a total of 94,175 accessible chromatin regions or putative CREs across the three populations; there were 68,002 in surface, 69,178 in Pachón and 73,762 in Tinaja (Supplementary Data 1). These CREs were distributed across the genomes, with 44.87% residing within 10 kilobases (kb) of a transcription start site (TSS) (Supplementary Fig. 1b).

We performed ChIP-seq (chromatin immunoprecipitation sequencing) for histone marks, and further characterized the candidate CREs using ChromHMM, an automated computational pipeline for learning chromatin states based on a hidden Markov model[20] (Supplementary Fig. 1c,d). Genomic regions marked with H3K4me3 (trimethylation of histone H3 at lysine 4) were classified as active promoters, and those marked with H3K27ac (acetylation of histone H3 at lysine 27) were classified as active enhancers (Supplementary Fig. 1c)[21]. In addition, regions with H3K27me3 (trimethylation of histone H3 at lysine 27) were classified as repressed, and those marked with all three epigenetic modifications were classified as poised[21]. Regions with low intensity of all histone

[1]Stowers Institute for Medical Research, Kansas City, MO, USA. [2]Department of Molecular and Integrative Physiology, University of Kansas Medical Center, Kansas City, KS, USA. [3]Present address: Department of Medicine, Knight Cardiovascular Institute, Oregon Health & Science University, Portland, OR, USA. [4]Present address: Institute for Evolution and Biodiversity, University of Münster, Münster, Germany. ✉e-mail: nro@stowers.org

marks were classified into low signal states for further analysis (Supplementary Fig. 1c). We observed that CREs marked as active and poised had accessible chromatin, whereas those marked as repressed and the low signal states had relatively inaccessible chromatin (Supplementary Fig. 1e). We performed bulk RNA-seq (RNA sequencing) on liver tissues, and observed that the expression of genes was positively correlated with active and poised marks and inversely correlated with repressed marks (Supplementary Fig. 1e,f and Supplementary Data 2). Together, these analyses demonstrate that chromatin accessibility can serve as a means of identifying putative CREs that correlate with the expression of the associated genes in this system.

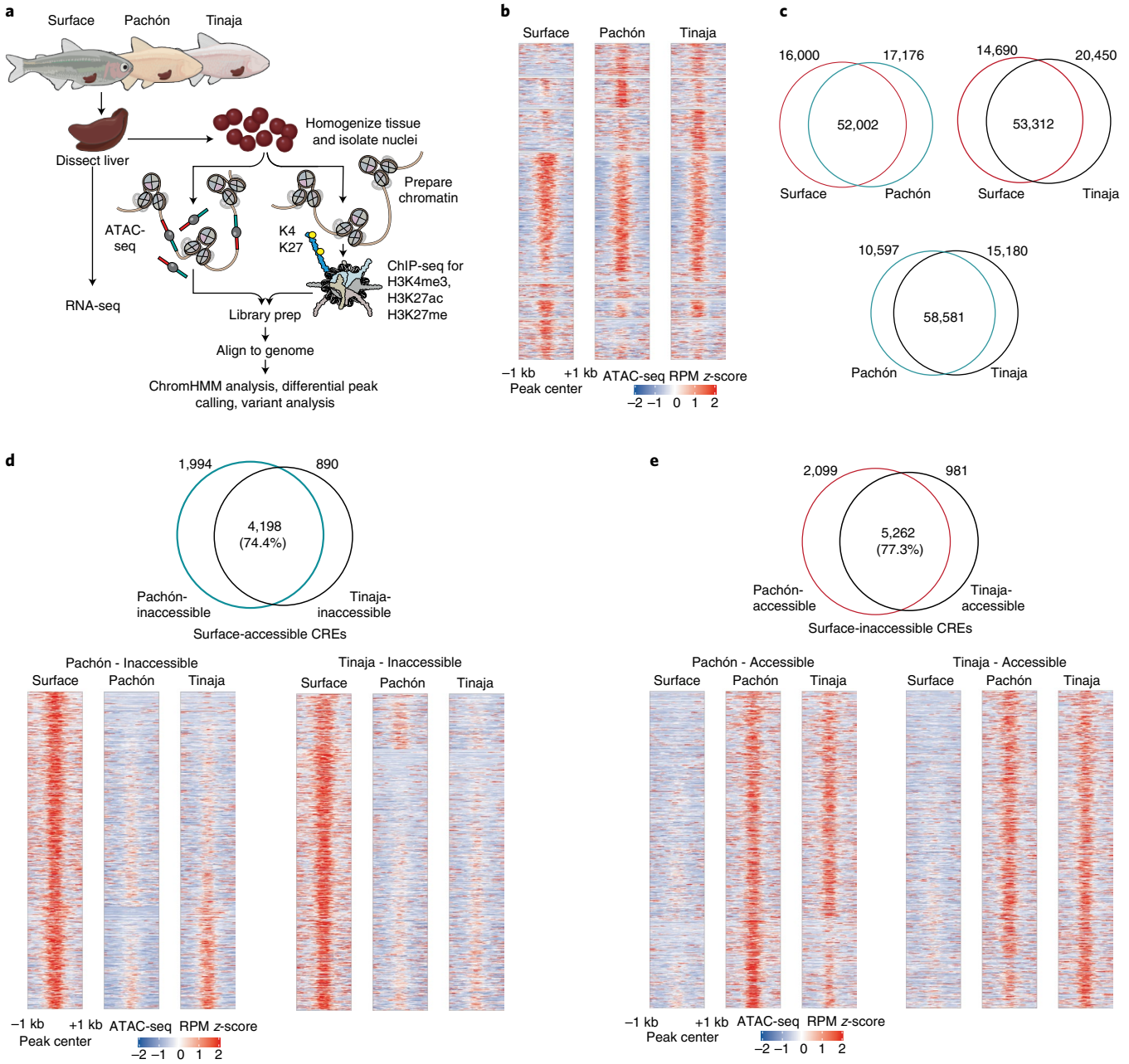
We also analyzed these open chromatin regions for sequence conservation across 11 fish species[22]. Regions marked by a chromatin feature were approximately 30% more conserved than randomly picked regions in the genome (Supplementary Fig. 2a; see example CRE in Supplementary Fig. 2b). We extended sequence conservation analyses to known regulatory regions in the human liver. We found that 94 out of 441 human liver-specific enhancers had sequence conservation and were marked as open chromatin in our dataset, indicating that some of the putative CREs that we identified are conserved in vertebrates and could be involved in gene regulatory networks that control conserved metabolic processes (Supplementary Table 1)[23].

Morphotype-biased CREs associate with key metabolic pathways. To understand how *cis*-regulatory networks have evolved during cave adaptation, we classified regions into similarly or differentially accessible among the three populations (Fig. 1b). As expected, many loci (47,241) were invariant, but 33,176 loci were differentially accessible between surface and Pachón and 35,140 loci were differentially accessible between surface and Tinaja (Fig. 1c). Interestingly, we identified only 25,777 differentially accessible regions between the two cave populations, which suggests a greater divergence between the surface and cave populations than between the cave populations (Fig. 1c). The global patterns for the active histone modifications also largely correlated with that of accessible chromatin (Supplementary Fig. 2c). We further analyzed regions that were highly differentially accessible and that were mapped near genes (<10 kb), and observed that 74.4% of the regions that were accessible in surface fish (surface-accessible CREs) lost accessibility in both of the cave populations (Fig. 1d,f). Similarly, 77.4% of the regions that gained accessibility in Pachón compared with surface also gained accessibility in Tinaja (cave-accessible CREs) (Fig. 1e,g). As shown in the heatmaps (Fig. 1d,e), CREs that were differentially accessible between surface and one of the cave morphotypes were also differentially accessible between surface and the other cave morphotype. These comparisons indicate that both of the cave populations have gained or lost accessible chromatin states with regulatory potential in a similar set of genomic regions during evolution. A similar trend was seen in the genome-wide pattern of histone modifications and gene expression between the two cavefish populations compared with surface fish (Supplementary Fig. 2d,e). This high degree of similarity in the chromatin profiles of the two cave populations points toward modifications of existing regulatory circuitry in surface fish and suggests that the system is robustly wired to use similar sets of gene regulatory mechanisms in independently derived populations, indicating repeated or parallel evolution on an epigenetic level. To further understand the repeated cave-adaptive *cis*-regulatory features, we focused on regions that were similarly biased in both cave populations (cave-accessible CREs (c-CREs)), and compared them with surface-accessible CREs (s-CREs).

Although changes in CREs do not show strict correlation with the expression changes of nearby genes, there is evidence that CREs tend to control nearby genes[24–29]. In addition, studies have shown that evolutionarily biased enhancers associated with transcriptional changes are often linked to phenotypic variations[3]. Therefore, to focus on stronger CRE–gene correlation, we analyzed CREs within 10 kb of the TSS of genes[17,27,28] (Fig. 2a). We observed a significant correlation ($P < 0.05$; see the Methods for details) between morphotype-biased CREs and the expression of nearby genes (Fig. 2a,b). We observed that genes associated with s-CREs were enriched in pathways that are involved in circadian rhythm, lipid metabolism and transforming growth factor β (TGF-β) signaling (Fig. 2c), and genes associated with c-CREs were enriched in pathways that are involved in lipid metabolism and immune function (Fig. 2d). Notably, lipid metabolism pathway genes that were enriched near s-CREs comprised catabolic genes (lipases and fatty acid binding proteins) (Fig. 2e) that are upregulated in surface fish, whereas lipid metabolism pathway genes that were near c-CREs comprised lipid signaling and anabolic genes (fatty acid synthase and acyl-CoA synthetases) that are upregulated in cavefish (Fig. 2f). These findings are in line with those of previous studies that show increased fat accumulation in cavefish[12,30]. Network analysis of the genes associated with c-CREs showed upregulation of lipid synthesis pathways; specifically, genes such as *fasn* (a fatty acid synthase gene), *g6pd* (encodes glucose-6-phosphate dehydrogenase) and *slc30a8* (encodes a zinc transporter involved in insulin function) were upregulated in cavefish (Supplementary Fig. 3a)[31–33].

Genetic changes in CREs cause differential functional output. The differential accessibility of the s-CREs and c-CREs could be due to either differences in *trans*-acting factors or differences in the sequence of the putative CRE itself. To understand the role of specific transcription factors (TFs) in differential accessibility, we analyzed the enriched binding motifs of TFs in s-CREs and c-CREs. In s-CREs, we observed enrichment for motifs of the nuclear receptors retinoic acid receptor (RAR), liver X receptor (LXR), and hepatocyte nuclear factor 4A (HNF4A), which are known to regulate glucose and lipid metabolism (Fig. 2g)[34–36], and confirmed the results from the pathway analysis (Fig. 2c). Similar analysis for c-CREs revealed enrichment of binding sites for nuclear transcription factor Y (NF-Y) and Krüppel-like factor 14 (KLF14) (Fig. 2h). NF-Y regulates lipid metabolism via the leptin pathway[37], and KLF14 represses TGF-β signaling[38], a pathway that is enriched in surface fish but not in cavefish (Fig. 2c,d). Notably, we found that the consensus motif for CCCTC-binding factor (CTCF), a regulatory element-binding and genome-organizing

Fig. 1 | Analysis of morphotype-biased accessible chromatin regions. a, Schematic of the experimental design. **b**, Heatmap comparing ATAC-seq signals between morphotypes. Although most loci have similar accessibility, others have varying degrees of accessibility across populations. RPM, reads per million. **c**, Two-way comparisons of accessible chromatin regions represented as Venn diagrams. **d**, Venn diagram showing overlap of ATAC-seq peaks between Pachón and Tinaja for surface-accessible CREs. Lower panels are heatmaps showing that most of the regions that are accessible in surface but inaccessible in Pachón are also inaccessible in Tinaja. **e**, Venn diagram showing overlap of ATAC-seq peaks between Pachón and Tinaja for surface-inaccessible CREs. Lower panels are heatmaps showing that most of the regions that are inaccessible in surface but accessible in Pachón are also accessible in Tinaja. **f**, UCSC Genome Browser screenshot showing the epigenetic profile for a surface-accessible CRE. The *y* axis denotes reads per million. **g**, UCSC Genome Browser screenshot showing the epigenetic profile for a cave-accessible CRE. The *y* axis denotes reads per million.


a
Surface
Pachón
Tinaja
Dissect liver
Homogenize tissue and isolate nuclei
Prepare chromatin
ATAC-seq
K4 K27
ChIP-seq for H3K4me3, H3K27ac H3K27me
RNA-seq
Library prep
Align to genome
ChromHMM analysis, differential peak calling, variant analysis
b
Surface
Pachón
Tinaja
−1 kb Peak center
+1 kb
ATAC-seq RPM z-score
−2 −1 0 1 2
c
16,000
17,176
52,002
Surface
Pachón
14,690
20,450
53,312
Surface
Tinaja
10,597
15,180
58,581
Pachón
Tinaja
d
1,994
890
4,198 (74.4%)
Pachón-inaccessible
Tinaja-inaccessible
Surface-accessible CREs
Pachón - Inaccessible
Tinaja - Inaccessible
Surface
Pachón
Tinaja
−1 kb Peak center
+1 kb
ATAC-seq RPM z-score
−2 −1 0 1 2
e
2,099
981
5,262 (77.3%)
Pachón-accessible
Tinaja-accessible
Surface-inaccessible CREs
Pachón - Accessible
Tinaja - Accessible
Surface
Pachón
Tinaja
−1 kb Peak center
+1 kb
ATAC-seq RPM z-score
−2 −1 0 1 2
f
19:157,983-165,578
2 kb
astMex_2.0
19:
159,000
160,000
161,000
162,000
163,000
164,000
165,000
Nostrin
Surface_ATAC
Pachón_ATAC
Tinaja_ATAC
Surface_H3K4me3
Pachón_H3K4me3
Tinaja_H3K4me3
Surface_H3K27ac
Pachón_H3K27ac
Tinaja_H3K27ac
g
17:20,801,033-20,807,714
2 kb
astMex_2.0
17:
20,802,000
20,803,000
20,804,000
20,805,000
20,806,000
20,807,000
iws1
Surface_ATAC
Pachón_ATAC
Tinaja_ATAC
Surface_H3K4me3
Pachón_H3K4me3
Tinaja_H3K4me3
Surface_H3K27ac
Pachón_H3K27ac
Tinaja_H3K27ac

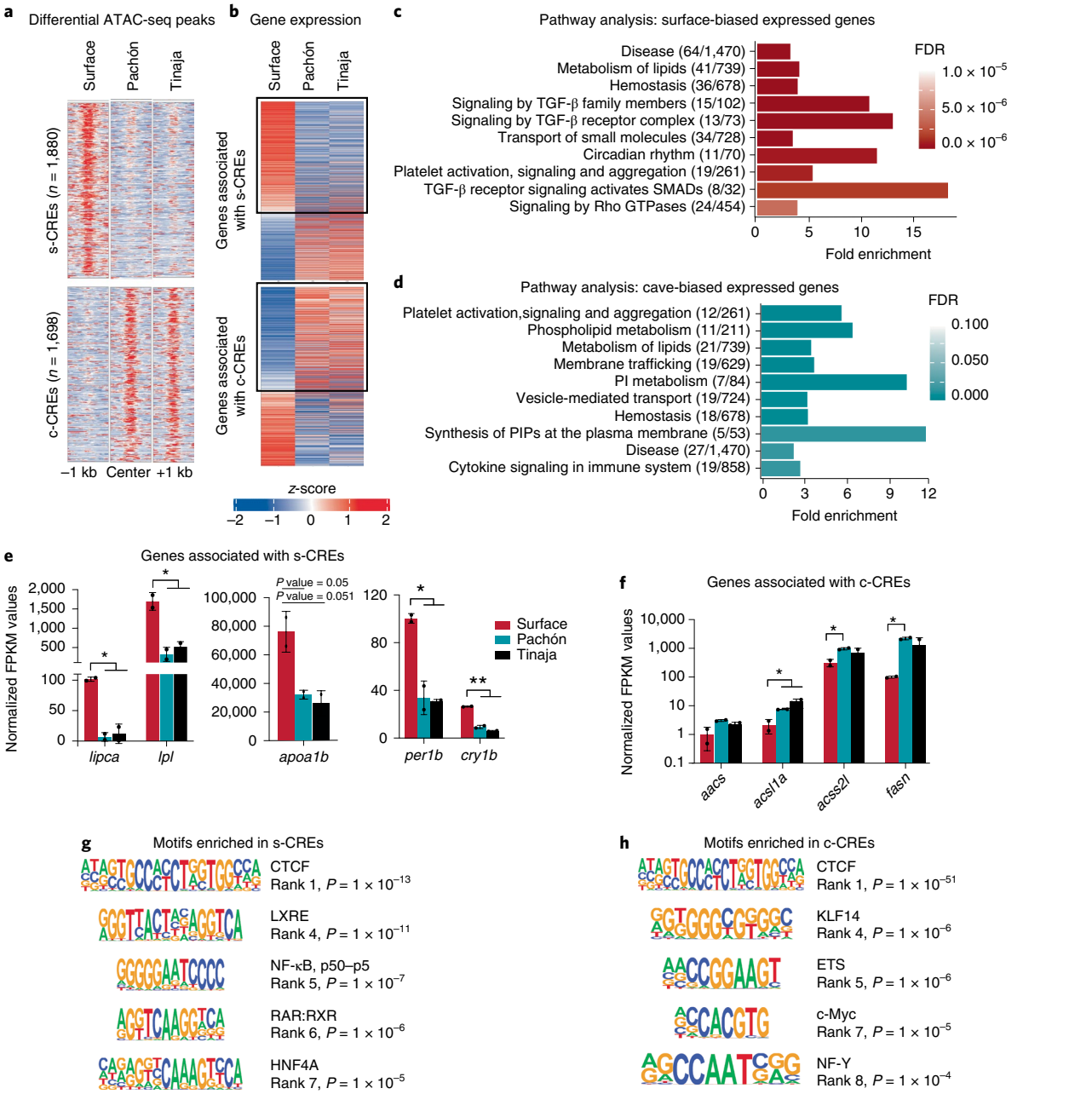


Fig. 2 | Morphotype-biased accessible chromatin regions associate with key metabolic pathway genes. a, Heatmaps depicting *z*-scores of the ATAC-seq signal at the s-CREs and at the c-CREs. *n*, number of peaks. **b**, Heatmaps depicting *z*-scores for the expression of genes associated with s-CREs and genes associated with c-CREs. **c**,**d**, Pathways or reactomes (using GSEA) enriched in genes in proximity (within 10 kb of the TSS) of s-CREs (**c**) and c-CREs (**d**) and showing the same bias in expression. The values in parentheses represent the number of genes in our list / total number of genes in that reactome. FDR, false discovery rate; PI, phosphatidylinositol; PIPs, phosphatidylinositol phosphates. **e**, Expression levels of genes associated with s-CREs belonging to the lipid metabolism (catabolism) and circadian rhythm reactomes. The graph shows mean values ± s.d. from $n = 2$ biologically independent RNA-seq experiments. *P* values are derived from edgeR and are adjusted for multiple hypothesis testing. $^*P < 0.05$; $^{**}P < 0.005$ using two-tailed Student's *t*-test. FPKM, fragments per kilobase of transcript per million mapped reads; *lipca*, hepatic triacylglycerol lipase a; *lpl*, lipoprotein lipase; *apoa1b*, apolipoprotein 1a; *per1b*, period 1b; *cry1b*, cryptochrome circadian regulator 1b. **f**, Expression levels of genes associated with c-CREs belonging to the lipid metabolism (anabolism) reactome. The graph shows mean values ± s.d. from $n = 2$ biologically independent RNA-seq experiments. *P* values are derived from edgeR and are adjusted for multiple hypothesis testing. $^*P < 0.05$ using two-tailed Student's *t*-test. *aacs*, acetoacetyl-CoA synthetase; *acsl1a*, long-chain-fatty-acid-CoA ligase 1; *acss2l*, acyl-CoA synthetase short chain family member 2; *fasn*, fatty acid synthase. **g**, Motifs enriched in s-CREs. CTCF, CCCTC-binding factor; LXRE, liver X receptor; NF-κB, nuclear factor kappa-light-chain-enhancer of activated B cells p50–p52 subunits; RAR:RXR, retinoic acid receptor:retinoic X receptor; HNF4A, hepatocyte nuclear factor 4A. **h**, Motifs enriched in c-CREs. KLF14, Krüppel-like factor 14; ETS, E26 transformation-specific; c-Myc, cellular myelocytomatosis oncogene; NF-Y, nuclear transcription factor Y.

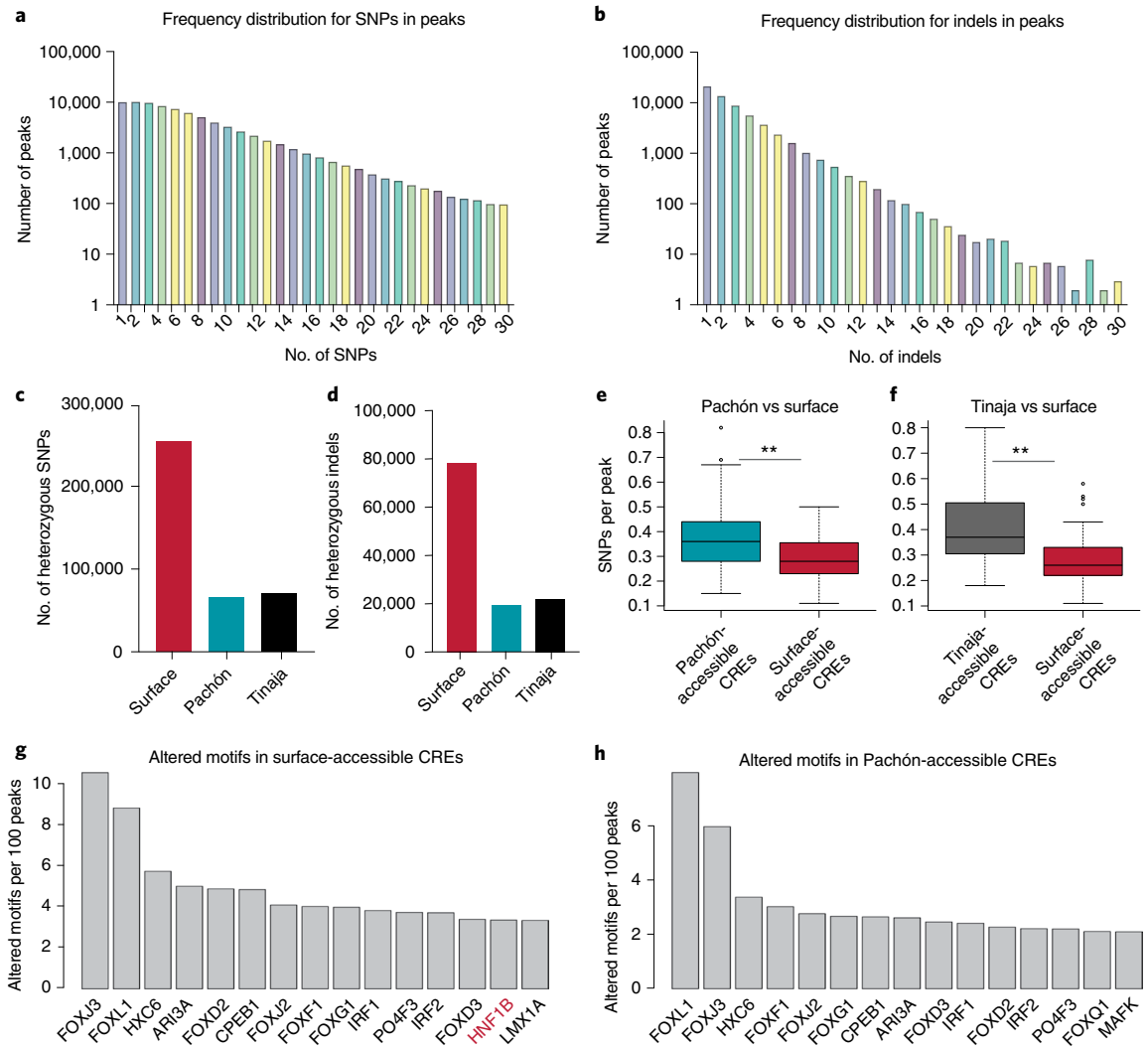


Fig. 3 | Analysis of genetic changes underlying accessible chromatin regions. a, Bar graph showing frequency distribution of SNPs under ATAC-seq peaks. **b**, Bar graph showing frequency distribution of indels under ATAC-seq peaks. **c**, Bar graph showing number of heterozygous SNPs in the different morphotypes. **d**, Bar graph showing number of heterozygous indels in the different morphotypes. **e**, Boxplot comparing SNP density per peak in Pachón-accessible and surface-accessible CREs. $^{**}P = 3.52 \times 10^{-4}$ using Mann–Whitney *U*-test. **f**, Boxplot comparing SNP density per peak in Tinaja-accessible and surface-accessible CREs. $^{**}P = 4.29 \times 10^{-10}$ using Mann–Whitney *U*-test. The center line in the boxplots indicates the median, box edges indicate the 25th and 75th percentiles, and whiskers indicate 1.5 × the interquartile distance, or the minima or maxima, whichever value is closest to the median. **g**,**h**, Bar graph showing number of altered TF motifs in surface-accessible CREs (**g**) and Pachón-accessible CREs (**h**). There are more altered motifs for HNF1B in surface-accessible CREs than in Pachón-accessible CREs.

factor, was enriched in both s-CREs and c-CREs, which supports their putative function as regulatory regions of the genome, as well as the role of CTCF in the three-dimensional organization of the genome during evolution[39,40]. Together, these analyses highlight key liver TFs and pathways that likely influence metabolism via the identified CREs.

The binding of TFs can directly impact the epigenetic status and function of CREs[18]. We reasoned that the differential accessibility of the s-CREs and c-CREs could be due to mutations and/or differential expression of the TFs that recognize the cognate motifs enriched in the analysis. Notably, a recent study identified a mutation in *hnf4a* (ref. [19]) that may lead to differences in its downstream targets between surface fish and cavefish. Indeed, network analysis showed that many direct and indirect targets of Hnf4a were differentially expressed between surface and Pachón (Supplementary Fig. 3b,c), which could be a consequence of changes to the CREs of these genes. Next, we queried the liver transcriptome data for the TFs with enriched motifs (Fig. 2g,h) and found little or no significant difference in expression between morphotypes (Supplementary Fig. 4a). Lastly, mutations in CREs themselves could result in changes in downstream gene expression[41].

To explore the causal role of polymorphisms or mutations in CREs in differential gene regulation, we did a comprehensive analysis of single nucleotide polymorphisms (SNPs) and insertion–deletion polymorphisms (indels) within all putative CREs. Using the Genome Analysis Toolkit (GATK) variant calling tool, we identified a total of 527,644 SNPs and 183,958 indels between morphotypes using raw reads of the ATAC-seq (assay for transposable-accessible chromatin using sequencing) datasets. First, we analyzed our SNP data and found that 27.4% of the peaks had no SNPs, whereas the rest of the peaks had anywhere between 1 and 169 SNPs per peak (Fig. 3a shows the distribution for 1–30 SNPs per peak). Similarly, 27.1% of peaks had no indels, whereas the rest of the peaks had between 1 and 44 indels (Fig. 3b shows the distribution for 1–30

indels per peak). There were 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 cave populations (Supplementary Fig. 4b). We observed a similar trend for indels (Supplementary Fig. 4c). The extent of heterozygosity was greater in surface for both SNPs and indels, which is in line with earlier observations of greater genetic diversity in surface fish and higher inbreeding in cave populations[42] (Fig. 3c,d). To assess whether genetic changes in CREs underlie differential chromatin accessibility, we compared the SNP frequency in surface-accessible peaks to that in Pachón-accessible and Tinaja-accessible peaks (differential peaks with $P < 0.001$). We observed a small but statistically significant increase in the SNP frequency in both Pachón-accessible and Tinaja-accessible peaks compared with that of 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 the R package motifbreakR and predicted altered motifs, which are TF motifs overlapping with SNP(s) that could potentially alter TF binding (Fig. 3g)[43]. In our entire dataset, we observed that 48.2% (254,554 SNPs) of the identified SNPs had the potential to alter TF motifs, resulting 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, whereas 41% of surface-accessible CREs contained at least one altered motif. Interestingly, hepatocyte nuclear factor 1B (HNF1B), which regulates glucose metabolism and is implicated in diabetes[44], was one of the top 15 TFs with altered motifs in surface-accessible CREs (Fig. 3g), but not in Pachón-accessible CREs (Fig. 3h). This analysis highlights the abundance of altered TF motifs occurring within CREs that could potentially alter CRE activity and could thereby have significant effects on downstream gene expression patterns between surface and Pachón.

To functionally validate our prediction on mutated CREs affecting gene expression, we used a luciferase reporter assay and examined activity of a selected set of surface-biased and cave-biased CREs. To select which CREs to test in reporter assays, we considered only differentially accessible CREs that carried at least one polymorphism (SNP or indel) between surface and either of the cave populations and were within 10 kb of a TSS. Next, to enable downstream characterization of the CRE and its associated gene, we selected CREs with associated genes that were annotated with gene names on Ensembl (BioMart)[45]. Lastly, to ensure better CRE–gene association, we selected CREs with associated gene expression that was biased in the same direction as the chromatin accessibility between morphotypes (Fig. 4a). For these 466 candidate CREs, we manually evaluated additional parameters such as the degree of difference in the expression of associated gene, maintenance of differential chromatin features in the flanking genomic regions, and relevance of the associated gene in metabolism (see the Methods for details). This resulted in a total of 25 differentially accessible CREs for functional validation (Fig. 4a and Supplementary Table 2).

To functionally test the CREs in vivo, we generated transient transgenic *Astyanax* (Supplementary Fig. 5a,b) and zebrafish (Supplementary Fig. 5a,c) embryos for several CREs by Tol2-mediated transgenesis in fertilized eggs as previously described[46,47]. The embryos were injected with the surface or cave constructs of CREs, and green fluorescent protein (GFP) expression was examined 3–5 days after fertilization (Supplementary Fig. 5). We observed enhanced GFP expression compared with vector control, suggesting robust enhancer activity for several of these CREs. To take a more quantitative approach for comparing activities of surface and cave elements, we switched to reporter assays in cell lines. In the absence of available *Astyanax* liver cell lines, we performed luciferase reporter assays in the zebrafish liver (ZFL) cell line and the human liver (HepG2) cell line (Fig. 4b–d). We tested surface and Pachón alleles for each of the 25 CREs in replicates. We found that 80% of the tested CREs (20 out of 25) (Fig. 4c) mediated expression twofold or higher compared with the empty vector control, indicating their ability to function as enhancers in ZFL cell lines. We observed that 32% of these CREs were also functional in human cells (Fig. 4d), suggesting conservation of regulatory function across large evolutionary distances. For 7 out of the 20 functional CREs, the surface and cave alleles displayed differential enhancer capabilities when tested in ZFL cells, whereas 5 out of 8 CRE enhancers were differential in HepG2 (Fig. 4c,d). These results show that our selection criteria are not sufficient to predict the outcome in the reporter assay in vitro. This can be due to several reasons. First, the reporter assays have been performed in cell lines from different species, and the cellular environment in cell lines could be different from that of adult *Astyanax* liver. Second, the CREs were being tested out of their genomic context. Third, not all SNPs or indels will necessarily result in a functional consequence. Nonetheless, these reporter assays are strong and robust tool to functionally annotate and identify CREs that differ in their activity solely because of underlying sequence differences. Combining the results from ZFL and HepG2, we identified that CRE-1, CRE-7, CRE-15 and CRE-20 maintain their differential reporter output in both cell lines tested, confirming that polymorphisms in the underlying DNA sequences of these CREs are causal for differences in their ability to drive reporter expression.

Deletion of repressor binding site increases CRE-15 activity. We further characterized the cave-biased CRE-15 and generated stable transgenic lines for both surface allele (S-*CRE-15*) and cave allele (P-*CRE-15*) *of CRE-15* in zebrafish (Supplementary Fig. 5d). We observed reporter expression in the anterior gut and liver region, suggesting a tissue-specific expression pattern for this CRE (hereafter *CRE-hpdb*) similar to that of its nearby gene (744 base pairs (bp) away) *hpdb* (ENSAMXG00000015502) (Fig. 5a). *hpdb* catalyzes the first unidirectional step in tyrosine catabolism, and is upregulated in Tinaja and the most upregulated gene in the transcriptome of Pachón (Supplementary Fig. 6a and Supplementary Data 2), which we also validated using quantitative PCR (qPCR) (Fig. 5b).

Next, we used Pachón–surface F_1 hybrids to assess whether the decreased expression of *hpdb* in surface fish is due to *cis*-mediated repression or due to the presence of some *trans*-acting factor(s)[48]. The qPCR quantification of *hpdb* expression in surface and Pachón matched the RNA-seq data, and the livers of the F_1 hybrid fish expressed intermediate levels of the RNA, indicating that the Pachón allele is incompletely dominant (Fig. 5b). We monitored allele-specific expression levels by taking advantage of the presence of a synonymous SNP in exon 12 of the *hpdb* coding region to distinguish the parental alleles (Fig. 5c). Although we detected both alleles in the DNA samples, we detected expression from only the Pachón allele in the mRNA samples. This suggests that the increased expression of *hpdb* in Pachón is mediated by changes in *cis*.

We identified several small deletions in the *CRE-hpdb* sequence in Pachón that are predicted to abolish a binding motif for the repressor protein IRF2 (Fig. 5a, lower panel). As there was no significant difference in *irf2* expression between surface and Pachón (Supplementary Fig. 6b), we explored the possibility that differential binding of the IRF2 protein could be linked to the regulatory difference in *CRE-hpdb* by using an in vitro electrophoretic mobility shift assay (EMSA) (Fig. 5d). We carried out the EMSA using γ-^{32}P-labeled 20-bp oligonucleotides from the surface fish enhancer containing the IRF2 binding site and the corresponding 20-bp region from the Pachón enhancer lacking that site. We used

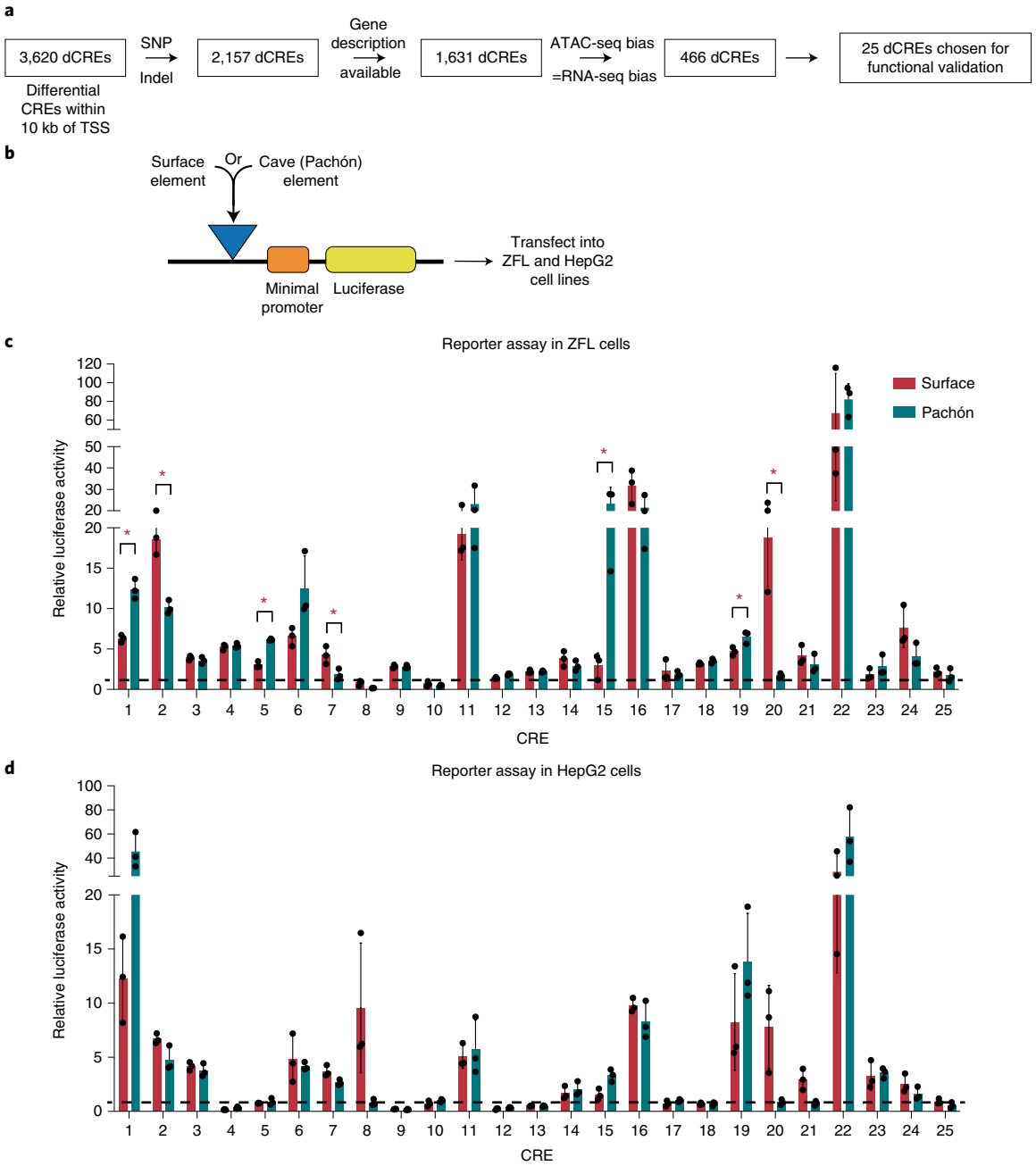


Fig. 4 | Functional validation of differentially accessible CREs. a, Flowchart showing the process of selection of candidate differential CREs (dCREs) for functional testing. **b**, Schematic of the reporter construct used for assaying enhancer activity in cell lines. **c**,**d**, Enhancer activity for the 25 enhancer candidates each from surface and Pachón tested using luciferase assay in zebrafish liver (ZFL) cells (**c**) and HepG2 cells (**d**). The asterisk (*) indicates candidates with enhancer activity that is significantly different between surface and Pachón CRE constructs (*$P < 0.05$ using two-tailed Student's *t*-test). The bars represents mean values, and error bars represent s.d. between three biological replicates. The horizontal dashed line marks activity of vector alone (normalized to 1).

recombinant human IRF2 protein, as the DNA-binding domain between *Astyanax* and human is highly conserved (Supplementary Fig. 6c). IRF2 binds to the oligo based on the surface fish sequence, whereas very weak or no binding was observed with the Pachón probe, showing altered affinity of this region for the IRF2 protein (Fig. 5d). Nonspecific competition with an unrelated oligo did not hamper the robust DNA–protein interaction (Fig. 5d), wheras addition of 200× (20 pmol) (lanes 4 and 7) and 400× (40 pmol) (lanes 5 and 8) unlabeled self-competitor weakened the interaction, confirming the specificity of the binding. These results suggest that the deletion in P-*CRE-hpdb* prevents the IRF2 repressor from binding, which in turn could affect the activity of this regulatory region.

Next, we tested whether the lack of an IRF2 site is sufficient to abolish repression in vitro. Using the cell-based luciferase assay, we found that specifically deleting the IRF2 motif in the surface allele of *CRE-hpdb* (surface$^{\text{IRF2-del}}$, now similar to the Pachón allele) restored expression (Fig. 5e). This suggests that IRF2 is indeed a repressor of *CRE-hpdb*. Next, we asked whether repression could be rescued by adding back the IRF2 binding site to the Pachón sequence. Surprisingly, we noted that the addition of the IRF2 site was not

sufficient to cause significant repression (Pachón$^{IRF2\text{-}ins}$) (Fig. 5e), suggesting that other sequence changes have occurred in Pachón that prevent repression by IRF2. We sequentially converted two more sites to the Pachón alleles (Supplementary Data 3) (Pachón$^{IRF2\text{-}ins\text{-}ATA}$ and Pachón$^{IRF2\text{-}ins\text{-}ATA\text{-}TT}$), resulting in restoration of the repressive activity of the IRF2 site insertion in P-*CRE-hpdb* (Fig. 5e). This suggests that, although deleting the IRF2 binding site is sufficient to release repression, other mutations have contributed to the differential activity of the CRE.

Deletion of IRF2 binding site as an adaptive trait. Tyrosine serves as a substrate for several important compounds in the cell, including melanin, dopamine and certain intermediates for ketone body formation and the TCA cycle (Fig. 5f). An accumulation of excess tyrosine in cells has been reported in cave populations that are mutant for melanin formation and in surface fish upon abrogation of melanin formation by knocking down *oca2*[49]. We also noticed in our liver transcriptome data that the expression of genes encoding enzymes that convert tyrosine to L-DOPA (tyrosinase and tyrosine hydroxylase) is very low (Supplementary Fig. 6d). In a nutrient-deprived condition, it could be a thrifty strategy to divert the excess tyrosine to produce TCA intermediates and ketone bodies for energy storage and production.

To investigate this possibility, we genotyped 23 wild-caught surface fish, 23 wild Pachón fish and one each of Tinaja, Yerbaniz, Piedras and Japonés cavefishes. We found that the IRF2 site deletion is fixed in all cavefish samples, except for the Japonés individual, which was heterozygous (Supplementary Fig. 6e), whereas all surface fishes had the wild-type sequence. This supports previously published data, in which we found the expression of *hpdb* from wild Pachón to be higher than that in wild surface fish (Fig. 5g)[50]. These observations suggest that the mutation is under positive selection[41,49]. We further analyzed the expression of other relevant enzymes in the tyrosine catabolism pathway and found that the genes *hpdb*, *hgd* and *fah*, which encode enzymes that catalyze unidirectional steps in the tyrosine catabolism pathway, also have higher expression in Pachón than in surface livers (Fig. 5b and Supplementary Fig. 6f). This supports our hypothesis that excess tyrosine in the cells is being repurposed by the TCA cycle.

Discussion

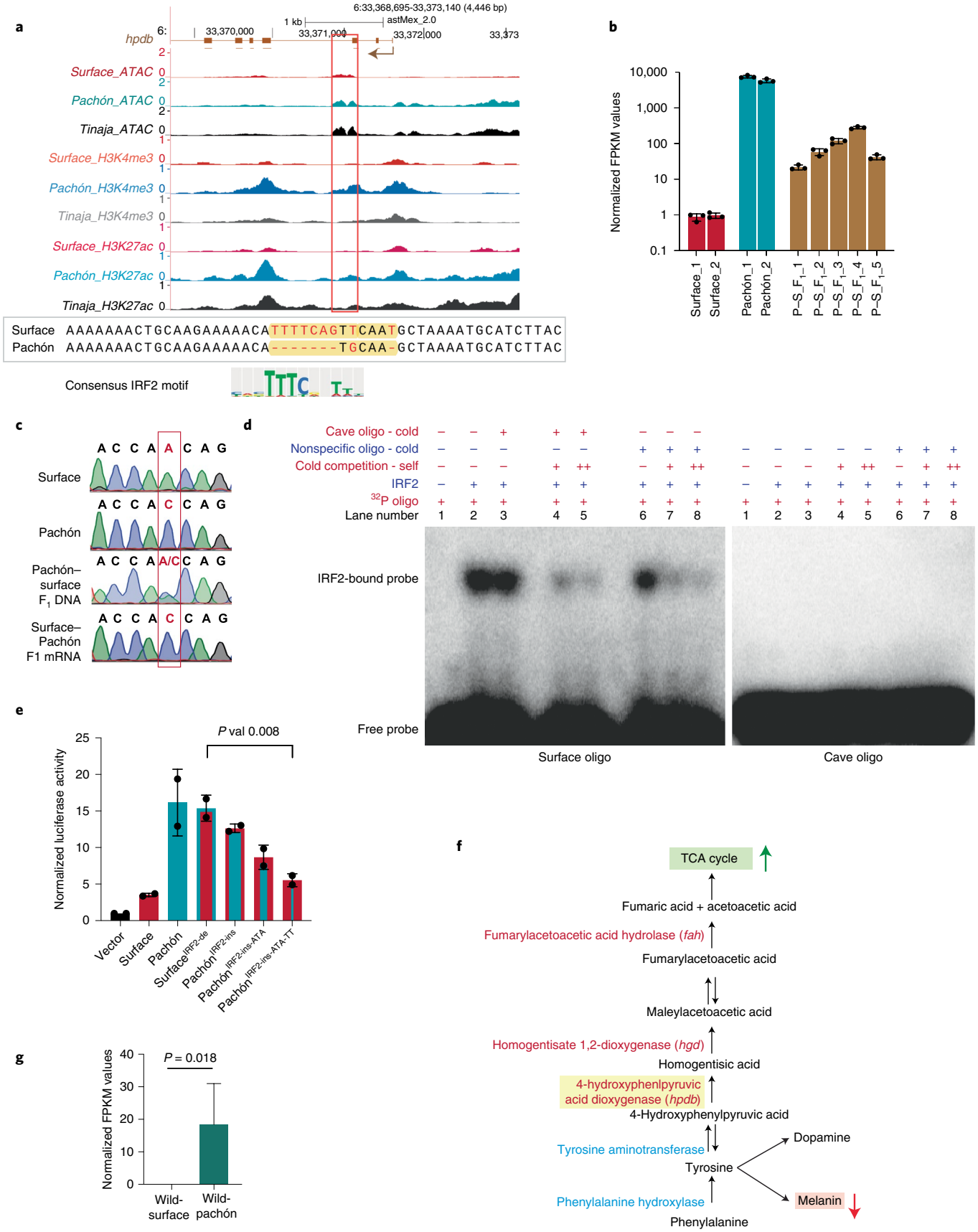
In this study, we used *Astyanax mexicanus*, an evolutionary genetic model system, to examine how changes in the CREs of the genome can help cavefish adapt to the extreme cave environment characterized by nutrient deprivation and the absence of light. Our analysis uncovered many putative CREs that are differentially regulated between surface and cave morphotypes. We found that surface-biased CREs were associated with key genes related to circadian rhythm, lipid catabolism and TGF-β signaling, whereas cave-biased CREs were associated with lipid anabolism genes and the immune system. The biased enrichment of motifs for key TFs such as LXR and HNF4A in s-CREs, and NF-Y and KLF14 in c-CREs, points to key TF networks that could be linked to changes in the above metabolic pathways. These global regulatory changes reveal how adaptation to the cave environment has modified the regulatory architecture of the genome to support physiological traits and metabolic processes in cavefish compared with that of surface fish.

Our study also highlights that the genome-wide chromatin architecture of the two cavefish morphotypes, Pachón and Tinaja, were more similar to each other than to the surface fish. This is in line with the previously observed phenotypic convergence in these independently derived cavefish populations such as loss of visible eyes and pigment, accumulation of excess fat, and insulin resistance[12,30,51]. One interesting hypothesis supporting this observation is that selection has repeatedly acted on existing transcriptional gene regulatory networks in the animals upon exposure to new environments.

Functional validation of our genome-wide analyses in cell culture assays revealed that a large proportion of CREs have altered activities, potentially due to *trans*-acting effects, whereas some arise due to underlying genetic changes. Our analyses of genetic variation underlying differentially accessible CREs suggest its contribution to the gain of accessible chromatin regions in cavefish. Detailed genetic and functional characterization of an enhancer of the *hpdb* gene affirmed the role of *cis*-regulatory changes in controlling the cave-biased expression of *hpdb* in the liver. The *hpdb* gene is part of the tyrosine metabolism pathway and converts tyrosine to the TCA cycle intermediate fumarate and the ketone body acetoacetate. It has been shown that the absence of melanin, as seen in cavefish, results in more tyrosine and dopamine in the body[49]. An intriguing possibility supported by the presence of this mutated CRE in multiple cave populations is that, in the liver, excess tyrosine can be converted to fumarate and acetoacetate as a measure to use any available nutrient for energy production in cavefish.

We expect that many other differentially accessible regions identified in our study are involved in the metabolic adaptation of cavefish to their low-nutrient environment. We propose that *Astyanax mexicanus*, with its contrasting morphotypes and independently derived cave populations, presents an effective system to unravel global gene regulatory pathways and networks that are important in the physiological adaptation of species to new and changing environments and could give insight toward a better understanding of conserved metabolic processes in vertebrate physiology.

Fig. 5 | Detailed characterization of CRE-15. a, UCSC Genome Browser screenshot showing various chromatin features at the genomic region around *CRE-hpdb*. The *y* axis denotes reads per million. The red box indicates the location of CRE-15 (744 bp from the TSS). The lower panel shows the sequence of the region of *CRE-hpdb* that contains the deletion of the predicted binding site. The cognate IRF2 motif is also shown. **b**, *hpdb* RNA levels using qPCR in livers of adult surface fish, Pachón cavefish and Pachón–surface (P–S) F_1 hybrids. Each bar represents data from one individual fish. Error bars represent s.d. between three technical replicates of qPCR. All values are relative to Surface_1 fish normalized to 1. **c**, Chromatograms from the sequencing of the SNP within exon 12 of the *hpdb* gene used to distinguish surface and Pachón alleles. **d**, Representative gel (for three independent experiments) showing gel shift assay for binding of recombinant IRF2 on surface fish and cavefish oligo spanning the IRF2 binding site. Excess ^{32}P-labeled oligo runs at the bottom, and IRF2-bound surface oligo runs slower. Self-competition was done using 200× (+) and 400× (++) unlabeled oligos. Cavefish oligomer and a random oligomer were used as nonspecific competitors. Radiolabeled surface oligo binds the protein and does not get competed out by excess unlabeled cave or nonspecific oligo. Cavefish oligo fails to bind any IRF2. **e**, Relative luciferase activities for vector alone and various alleles of *CRE-hpdb*: surface, Pachón, S-*CRE-hpdb* without IRF2 binding site (surface$^{IRF2\text{-}del}$), P-*CRE-hpdb* with the IRF2 binding site restored (Pachón$^{IRF2\text{-}ins}$) and Pachón$^{IRF2\text{-}ins}$ with additional mutations that convert the Pachón allele to a surface allele (Pachón$^{IRF2\text{-}ins\text{-}ATA}$ and Pachón$^{IRF2\text{-}ins\text{-}ATA\text{-}TT}$). The *P* value was calculated using a two-tailed Student's *t*-test. The graph shows mean values ± s.d. from $n = 3$ biological replicates. **f**, Schematic of pathways that use tyrosine in the cell. A decreased demand for melanin in cavefish could in principle lead to increased availability of tyrosine for other pathways. **g**, *hpdb* RNA levels in wild-caught surface fish and Pachón cavefish livers (RNA-seq data)[50]. The graph shows mean values ± s.d. from $n = 3$ biologically independent RNA-seq experiments. The *P* value was calculated using a two-tailed Student's *t*-test.


a
6:33,368,695-33,373,140 (4,446 bp)
1 kb
astMex_2.0
hpdb
Surface_ATAC
Pachón_ATAC
Tinaja_ATAC
Surface_H3K4me3
Pachón_H3K4me3
Tinaja_H3K4me3
Surface_H3K27ac
Pachón_H3K27ac
Tinaja_H3K27ac
Surface AAAAAAACTGCAAGAAAAACATTTTCAGTTCAATGCTAAAATGCATCTTAC
Pachón AAAAAAACTGCAAGAAAAACA-------TGCAA-GCTAAAATGCATCTTAC
Consensus IRF2 motif
b
Normalized FPKM values
Surface_1
Surface_2
Pachón_1
Pachón_2
P–S_F1_1
P–S_F1_2
P–S_F1_3
P–S_F1_4
P–S_F1_5
c
Surface
Pachón
Pachón–surface F1 DNA
Surface–Pachón F1 mRNA
d
Cave oligo - cold
Nonspecific oligo - cold
Cold competition - self
IRF2
32P oligo
Lane number
IRF2-bound probe
Free probe
Surface oligo
Cave oligo
e
P val 0.008
Normalized luciferase activity
Vector
Surface
Pachón
SurfaceIRF2-del
PachónIRF2-ins
PachónIRF2-ins-ATA
PachónIRF2-ins-ATA-TT
f
TCA cycle
Fumaric acid + acetoacetic acid
Fumarylacetoacetic acid hydrolase (fah)
Fumarylacetoacetic acid
Maleylacetoacetic acid
Homogentisate 1,2-dioxygenase (hgd)
Homogentisic acid
4-hydroxyphenlpyruvic acid dioxygenase (hpdb)
4-Hydroxyphenylpyruvic acid
Tyrosine aminotransferase
Tyrosine
Dopamine
Phenylalanine hydroxylase
Melanin
Phenylalanine
g
P = 0.018
Normalized FPKM values
Wild-surface
Wild-pachón

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-022-01049-4.

Received: 12 August 2020; Accepted: 9 March 2022;
Published online: 12 May 2022

References

1. Wittkopp, P. J. & Kalay, G. *Cis*-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat. Rev. Genet.* **13**, 59–69 (2012).
2. Long, H. K., Prescott, S. L. & Wysocka, J. Ever-changing landscapes: transcriptional enhancers in development and evolution. *Cell* **167**, 1170–1187 (2016).
3. Prescott, S. L. et al. Enhancer divergence and *cis*-regulatory evolution in the human and chimp neural crest. *Cell* **163**, 68–83 (2015).
4. Thompson, A. C. et al. A novel enhancer near the *Pitx1* gene influences development and evolution of pelvic appendages in vertebrates. *eLife* **7**, e38555 (2018).
5. Partha, R. et al. Subterranean mammals show convergent regression in ocular genes and enhancers, along with adaptation to tunneling. *eLife* **6**, e25884 (2017).
6. Gore, A. V. et al. An epigenetic mechanism for cavefish eye degeneration. *Nat. Ecol. Evol.* **2**, 1155–1160 (2018).
7. Jeffery, W. R. *Astyanax* surface and cave fish morphs. *EvoDevo* **11**, 14 (2020).
8. Krishnan, J. & Rohner, N. Sweet fish: fish models for the study of hyperglycemia and diabetes. *J. Diabetes* **11**, 193–203 (2019).
9. Bradic, M., Beerli, P., García-de León, F. J., Esquivel-Bobadilla, S. & Borowsky, R. L. Gene flow and population structure in the Mexican blind cavefish complex (*Astyanax mexicanus*). *BMC Evol. Biol.* **12**, 9 (2012).
10. Herman, A. et al. The role of gene flow in rapid and repeated evolution of cave-related traits in Mexican tetra, *Astyanax mexicanus*. *Mol. Ecol.* **27**, 4397–4416 (2018).
11. Coghill, L. M., Hulsey, C. D., Chaves-Campos, J., García de Leon, F. J. & Johnson, S. G. Next generation phylogeography of cave and surface *Astyanax mexicanus*. *Mol. Phylogenet. Evol.* **79**, 368–374 (2014).
12. Riddle, M. R. et al. Insulin resistance in cavefish as an adaptation to a nutrient-limited environment. *Nature* **555**, 647–651 (2018).
13. Aspiras, A. C., Rohner, N., Martineau, B., Borowsky, R. L. & Tabin, C. J. Melanocortin 4 receptor mutations contribute to the adaptation of cavefish to nutrient-poor conditions. *Proc. Natl Acad. Sci. USA* **112**, 9668–9673 (2015).
14. Rui, L. Energy metabolism in the liver. *Compr. Physiol.* **4**, 177–197 (2014).
15. Dowling, T. E., Martasian, D. P. & Jeffery, W. R. Evidence for multiple genetic forms with similar eyeless phenotypes in the blind cavefish, *Astyanax mexicanus*. *Mol. Biol. Evol.* **19**, 446–455 (2002).
16. Gross, D. S. & Garrard, W. T. Nuclease hypersensitive sites in chromatin. *Annu. Rev. Biochem.* **57**, 159–197 (1988).
17. Daugherty, A. C. et al. Chromatin accessibility dynamics reveal novel functional enhancers in *C. elegans*. *Genome Res.* **27**, 2096–2107 (2017).
18. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nat. Rev. Genet.* **20**, 207–220 (2019).
19. Warren, W. C. et al. A chromosome-level genome of *Astyanax mexicanus* surface fish for comparing population-specific genetic differences contributing to trait evolution. *Nat. Commun.* **12**, 1447 (2021).
20. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).
21. Zhou, V. W., Goren, A. & Bernstein, B. E. Charting histone modifications and the functional organization of mammalian genomes. *Nat. Rev. Genet.* **12**, 7–18 (2011).
22. Cooper, G. M. et al. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15**, 901–913 (2005).
23. Villar, D. et al. Enhancer evolution across 20 mammalian species. *Cell* **160**, 554–566 (2015).
24. Schmidt, D. et al. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328**, 1036–1040 (2010).
25. Hong, J.-W., Hendrix, D. A. & Levine, M. S. Shadow enhancers as a source of evolutionary novelty. *Science* **321**, 1314 (2008).
26. Wong, E. S. et al. Decoupling of evolutionary changes in transcription factor binding and gene expression in mammals. *Genome Res.* **25**, 167–178 (2015).
27. Hariprakash, J. M. & Ferrari, F. Computational biology solutions to identify enhancers-target gene pairs. *Comput. Struct. Biotechnol. J.* **17**, 821–831 (2019).
28. Mifsud, B. et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.* **47**, 598–606 (2015).
29. Fishilevich, S. et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database* **2017**, bax028 (2017).
30. Xiong, S., Krishnan, J., Peuß, R. & Rohner, N. Early adipogenesis contributes to excess fat accumulation in cave populations of *Astyanax mexicanus*. *Dev. Biol.* **441**, 297–304 (2018).
31. Flannick, J. et al. Loss-of-function mutations in *SLC30A8* protect against type 2 diabetes. *Nat. Genet.* **46**, 357–363 (2014).
32. Eissing, L. et al. De novo lipogenesis in human fat and liver is linked to ChREBP-β and metabolic health. *Nat. Commun.* **4**, 1528 (2013).
33. Ham, M. et al. Glucose-6-phosphate dehydrogenase deficiency improves insulin resistance with reduced adipose tissue inflammation in obesity. *Diabetes* **65**, 2624–2638 (2016).
34. He, Y. et al. The role of retinoic acid in hepatic lipid homeostasis defined by genomic binding and transcriptome profiling. *BMC Genomics* **14**, 575 (2013).
35. Laurencikiene, J. & Rydén, M. Liver X receptors and fat cell metabolism. *Int. J. Obes.* **36**, 1494–1502 (2012).
36. Weissglas-Volkov, D. et al. Common hepatic nuclear factor-4α variants are associated with high serum lipid levels and the metabolic syndrome. *Diabetes* **55**, 1970–1977 (2006).
37. Lu, Y.-H., Dallner, O. S., Birsoy, K., Fayzikhodjaeva, G. & Friedman, J. M. Nuclear Factor-Y is an adipogenic factor that regulates leptin gene expression. *Mol. Metab.* **4**, 392–405 (2015).
38. Truty, M. J., Lomberk, G., Fernandez-Zapico, M. E. & Urrutia, R. Silencing of the transforming growth factor-β (TGFβ) receptor II by Krüppel-like factor 14 underscores the importance of a negative feedback mechanism in TGFβ signaling. *J. Biol. Chem.* **284**, 6291–6300 (2009).
39. Phillips-Cremins, J. E. et al. Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* **153**, 1281–1295 (2013).
40. Kentepozidou, E. et al. Clustered CTCF binding is an evolutionary mechanism to maintain topologically associating domains. *Genome Biol.* **21**, 5 (2020).
41. Chan, Y. F. et al. Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a *Pitx1* enhancer. *Science* **327**, 302–305 (2010).
42. Bradic, M., Teotónio, H. & Borowsky, R. L. The population genomics of repeated evolution in the blind cavefish *Astyanax mexicanus*. *Mol. Biol. Evol.* **30**, 2383–2400 (2013).
43. Coetzee, S. G., Coetzee, G. A. & Hazelett, D. J. motifbreakR: an R/Bioconductor package for predicting variant effects at transcription factor binding sites. *Bioinformatics* **31**, 3847–3849 (2015).
44. Raile, K. et al. *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 (2008).
45. Howe, K. L. et al. Ensembl 2021. *Nucleic Acids Res.* **49**, D884–D891 (2021).
46. Fisher, S. et al. Evaluating the biological relevance of putative enhancers using Tol2 transposon-mediated transgenesis in zebrafish. *Nat. Protoc.* **1**, 1297–1305 (2006).
47. Parker, H. J., Bronner, M. E. & Krumlauf, R. A *Hox* regulatory network of hindbrain segmentation is conserved to the base of vertebrates. *Nature* **514**, 490–493 (2014).
48. Wittkopp, P. J., Haerum, B. K. & Clark, A. G. Evolutionary changes in *cis* and *trans* gene regulation. *Nature* **430**, 85–88 (2004).
49. Bilandžija, H., Ma, L., Parkhurst, A. & Jeffery, W. R. A potential benefit of albinism in *Astyanax* cavefish: downregulation of the *oca2* gene increases tyrosine and catecholamine levels as an alternative to melanin synthesis. *PLoS ONE* **8**, e80823 (2013).
50. Krishnan, J. et al. Comparative transcriptome analysis of wild and lab populations of *Astyanax mexicanus* uncovers differential effects of environment and morphotype on gene expression. *J. Exp. Zool. B Mol. Dev. Evol.* **334**, 530–539 (2020).
51. Jeffery, W. R. Regressive evolution in *Astyanax* cavefish. *Annu. Rev. Genet.* **43**, 25–47 (2009).

Methods

***Astyanax* husbandry.** *Astyanax* are housed in polycarbonate or glass fish tanks on racks (Pentair Aquatic Eco-Systems). The surface fish and the cavefish are maintained under the exact same laboratory conditions. They are provided with the same food, and both kept under the same 14/10 h light/dark cycle. As we are mainly interested in understanding the genetic changes between the morphotypes, maintaining them under the same conditions helps negate potential environmental effects. Each rack uses an independent recirculating aquaculture system with mechanical, chemical and biologic filtration, and UV disinfection. Water quality parameters are maintained within safe limits (upper limit of total ammonia nitrogen range, 1 mg l^{-1}; upper limit of nitrite range, 0.5 mg l^{-1}; upper limit of nitrate range, 60 mg l^{-1}; temperature set-point, 22 °C; pH, 7.65, specific conductance, 800 µS cm^{-1}; dissolved oxygen, >90%). Water changes range from 20% to 30% daily (supplemented with Instant Ocean Sea Salt (Spectrum Brands)). Adult fish are fed three times per day during breeding weeks and one time per day during nonbreeding weeks on a diet of Mysis shrimp (Hikari Sales USA) and Gemma 800 (Skretting USA). Animal husbandry followed protocol 2019-084 approved for the Stowers Institute for Medical Research.

Zebrafish husbandry. Zebrafish are housed in polycarbonate fish tanks on racks (Pentair Aquatic Eco-Systems) with a 14/10 h light/dark photoperiod. Racks are supplied by two recirculating aquaculture systems with mechanical, chemical and biological filtration, and UV disinfection. Water quality parameters are maintained within safe limits (upper limit of total ammonia nitrogen range, 0.5 mg l^{-1}; upper limit of nitrite range, 0.5 mg l^{-1}; upper limit of nitrate range, 40 mg l^{-1}; temperature set-point, 28.5 °C; pH, 7.60, specific conductance, 500 µS cm^{-1}; dissolved oxygen, >85%). Water changes range from 20% to 30% daily (supplemented with Instant Ocean Sea Salt). Adult zebrafish are fed twice daily with one feed of hatched *Artemia* (first instar) (Brine Shrimp Direct) and one feed of Zeigler Adult Zebrafish Diet (Zeigler Bros). Embryos up to 5 days after fertilization were maintained at 28.5 °C in E2 embryo media. Animal husbandry followed protocol 2019-078 (zebrafish) approved for the Stowers Institute for Medical Research.

ATAC-seq. For harvesting tissues, the fish were euthanized in tricaine methanesulfonate (MS-222) and immediately dissected. Dissections were performed in the morning (3 h after light turns on) after fasting the fish overnight. Livers were dissected from three fish each for surface, Pachón and Tinaja populations and were divided into two parts for RNA-seq and ATAC-seq. ATAC-seq was performed as per ref. [52] with some modifications to accommodate the use of whole tissues as starting material instead of cells. Livers were homogenized (30–40 strokes) using the loose pestle of a Dounce homogenizer in lysis buffer (10 mM Tris·Cl, pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% (v/v) Igepal CA-630) to prepare nuclei. The rupturing of the cell membrane and obtainment of nuclei was confirmed using Trypan blue staining under a phase contrast microscope. Nuclei were counted under a microscope, and ~70,000 nuclei were taken and spun down at 1,500 *g* at 4 °C. The nuclei were resuspended in transposition mixture (25 µl TD of 2X reaction buffer from a Nextera kit, 2.5 µl TDE1 Tn5 Transposase from a Nextera kit, 22.5 µl nuclease-free water) and incubated at 37 °C for 22–24 min. The reaction was purified, and the library was prepared as per the original protocol[52]. Paired-end sequencing was performed using the Illumina NextSeq Mid-Output mode. Quality control data consisting of the number of reads per sample and the number of peaks obtained are available in Supplementary Table 3.

RNA-seq. The part of the livers that was used for RNA-seq was frozen and later used for RNA extraction using the Qiagen RNeasy kit. RNA-seq and ATAC-seq were performed using the same liver samples to get maximum possible correlation between chromatin accessibility. Libraries were prepared according to the manufacturer's instructions using the TruSeq Stranded mRNA Prep Kit (Illumina). The resulting libraries were purified using the Agencourt AMPure XP system (Beckman Coulter), then quantified using a Bioanalyzer (Agilent Technologies) and a Qubit fluorometer (Life Technologies). Libraries were requantified, normalized, pooled and sequenced on an Illumina HiSeq 2500 instrument as 50-bp single read. Following sequencing, Illumina Real-Time Analysis v1.18.64 and bcl2fastq2 v2.20 were run to demultiplex reads and generate FASTQ files.

ChIP-seq. Adult livers were dissected and fixed in 1% formaldehyde for 15 min. Chromatin was prepared using the MAGnify Low Cell ChIP-Seq Kit (Thermo Fisher). Antibodies were validated by western blotting (Supplementary Fig. 6g). Antitubulin antibody was used for loading control. Antihistone antibodies were used in a 1:1,000 dilution, and antitubulin was used in a 1:10,000 dilution. ChIP-seq was done for H3K4me3 (Millipore, 07-473), H3K27ac (Abcam, ab4729) and H3K27me3 (Abcam, ab195477). Approximately 200,000 cells and 1 µg of antibody were used for each sample. Libraries were prepared using the KAPA HTP Library Preparation Kit for Illumina and Bioo Scientific NEXTflex DNA barcodes. The resulting libraries were purified using the Agencourt AMPure XP system, then quantified using a Bioanalyzer and a Qubit fluorometer. Postamplification size selection was performed on all libraries using a Pippin Prep (Sage Science). Libraries were requantified, normalized, pooled and sequenced on Illumina HiSeq 2500 instrument as 50-bp single read. Following sequencing, Illumina Real-Time Analysis v1.18.64 and bcl2fastq2 v2.18.0.12 were run to demultiplex reads and generate FASTQ files.

RNA-seq analysis. Reads were aligned to astMex_2.0 and gene models from Ensembl 96 with STAR v2.6.1c[53] with –outFilterMultimapNmax 2 and–quantMode GeneCounts to generate a count table. The gene counts were analyzed in R using the edgeR library[54] to identify genes differentially expressed between Pachón vs surface and Tinaja vs surface. *P* values were adjusted for multiple hypothesis testing using the method in ref. [55].

ChIP-seq and ATAC-seq analyses. The surface fish genome, which has been publicly available on Ensembl since 2017 (astMex_2.0), was used as the reference genome for all analyses in the study unless otherwise mentioned. Approximately 40 million reads were produced for each sample. For the analysis of chromatin marks, FASTQ files were aligned to astMex_2.0 using Bowtie 2 under default parameters. Reads mapping to the mitochondrial chromosome were filtered out using samtools prior to calling peaks with MACS2 under default parameters. ATAC-seq data were aligned to astMex_2.0 using Bowtie 2 with the following parameters: -X 2000 –very-sensitive. More than 80% of the reads successfully mapped to the reference genome (Supplementary Table 3). The duplication rate was assessed at less than 0.4% per sample using Picard MarkDuplicates. Fragment size distribution was examined with ATACseqQC[56] from Bioconductor, which indicated a mean fragment length of 200 bases, decreasing in frequency with a periodicity of 200 bp. Peaks were called with MACS2 using default parameters. Supplementary Table 3 lists the number of peaks obtained for each fish type. For each chromatin mark and the ATAC-seq dataset, a reference set of peaks was created for each type of fish by reducing peaks from multiple replicates to a single peak set, and keeping only peaks observed in at least two replicates. The ATAC-seq peak reference loci were quantified across all samples by tallying read overlaps in R, and were analyzed for differential accessibility between fish using edgeR[54] to calculate $\log_2$[fold change] ratios and *P* values for each locus. ATAC-seq peak loci with mean counts per million less than 1 were not considered for further analysis. *P* values were adjusted using the method in ref. [55], and those with a value less than 0.0001 were considered differentially open between the two morphotypes. ATAC-seq loci were then scored for overlap (1 or 0) with each chromatin mark, and were mapped to the nearest gene using the GenomicRanges library in R. For analysis of ChromHMM, pathway and motif enrichment, we used CREs within 10 kb of the nearest TSS.

Heatmaps. Heatmaps for signals around TSS regions (promoters) show the number of reads per million normalized to ChIP-seq signals around TSS regions (2 kb upstream and downstream) of 18,836 transcripts. The criteria for transcript selection are Ensembl 94 protein-coding genes, the longest transcript for each gene, and transcripts that did not overlap. The 4-kb TSS regions were binned into 100 bins (40 bp per bin), and the mean number of reads per million signal was calculated for each bin. TSS regions were ordered based on average ATAC-seq signals in decreasing order.

Heatmaps for differentially accessible CREs consist of 1,698 cave-biased and 1,880 surface-biased peaks with a *q* value less than 0.05. In addition, the differential peaks had *q* values less than 0.05 for both Pachón vs surface and Tinaja vs surface. Conversely, invariant peaks represented in the heatmap had *q* values (Pachón vs surface, Tinaja vs surface, and Pachón vs Tinaja) that were all greater than 0.90 (1,991 peaks). We plotted the number of reads per million normalized ATAC-seq signals (row-scaled as *z*-scores) in ±1-kb regions around peak centers using the R package EnrichedHeatmap[57]. We plotted the expression (normalized transcripts per million, row-scaled) of the nearest gene (within a 10-kb region) for differentially accessible CREs in the three morphotypes. Genes are ordered based on mean fold change between surface fish and cavefish. For surface-biased peaks, genes have higher expression in surface fish than in cavefish (one-sided *t*-test *P* values are 0.014 for surface vs Pachón and 0.052 for surface vs Tinaja); for cave-biased peaks, genes have higher expression in cavefish than in surface fish (one-sided *t*-test *P* values are 0.033 for Pachón vs surface and 0.031 for Tinaja vs surface).

Conservation. All conservation analyses were done using the astMex1 genome. Genomic Evolutionary Rate Profiling (GERP) regions and their conservation scores for 11 fish conservation were obtained from Ensembl. For obtaining evolutionarily conserved CREs, overlap was seen between open chromatin regions and evolutionarily constrained regions from GERP. We omitted exons from the GERP regions to prevent bias from the highly conserved exonic regions. To obtain the background level of conservation, we obtained size-matched random sets of regions from the genome, and the process was iterated 1,000 times. To find orthologs of known human liver enhancers, we downloaded human multiz100way genome alignments from the UCSC Genome Browser, and pairwise alignments between human (hg19) and cavefish (astMex1) were extracted using UCSC Genome Browser utilities (mafSpeciesSubset). Then, bedtools v2.26.0 was used to map human enhancers to their nearest conserved regions[58].

ChromHMM. We applied ChromHMM v1.15 on H3K27ac, H3K4me3 and H3K27me3 histone marks for three morphotypes types (surface, Pachón and Tinaja). The cavefish genome for ChromHMM was built using Ensembl 94 cavefish genome and annotation release. ChIP-seq BAM files were binarized using the BinarizeBam command with a bin size (-b option) of 200 and Poisson threshold (-p option) of 0.001. Next, we built a seven-state hidden Markov model (HMM) using learnModel with default parameters.

ChromHMM state distribution. ATAC-seq consensus peaks were resized to 400 bp around the peak center. The cavefish genome was binned into 200-bp nonoverlapping bins using the bedtools makewindows command. Then, we assigned a ChromHMM state to each ATAC-seq peak or genome bin, requiring that the state covered at least 50% of the ATAC-seq peak or genome bin. Next, ChromHMM state distribution was calculated for ATAC-seq peaks and genome bins by counting the occurrences of each assigned state.

ChromHMM state enrichment. We randomly placed the ATAC-seq consensus peaks (resized to 400 bp) onto the cavefish genome using the bedtools shuffle command (with the -noOverlapping option). This shuffling process was repeated 1,000 times. For each shuffled ATAC-seq peak, we computed its ChromHMM state distribution in the same way as mentioned above and calculated the $\log_2$ enrichment ratio between the true and shuffled distributions. The bar graph in Supplementary Fig. 1e shows the mean enrichment ratio and s.d. ($n = 1{,}000$) for each ChromHMM state.

Motif and pathway analysis. Motif enrichment analysis was done using HOMER[59] with the command 'findMotifsGenome.pl -size given', and known vertebrate motifs were analyzed. Pathway (reactome) analysis was performed using the Molecular Signatures Database of the Gene Set Enrichment Analysis (GSEA)[60,61] with default parameters after converting *Astyanax* Ensembl gene IDs to human IDs.

Variant calling. Reads from paired-end ATAC-seq FASTQ files were aligned to the astMex_2.0 surface fish genome from Ensembl and marked by Read Group with BWA mem v0.7.17. BAM files were merged and deduplicated using GATK[62]. Reads were processed for calling variants with HaplotypeCaller using a GATK best practices pipeline implemented with Snakemake. Two rounds of bootstrapping with filtering were used to create a reference set of variants for Base Quality Score Recalibration prior to calling variants against astMex_2.0. SNPs and indels were selected and hard filtered based on GATK generic hard filtering recommendations (in brief, for SNPs, QD < 2.0, FS > 60.0, MQ < 40.0, ReadPosRankSum < −8.0; for indels, QD < 2.0, FS > 200.0, ReadPosRankSum < −20.0). To normalize for varying peak width, we focused on a 200-bp region spanning the peak center. In addition, we ensured that all SNPs considered for this analysis were called in both surface and cave sequences to avoid any bias due to lack of SNP call. These fish are laboratory stocks and are probably somewhat inbred; therefore, we had little power to assign allele frequencies between cave and surface pooled samples. For all analyses, we concentrate on differences that were completely fixed in our dataset between the surface genotypes and the cave genotypes.

motifbreakR. SNPs in cavefish and surface fish were annotated for their potential effect on TF binding motifs from the HOCOMOCO database from MotifDB 1.30 in Bioconductor using the R package motifbreakR[43,63]. The main function of the package was run with the following parameters: filterp=TRUE, threshold=1e-4, method = "ic", and bkg=c(A = 0.25, C = 0.25, G = 0.25, T = 0.25).

Selection of candidates for functional testing. We set a series of criteria by which we selected differential CRE candidates for functional testing. After selecting for candidates with polymorphisms, a well-annotated neighboring gene and a biased gene expression, we were left with 466 candidates to choose from. First, we looked at the details of the genomic context of the CRE. We checked whether the flanking regions of the CRE maintained the biased epigenetic signature and that there were no major unbiased peaks in the immediate flanking regions. Next, we looked for various characteristics of the neighboring gene. We selected CREs that were associated with genes with highly differential expression levels in our RNA-seq analysis, for example, *Nos* and *hpdb*. Last, we reviewed literature and focused on CREs associated with genes that were involved in metabolic processes (for example, carbohydrate or fat metabolism) or in pathways that maintain the health of tissues (for example, the inflammatory pathway). Supplementary Table 2 lists the final candidates tested in reporter assays, as well as details of the epigenetic signature and the associated genes.

Cloning and reporter assays. Candidate CREs from surface, Pachón and Tinaja genomes were amplified from the genome or synthesized commercially (GenScript) and cloned into pGL4.23 (Promega) and HLC[47] vectors using Gibson assembly (New England Biolabs, E2611). All primers used in the study, as well as their descriptions, are listed in Supplementary Table 4.

Reporter assays were performed in *Astyanax* surface fish embryos and zebrafish embryos, the adult zebrafish liver cell line ZFL and the human liver cell line HepG2. For zebrafish microinjections, differential CRE candidates were cloned into a Tol2-based vector with minimal c-*fos* promoter and downstream GFP. In general, a minimum of 100 embryos were injected to monitor activity for each construct because of mosaicism and position effects of integration. Embryos, representing >40% of the injected ones, are depicted in the Supplementary Fig. 5 pictures. *Astyanax* and zebrafish larvae were anesthetized using buffered MS-222, immersed in 3% methyl cellulose in a depression slide and imaged using a Leica stereomicroscope.

For luciferase reporter assays, the CRE candidates were cloned into the pGL4.23 vector at the EcoRV site upstream to a minimal promoter that drives the firefly luciferase gene. To control for transfection efficiency, we co-transfected a pRLTK vector (Promega) that expressed *Renilla* luciferase gene. All transfections were done using Lipofectamine LTX with Plus reagent (15338030) in 24-well plates with 350 ng of test construct and 150 ng of control plasmid. Luciferase activity was measured 48 h after transfection using a luminometer (Victor X Light, PerkinElmer). All constructs were done in two or more replicates, and relative enhancer activity was calculated by normalizing the empty vector to 1. Significance was determined using a two-tailed Student's *t*-test.

Electrophoretic mobility shift assays. Gel shift assays were performed using recombinant human IRF2 protein (Sigma, SRP6338). 10 pmol of single-stranded oligos for surface fish IRF2 binding site and corresponding mutant cavefish IRF2 site were end-labeled with a ^{32}P radioisotope using polynucleotide kinase, annealed with complementary strands to make double-stranded oligos and purified using G-25 spin columns (GE Healthcare, 27-5325-01). Binding reactions were set up using 1X binding buffer (5X binding buffer: 50 mM Tris 7.5, 375 mM NaCl, 5 mM EDTA, 30% glycerol, 15 mM spermidine, 5 mM DTT), 50 ng µl^{-1} polydI•dC, 0.25 µg of protein, 100 fmol of labeled target probe and 25 pmol of mutant or nonspecific oligo in a final volume of 20 µl. The binding reaction was set up in cold and then incubated at room temperature 22 °C for 20 min. DNA–protein complexes were separated on nondenaturing 6% DNA retardation gels (Invitrogen, EC63655BOX) at 100 V constant voltage for 50 min. After the run, the gel was fixed in gel fixation solution (40% v/v methanol, 10% v/v acetic acid) for 30 min, exposed to a Phosphor Imaging screen overnight and imaged using a Typhoon scanner.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Original data underlying this manuscript can be accessed from the Stowers Original Data Repository at http://www.stowers.org/research/publications/libpb-1538. The ATAC-seq, ChIP-seq and RNA-seq data can be found at GEO accession number GSE153052. Source data are provided with this paper.

Code availability

All of the code used for the analysis can be accessed from the Stowers Original Data Repository at http://www.stowers.org/research/publications/libpb-1538.

References

52. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.* **109**, 21.29.1–21.29.9 (2015).
53. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
54. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
55. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.: Ser. B (Methodol.)* **57**, 289–300 (1995).
56. Ou, J. et al. ATACseqQC: a Bioconductor package for post-alignment quality assessment of ATAC-seq data. *BMC Genomics* **19**, 169 (2018).
57. Gu, Z., Eils, R., Schlesner, M. & Ishaque, N. EnrichedHeatmap: an R/Bioconductor package for comprehensive visualization of genomic signal associations. *BMC Genomics* **19**, 234 (2018).
58. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
59. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
60. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
61. Mootha, V. K. et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34**, 267–273 (2003).

62. Van der Auwera, G. A. et al. From FastQ data to high-confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinf.* **43**, 11.10.1–11.10.33 (2013).
63. Kulakovskiy, I. V. et al. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.* **46**, D252–D259 (2018).

Acknowledgements

We are grateful to the cavefish and aquatics core facilities at the Stowers Institute for Medical Research for the support and husbandry of the cavefish and zebrafish. DNA samples for wild-caught Tinaja, Yerbaniz, Piedras and Japonés cavefish were generously provided by R. Borowsky, and B. Jeffery provided *Astyanax* liver samples for preliminary ChIP experiments. We thank M. Cook for assistance with the motif analysis software, K. Weaver for assistance with high-throughput genotyping, and M. Miller for illustrations. We thank R. Krumlauf and J. Zeitlinger for useful input throughout the study and critical reading of the manuscript. N.R. is supported by institutional funding, National Institutes of Health (NIH) grants 1DP2AG071466-01 and R01 GM127872, and the National Science Foundation (NSF) EDGE award 1923372. R.P. was supported by a grant (no. PE 2807/1-1) from Deutsche Forschungsgemeinschaft.

Author contributions

J.K. and N.R. designed the study. J.K. performed experiments with critical input from N.P.S. and J.W.C., and support from S.X. R.P. and A.K. collected wild *Astyanax* samples. J.K., C.W.S., N.Z. and J.V. performed the analyses with support from H.L. J.K. and N.R. wrote the manuscript. All authors read and approved of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-022-01049-4.

Correspondence and requests for materials should be addressed to Nicolas Rohner.

Peer review information *Nature Genetics* thanks Paul Wade and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

nature research

Corresponding author(s): Nicolas Rohner

Last updated by author(s): Nov 9, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	For ChIP-seq, libraries were prepared using the KAPA HTP Library Prep Kit for Illumina and Bioo Scientific NEXTflex DNA barcodes. The Data analysis resulting libraries were purified using the Agencourt AM Pure XP system (Beckman Coulter) then quantified using a Bioanalyzer (Agilent Technologies) and a Qubit fluorometer (Life Technologies). Post amplification size selection was performed on all libraries using a Pippin Prep (Sage Science). Libraries were re-quantified, normalized, pooled and sequenced on Illumina HiSeq 2500 instrument as 50-bp single read. Following sequencing, Illumina Real Time Analysis version 1.18.64 and bcl2fastq2 v2.18.0.12 were run to demultiplex reads and generate FASTQ files. For RNA-seq, Libraries were prepared according to manufacturer's instructions using the TruSeq Stranded mRNA Prep Kit (Illumina). The resulting libraries were purified using the Agencourt AM Pure XP system (Beckman Coulter) then quantified using a Bioanalyzer (Agilent Technologies) and a Qubit fluorometer (Life Technologies). Libraries were re-quantified, normalized, pooled and sequenced on an Illumina HiSeq 2500 instrument as 50-bp single read. Following sequencing, Illumina Real Time Analysis version 1.18.64 and bcl2fastq2 v2.20 were run to demultiplex reads and generate FASTQ files. For ATAC-seq, the library was prepared as per the original protocol (Buenrostro et al., 2015). Paired-end sequencing was performed using the Illumina Next-seq Mid-output mode. UCSC multiz100way alignments download: UCSC utilities (mafSpeciesSubset, 2016-08-23),
Data analysis	The surface fish genome, publicly available on Ensembl since 2017 (astMex_2.0), was used as the reference genome for all analysis in the study unless otherwise mentioned. Approximately 40 million reads were produced for each sample. For analysis of chromatin marks, FASTQ files were aligned to astMex_2.0 using bowtie2 under default parameters. Reads mapping to the mitochondrial chromosome were filtered out using samtools prior to calling peaks with MACS2 version 2.1.2 under default parameters. ATAC-seq data was aligned to astMex_2.0 using bowtie2 version 2.3.4.1 with the following parameters: -X 2000-very-sensitive. More than 80% of the reads successfully mapped to the reference genome (Supplementary Information S1-6). Duplication rate was assessed at less than 0.4% per sample using Picard MarkDuplicates. Fragment size distribution was examined with ATACseqQC (Ou et al., 2018) from Bioconductor, indicating mean fragment length of 200 bases, and

decreasing in frequency with a periodicity of 200 bp. Peaks were called with MACS2 using default parameters. Supplementary Information S1-6 gives the number of peaks obtained for each fish type. For each chromatin mark and the ATAC data set, a reference set of peaks was created for each type of fish
by reducing peaks from multiple replicates to a single peak set, and only keeping peaks observed in at least two replicates. The ATAC peak reference loci were quantified across all samples by tallying read overlaps in R and analyzed for differential accessibility between fish using edgeR (Robinson et al., 2010) to calculate logFC ratios and p-values for each locus. ATAC Peak loci with average CPM values less than 1 were not considered for further analysis. P-values were adjusted by Benjamini-Hochberg (Benjamini and Hochberg, 1995), and those with a value less than 0.0001 were considered differentially open between the 2 morphotypes. ATAC-seq loci were then scored for overlap (1 or 0) with each chromatin mark and mapped to the nearest gene using the GenomicRanges library in R (version R 3.6.1). For analysis of ChromHMM, pathway and motif enrichment we used CR Es within 10kb of the nearest TSS.
Programs use: Heatmaps: R (3.6.1), EnrichedHeatmap (1.16.0) Chromatin states: ChromHMM (1.15), bedtools (2.26.0), STAR (STAR_2.6.1c), bwa version: 0.7.17-r1188, GATK (GenomeAnalysisTK-3.8-0).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Original data underlying this manuscript can be accessed from the Stowers Original Data Repository at http://www.stowers.org/research/publications/libpb-1538. The ATAC-seq, ChIP-seq and RNA-seq data can be found at GEO accession number GSE153052. Databases used in the study includes Ensembl.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes for Next-generation sequencing experiments were determined based on published guidelines from EnCODE consortium (PMID: 22955991).
Data exclusions	No data was excluded.
Replication	All the experiments were performed with minimum two replicates. qPCR was performed on 5 F1 fishes. All replicates in the study were successful.
Randomization	Randomization was not needed in this study as there was only one treatment group (Normal fed fishes). The individual fish used in the study were not pre-determined except for choosing relatively large sized fish to obtain enough tissue for ChIP-seq and ATAC-seq experiments.
Blinding	The groups included Surface, Pachon and Tinaja morphotypes of Astyanax. As the study was based on a comparative analysis (cavefish with respect to surface fish), the analysts were not blinded to this fact.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☐	☒ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used: H3K4me3 (Millipore, # 07-473), H3K27ac (Abcam, # ab4729), H3K27me3 (Abcam, # ab195477). 1 ug antibody was used per sample.

Validation: All experiments used antibodies for epitopes that are highly conserved in all eukaryotes. Antibodies were sourced from commercial sources, validated using western blotting (See supplementary figure 6).

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s): HepG2 (ATCC), ZFL (ATCC)

Authentication: The cell lines were not authenticated.

Mycoplasma contamination: All cell lines were tested negative for mycoplasma.

Commonly misidentified lines (See ICLAC register): No commonly misidentified cell line was used.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals: Larvae of Zebrafish (Danio rerio) and, larvae and adults of Cavefish (Astyanax mexicanus) were used in the study. For ATAC-seq, a mix of male and females fishes were used. All the fishes used were in the age group of 9-12 months old.

Wild animals: Wild-caught animals reported in the study have been described in PMID: 32690906 and PMID: 32017448.

Field-collected samples: No field collected samples were used in this study.

Ethics oversight: Animal husbandry was according to IACUC approved protocol 2019-078 (zebrafish) and 2019-084 (cavefish) approved for Stowers Institute for
Medical Research.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

ChIP-seq

Data deposition

☒ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☒ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links
May remain private before publication.

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE153052

Files in database submission

GSE153052 Genome wide analysis of cis-regulatory elements reveals gene regulatory networks in metabolic adaptation of cavefish FASTQ

ATAC-Seq Files:
GSE153048 Genome-wide maps of open chromatin in liver cells of Astyanax mexicanus [ATAC-seq]
GEO_ID Sample_Name File_Type
GSM4633540 Pachon_1 [ATAC-seq] FASTQ
GSM4633542 Pachon_2 [ATAC-seq] FASTQ
GSM4633544 Surface_1 [ATAC-seq] FASTQ

GSM4633546 Surface_2 [ATAC-seq] FASTQ
GSM4633548 Tinaja_1 [ATAC-seq] FASTQ
GSM4633549 Tinaja_2 [ATAC-seq] FASTQ
BED files of ATAC peaks for Pachon, Surface, Tinaja (3 total).

GSE153049 Genome-wide maps of chromatin state in liver cells of Astyanax mexicanus [ChIP-seq]
ChIP-Seq Files:
GEO_ID Sample_Name File_Type
GSM4633520 Pachon_H3K27ac_1ug_ip_1889 FASTQ
GSM4633521 Pachon_H3K4me3_1ug_ip_1889 FASTQ
GSM4633522 Pachon_tc_1889 FASTQ
GSM4633523 Surface_H3K27ac_ip_1930_1 FASTQ
GSM4633524 Surface_H3K27ac_ip_1930_2 FASTQ
GSM4633525 Surface_H3K4me3_ip_1930_1 FASTQ
GSM4633526 Surface_H3K4me3_ip_1930_2 FASTQ
GSM4633527 Surface_tc_1930 FASTQ
GSM4633528 Pachon_H3K27ac_1ug_ip_1934_b FASTQ
GSM4633529 Pachon_H3K4me3_1ug_ip_1934_b FASTQ
GSM4633530 Tinaja_H3K27ac_ip_1934_b FASTQ
GSM4633531 Tinaja_H3K4me3_ip_1934_a FASTQ
GSM4633532 Tinaja_tc_1934 FASTQ
GSM4633533 Tinaja_H3K27ac_ip_2005_A FASTQ
GSM4633534 Tinaja_H3K27ac_ip_2005_B FASTQ
GSM4633535 Tinaja_H3K4me3_ip_2005_A FASTQ
GSM4633536 Tinaja_H3K4me3_ip_2005_B FASTQ
GSM4633537 Pachon_H3K27me3_ip_2399_a FASTQ
GSM4633538 Pachon_H3K27me3_ip_2399_b FASTQ
GSM4633539 Pachon_tc_2399 FASTQ
GSM4633541 Surface_H3K27me3_ip_2399_a FASTQ
GSM4633543 Surface_H3K27me3_ip_2399_b FASTQ
GSM4633545 Surface_tc_2399 FASTQ
GSM4633547 Tinaja_H3K27me3_ip_2399_a FASTQ
GSM4633550 Tinaja_H3K27me3_ip_2399_b FASTQ
GSM4633551 Tinaja_tc_2399 FASTQ
BED files of H3K27ac, H3K4me3, H3K27me3 peaks for Pachon, Surface, Tinaja (9 total).

GSE153050 RNA-Seq Analysis of liver cells of Astyanax mexicanus. Jun 01, 2021
RNA-Seq Files:
GEO_ID Sample_Name File_Type
GSM4633552 Pachon_1 [RNA-seq] FASTQ
GSM4633553 Pachon_2 [RNA-seq] FASTQ
GSM4633554 Surface_1 [RNA-seq]FASTQ
GSM4633555 Surface_2 [RNA-seq]FASTQ
GSM4633556 Tinaja_1 [RNA-seq] FASTQ
GSM4633557 Tinaja_2 [RNA-seq] FASTQ
Table of Read Counts on Genes in AstMex2 for Ensembl 96 (1 tsv file)

Genome browser session
(e.g. UCSC)

https://genome.ucsc.edu/s/jak/AstMex2_All_data_Dec%2719

Methodology

Replicates

All ChIP-seq, RNA-seq and ATAC-seq were done in two or more biological replicates.

Sequencing depth

All sequencing for ChIP-seq and RNA-seq were done as 51bp single end reads using Illumina HiSeq 2500. Sequencing were done to achieve 30 million reads per biological replicate.

Paired-end sequencing was performed using the Illumina Next-seq Mid-output mode.

Antibodies

H3K4me3 (Millipore, # 07-473), H3K27ac (Abcam, # ab4729), H3K27me3 (Abcam, # ab195477).

Peak calling parameters

Following sequencing, Illumina Real Time Analysis version 1.18.64 and bcl2fastq2 v2.18.0.12 were run to demultiplex reads and generate FASTQ files. FASTQ files for each chromatin mark were aligned to astMex_2.0 using bowtie2 under default parameters. Reads mapping to the mitochondrial chromosome were subsequently filtered out using samtools prior to calling peaks with MACS2 under default parameters. ATAC-seq data was aligned to astMex_2.0 with bowtie2 with the following parameters: -X 2000 –very-sensitive. Peaks were called with MACS2 using default parameters. For each chromatin mark and ATAC peak, a reference set of peaks was created for each type of fish by reducing peaks from each replicate to a single peak set, and then keeping only peaks observed in at least two replicates. These loci were quantified and analyzed for differential openness between fish using edgeR to calculate logFC ratios and p-values. ATAC-seq loci were then scored for overlap (1 or 0) with each chromatin mark and mapped to the nearest gene using the GenomicRanges library in R. Peaks were filtered for logCPM values greater than 0. For differential peaks, the p-value for the difference in the peak intensities between 2 morphotypes was <0.0001.

Data quality

For each chromatin mark and ATAC peak, a reference set of peaks was created for each type of fish by reducing peaks from each

Data quality

replicate to a single peak set, and then keeping only peaks observed in at least two replicates. These loci were quantified and analyzed for differential openness between fish using edgeR to calculate logFC ratios and p-values.

Software

bowtie2 version 2.3.4.1
MACS version 2.1.2
STAR STAR_2.6.1c
R 3.6.1
Bioconductor and the GenomicRanges library
bwa version: 0.7.17-r1188
gatk version: GenomeAnalysisTK-3.8-0
Gene Set Enrichment Analysis (Broad Institute)
Heatmaps: R (3.6.1)
EnrichedHeatmap (1.16.0)
Chromatin states: ChromHMM (1.15)
bedtools (2.26.0)