doi:10.1038/nature26136

We reexamined the available whole genome data from different cave and surface populations (McGaugh et al, unpublished) to investigate whether *insra* exhibited any indication that it has experienced exceptional divergence between cavefish and surface fish relative to the empirical distribution of genes in the genome.

For population genomic measures, we included a core set of samples sequenced with Illumina HiSeq 100bp paired-end reads. Our sampling included following individuals: Pachón, N = 10 (9 newly resequenced + the reference reads mapped back to the reference genome); Tinaja N =10; Molino N = 9; Rascon N = 6; and Río Choy N = 9 and required six or more individuals have data for a particular site.

Alignments of Illumina data to the reference genome were created with the BWA-mem algorithm [1] in *bwa-0.7.1* [2,3]. We utilized both Genome Analysis Toolkit v3.3.0 (GATK). The Genome Analysis Toolkit v.3.3.0 (GATK) and Picard v1.83 (http://broadinstitute.github.io/picard/) were used for downstream manipulation of alignments, according to GATK Best Practices[4-6]. For all population genomic measures, we excluded masked repetitive elements, indels (if present in any of the core set of samples), and 10bp surrounding the bases affected by each indel by using the masking_coordinates.gz file available for the *Astyanax mexicanus* genome v1.0.2 though NCBI genomes FTP. Alignments of paired-end and orphaned reads for each individual were sorted and merged using Picard. Duplicate reads that may have arisen during PCR were marked using Picard's MarkDuplicates tool, and filtered out of downstream analyses resulting in a mean coverage of 9.28 fold across samples. GATK's IndelRealigner (IR) and RealignerTargetCreator (RTC) tools were then used to realign reads that may have been errantly mapped around indels (insertions/deletions). IR was first used to realign around indels in intervals determined by RTC for each individual, and then in intervals determined by RTC for each population. This two-step process ensured that realignment occurred around a majority of the possible indels. RTC was run on each individual so that rare indels can trigger realignment, as they may be overlooked when input with other individuals that do not contain evidence of an indel at that location. Using all samples in a population added power, and may allow for realignment around indels in areas that were too low in coverage to trigger the need for realignment when using each individual alone.

The HaplotypeCaller tool in GATK v3.5.0 was used to create GVCFs of genotype likelihoods for each individual. The tool GenotypeGVCFs was used to create a combined variant call format (VCF) of raw variant calls for all samples. Hard filters were applied separately to SNPs and indels/mixed sites using the

VariantFiltration and SelectVariants tools. Filtering variants was performed to remove low confidence calls from the dataset.

Measures of divergence for each cave population compared to its most closely related-surface population. We also included Río Choy surface-Pachón cave and Río Choy surface –Tinaja cave as there is some evidence that this surface population interbreeds with these two caves. We used VCFtools v0.1.13[7] to calculate π, Tajima's D, F_{ST} and d_{XY} and custom python scripts to calculate these metrics on a per gene basis. d_{XY}[8 9] is an absolute measure of divergence that is independent of diversity levels, F_{ST}[10] is one of the most common ways to measure population differentiation, but is strongly influenced by within-population diversity. hapFLK[11] is a measure of haplotype divergence that takes into account hierarchical population structure. hapFLK is among the most sensitive measures for detecting positive selection[12]. Measures of diversity include: H-SCAN[13], which is measured here as the actual physical distance in number base pairs between the two delimiting SNPs that terminate a pairwise homozygosity tract at each end and π, which is pairwise nucleotide diversity within populations. Tajima's D is the measure of π (observed) – π (expected) based on the number of variable sites in the data and is a measure of deviation from neutrality. Negative values may be caused by selective sweeps or recent population expansion, whereas positive values may be caused by balancing selection or a recent bottleneck[14].

We found that *insra* is not exceptional in its divergence between cavefish and surface fish using multiple independent statistical methods (FST, DXY, of hapFLK results; Extended Data Table 1). In addition, we did not observe extreme reductions in diversity or exceptionally long tracts of homozygosity which would be indicative of recent hard selective sweeps (π, HSCAN; Supplementary Table 1). Thus, this gene is not exceptional for these metrics when compared to all other genes in the genome with available data.

Supplementary Table 1

Total sites refers to total sites with data for six or more individuals per population. Fixed = fixed differences between populations, invariant = invariant sites, percentile in the genome is relative to all other genes in the genome with data available. Data for hapFLK refers to the mean, median, maximum, minimum and standard deviation of p-values for the gene (total of N = 54 sites with p-values in the coding region of the gene). % in the genome is the percentile relative to all other genes in the genome with available data.

A. Measures of divergence

D_{XY}	Total sites	Mean	Stdev	Fixed	Invariant	% in genome
Rascon-Tinaja	4033	0.001073	0.023402	0	4016	26.51
Rascon-Pachón	4033	0.001021	0.025561	0	4019	22.71
Río Choy-Pachón	4046	0.002176	0.035289	1	4018	44.97
Río Choy-Tinaja	4046	0.002175	0.032547	0	4014	42.15

F_{ST}	Total sites	Mean	Stdev	% in genome
Rascon-Tinaja	12	0.207804	0.291358	38.84
Rascon- Pachón	12	0.269965	0.356356	45.21
Río Choy-Pachón	27	0.245772	0.310325	52.82
Río Choy-Tinaja	25	0.234978	0.271977	52.25

hapFLK	Median	Max	Min	Stdev	# of p.values<0.05
hapFLK p-values	0.328636	0.788896	0.086054	0.258214	0

B. Measures of Diversity

HSCAN	Mean_HSCAN	Percentile in genome (smaller = shorter homozygosity runs)
Tinaja	1617.131611	33.97
Pachón	10877.28727	48.51
Rascon	363.8201	34.83
Río Choy	328.45057	47.32

π	Total Sites	Mean	Stdev	% in genome
Tinaja	4047	0.00078	0.01431	53.43
Pachón	4047	0.00031	0.00719	27.22
Rascon	4033	0.00055	0.01149	14.38
Río Choy	4046	0.00182	0.02584	52.33

Tajima'sD	# of SNPS	Tajima'sD	% in genome
Tinaja	15	-0.924334	29.89
Pachón	8	-1.50598	15.32
Rascon	11	-1.40961	11.62
Río Choy	25	0.159596	76.84

1 Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997* (2013).

2 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754-1754 (2009).

3 Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-589 (2010).

4 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303 (2010).

5 DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43**, 491-498 (2011).

6 Auwera, G. A. *et al.* From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current protocols in bioinformatics*, 11.10. 11-11.10. 33 (2013).

7 Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158 (2011).

8 Nei, M. *Molecular Evolutionary Genetics*. (Columbia University Press, 1987).

9 Nei, M. *Mathematical model for studying genetic variation in terms of restriction endonucleases.* . Vol. 76 (Proceedings of the National Academy of Sciences, 1979).

10 Charlesworth, B. *Measures of divergence between populations and the effect of forces that reduce variability*. Vol. 15 538-543 (1988).

11 Fariello, M. I., Boitard, S., Naya, H., SanCristobal, M., and Servin, B. in *Genetics* Vol. 193 Ch. 929-941, (2013).

12 Schlamp, F., et al. in *Mol. Ecol.* Vol. 25 342-356 (2016).

13 Messer, P. W. *H-scan: Detecting hard and soft sweeps in population genomic data*, <https://messerlab.org/resources/> (2015).

14 Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585-595 (1989).