Supplement for:

Expression analyses of cave mollies (*Poecilia mexicana*) reveal key genes involved in the early evolution of eye regression

Kerry L. McGowan[1*], Courtney N. Passow[2,3], Lenin Arias-Rodriguez[4], Michael Tobler[3], Joanna L. Kelley[1]

[1]School of Biological Sciences, Washington State University, Pullman, WA, USA

[2]Department of Ecology, Evolution and Behavior, University of Minnesota – Twin Cities, St. Paul, MN, USA

[3] Division of Biology, Kansas State University, Manhattan, KS, USA

[4] División Académica de Ciencias Biológicas, Universidad Juárez Autónoma de Tabasco, Villahermosa, Tabasco, México

[*]Corresponding author: Kerry L. McGowan, School of Biological Sciences, Washington State University, Pullman, WA, USA

Email: kerry.mcgowan@wsu.edu

Supplementary Methods

Sample collection for relative eye size of cave mollies compared to surface fish

Poecilia mexicana adults and juveniles were collected using seines and dip nets from a non-sulfidic surface site (Arroyo Bonita), a sulfidic surface site (El Azufre), a non-sulfidic cave (Cueva Luna Azufre), and a sulfidic cave (Cueva del Azufre) in the Río Tacotalpa drainage near Tapijulapa, Tabasco, Mexico (figure S1). After measurements were taken, fish were released at the original collection site.

Sample collection for RNA-sequencing

Poecilia mexicana individuals were collected from the same four sites as above (see also figure S1). See table S1 for sampling details including site coordinates, average standard length, and average mass of sampled individuals. Individuals ($N = 4$ per site) were captured using seine nets. The individuals were also sampled for brain, gill, and liver tissues (1), but eye tissue data were not analyzed or published until this study. Cave animals were brought to the cave entrance in a closed dark container and quickly sacrificed to minimize light exposure. Whole eyes were immediately extracted using sterilized scissors and forceps and stored in RNAlater (Invitrogen). Procedures for all experiments were approved by the Institutional Animal Care and Use Committee at Kansas State University (Protocol #3418).

RNA extraction

For each sample, both eyes were inserted into a Covaris TT1 tissue TUBE, flash frozen in liquid nitrogen, then pulverized using a Covaris cryoPREP on setting 3. Total RNA was extracted using the Qiagen RNeasy Plus Mini Kit and quantified with the Thermo Fisher Scientific Qubit RNA

Assay Kit and Agilent RNA 6000 Nano Total RNA Kit on an Agilent 2100 Bioanalyzer. Ribo-zero depletion was accomplished with an Epicentre Ribo-Zero Magnetic Gold Kit (human/mouse/rat). The remaining mRNA was cleaned twice with Beckman Coulter Life Sciences Agencourt RNAClean XP beads. RNA was eluted and fragmented to 400 base pairs (bp) in length using the New England BioLabs (NEB) RNA fragmentation buffer for 4 min at 94°C.

cDNA library preparation and sequencing

For first-strand cDNA synthesis, fragmented mRNA was mixed with 1 µL random hexamers: oligo-dT primers (3 µg:1 µg), 4 µL of Invitrogen 5x first-strand reaction buffer, 2 µL of Invitrogen 0.1 M DTT, and 1 µL of NEB 10 mM dNTP mix. 1 µL Invitrogen SuperScript III Reverse Transcriptase was then added followed by incubation at 25°C for 50 min. Samples were then immediately placed on ice.

For second-strand cDNA synthesis, 5 µL Invitrogen 5x first-strand buffer, 1 µL Applied Biosystems DTT, 2 µL 10 mM dNTP mix with dUTP, 15 µL Invitrogen 5x second-strand reaction buffer, and 3.75 µL NEB second-strand enzyme mix were added to each sample. Samples were incubated for 2 hr at 16°C.

The resulting double-stranded cDNA libraries were cleaned using Beckman Coulter Life Sciences Agencourt RNAclean XP beads and eluted into 50 µL nuclease-free H_2O. KAPA Biosystems KAPA HTP Library Preparation Kit was used for end-repair, A-tailing, adapter ligation using Illumina TruSeq barcoded adapters, and library amplification. For library amplification, samples were first denatured for 45 seconds (s) at 98°C, then 12 cycles: 15 s at 98°C, 30 s at 60°C, and 30 s at 72°C, followed by a final extension for 60 s at 72°C. Thermo

Fisher Scientific Qubit dsDNA HS Assay Kit and Agilent High Sensitivity DNA Analysis Kit on the 2100 Bioanalyzer were used to quantify the cDNA libraries. Libraries were then pooled based on nanomolar concentrations quantified by the 2100 Bioanalyzer.

Paired-end sequencing was performed using an Illumina HiSeq 2000 with 101 bp reads. Some samples were rerun on another Illumina HiSeq 2000 lane due to low coverage from the initial run. Reads were concatenated for each sample for all downstream analyses. Raw read quality was assessed with FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Raw reads were adapter and quality trimmed using Trim Galore! (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/).

Identifying differentially expressed genes in the eyes of cave mollies compared to surface fish

A gene counts matrix was produced using StringTie and the associated Python script (prepDE.py). Analyses were performed in R. Genes with zero counts across all samples were removed from the gene counts matrix. Normalization factors were used to control for effective library sizes in downstream analyses. Common, trended, and tagwise dispersions were estimated using the Cox-Reid profile-adjusted likelihood method to control for biological variability across samples (2). Dispersions were then used to fit a negative binomial model based on the assumption that counts followed a gamma distribution.

Functional annotation of gene expression based on environment

A total of 1,000 permutations were run to calculate nominal *p*-values for the ranked list, followed by calculation of the false discovery rate (retained FDR < 0.05) to account for the number of gene ontologies (GO) in the GO Biological Process dataset (c5.bp.v6.2.symbols.gmt).

Identifying co-expressed gene modules correlated with cave adaptation

A soft thresholding power adjacency function ($\beta = 4$, corresponding to an R^2 of 0.823) was used to calculate an adjacency matrix. Genes were clustered into modules based on topological overlap matrix-based dissimilarity and the resulting dendrogram was split using a minimum module size of 20 genes. Modules with similar expression profiles were merged together. Modules were then related to habitat characteristics (cave vs. surface, sulfidic vs. non-sulfidic) using Pearson correlations.

Table S1. Population data. Location and size data ± standard error for populations of *P. mexicana* [sampled in (1)]. All individuals were female. N = 4 individuals per site.

Population	Site	GPS Coordinates	Average Standard Length (mm)	Average Mass (g)
Surface, non-sulfidic	Arroyo Bonita	17.427, -92.752	45.25 ± 5.50	1.93 ± 0.50
Surface, sulfidic	El Azufre II	17.439, -92.775	32.75 ± 0.96	0.92 ± 0.11
Cave, non-sulfidic	Cueva Luna Azufre[1]	17.441, -92.773	39.00 ± 5.35	1.71 ± 0.64
Cave, sulfidic	Cueva del Azufre[2]	17.442, -92.775	43.25 ± 4.27	1.97 ± 0.37

[1]Small, un-numbered pool in the cave chamber.
[2]Chamber V in the Cueva del Azufre system (3).

Table S2. Read counts. Reads ± standard error pre-trimming (raw reads) and post-trimming.

Population	Average Number of Reads	
	Pre-trimming	**Post-trimming**
Surface, non-sulfidic	8,783,666 ± 849,961	7,844,055 ± 675,623
Surface, sulfidic	9,172,022 ± 818,036	8,313,186 ± 684,809
Cave, non-sulfidic	9,849,840 ± 279,017	8,984,448 ± 287,894
Cave, sulfidic	10,665,563 ± 2,553,502	9,833,530 ± 2,367,666

Table S3. Mapping statistics. Reads counts and percentages ± standard error (SE) that mapped to the *P. mexicana* reference genome.

Population	Average Counts of Reads Mapped	Average % of Reads Mapped
Surface, non-sulfidic	2,565,909 ± 806,593	31.67 ± 7.72
Surface, sulfidic	2,631,180 ± 913,609	31.06 ± 9.26
Cave, non-sulfidic	2,205,920 ± 638,596	25.30 ± 7.98
Cave, sulfidic	2,199,701 ± 783,600	20.56 ± 3.16

Table S4. Sulfidic versus non-sulfidic. Significantly differentially expressed genes in eye tissues from sulfidic populations of *P. mexicana* compared to non-sulfidic populations.

Gene ID	Protein/RNA	Full Protein Name	Function	logFC	FDR
STRG.13368* No BLAST hit	16S rRNA	16S ribosomal RNA	Component of 30S subunit, prokaryotic	-3.04	3.65×10^{-4}
106910732	LSU rRNA	Large subunit ribosomal RNA	Component of 60S subunit	-1.52	4.60×10^{-27}

*The "STRG" identifier is from the program StringTie.

Table S5. GSEA results. NAME = Gene ontology; ACCESSION = Accession number; SIZE = Number of genes in the gene set after filtering out genes not expressed in the dataset; ES = Enrichment score; NES = Normalized enrichment score; NOM p-val = Nominal *p*-value; FDR q-val = False discovery rate; FWER p-val = Familywise-error rate; RANK AT MAX = Position in the ranked list where maximum enrichment score occurred; LEADING EDGE = Statistics describing the leading-edge subset of the gene set.

See separate excel spreadsheet.

Table S6. WGCNA results.

See separate excel spreadsheet.

Table S7. Relative eye size data. H_2S = -1 if nonsulfidic (NS), 1 if sulfidic (S). Light = 1 if present (Surface), -1 if absent (Cave). Sex = 0 if female, 1 if male.

See separate excel spreadsheet.

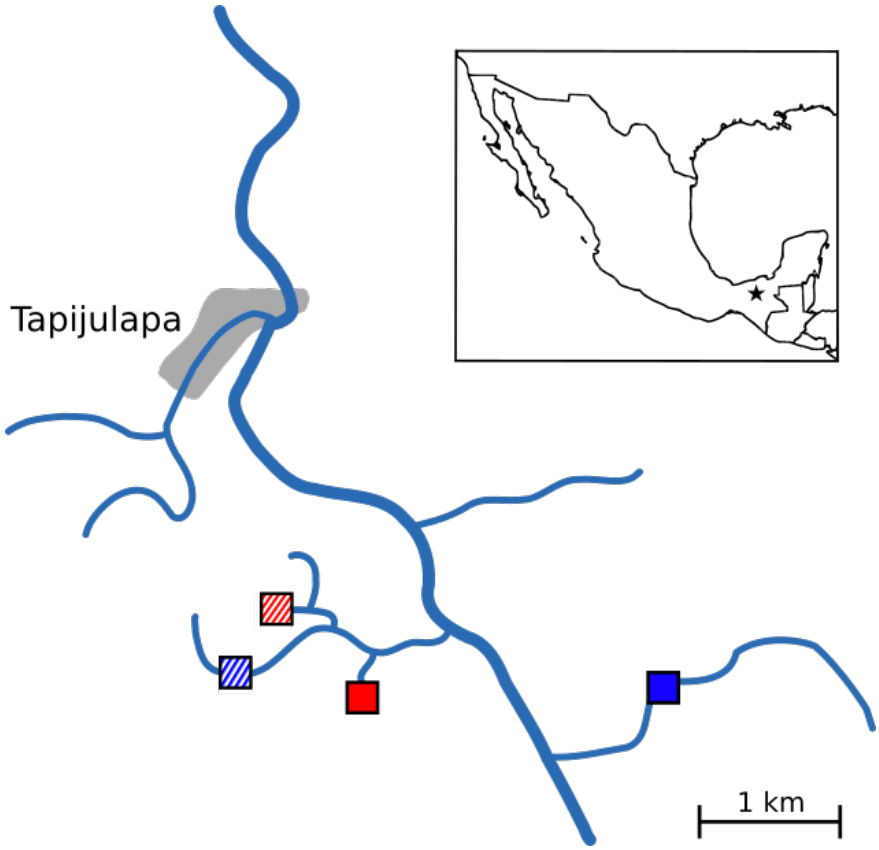

Figure S1. Geography of the Tacotalpa drainage near Tapijulapa, Tabasco, Mexico (star). Blue = surface populations, red = cave. Filled squares = non-sulfidic populations, hashed squares = sulfidic. See Table S1 for GPS coordinates.

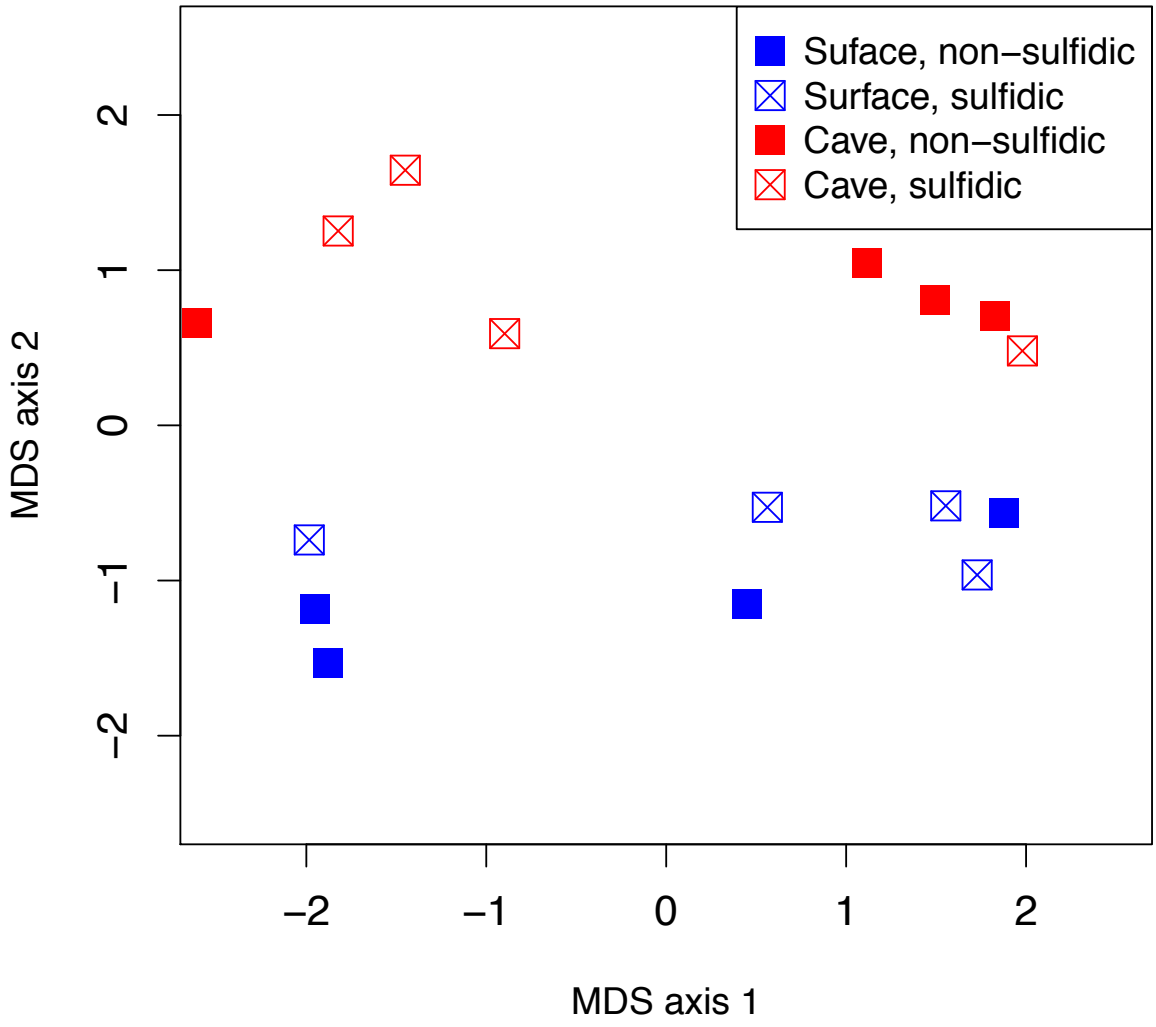

Figure S2. Samples separate by library size and environment. Multi-dimensional scaling (MDS) plot of the top 500 differentially expressed genes in *P. mexicana* eye tissues. MDS axis 1 separated samples by library size. MDS axis 2 separated samples by cave versus surface environment.

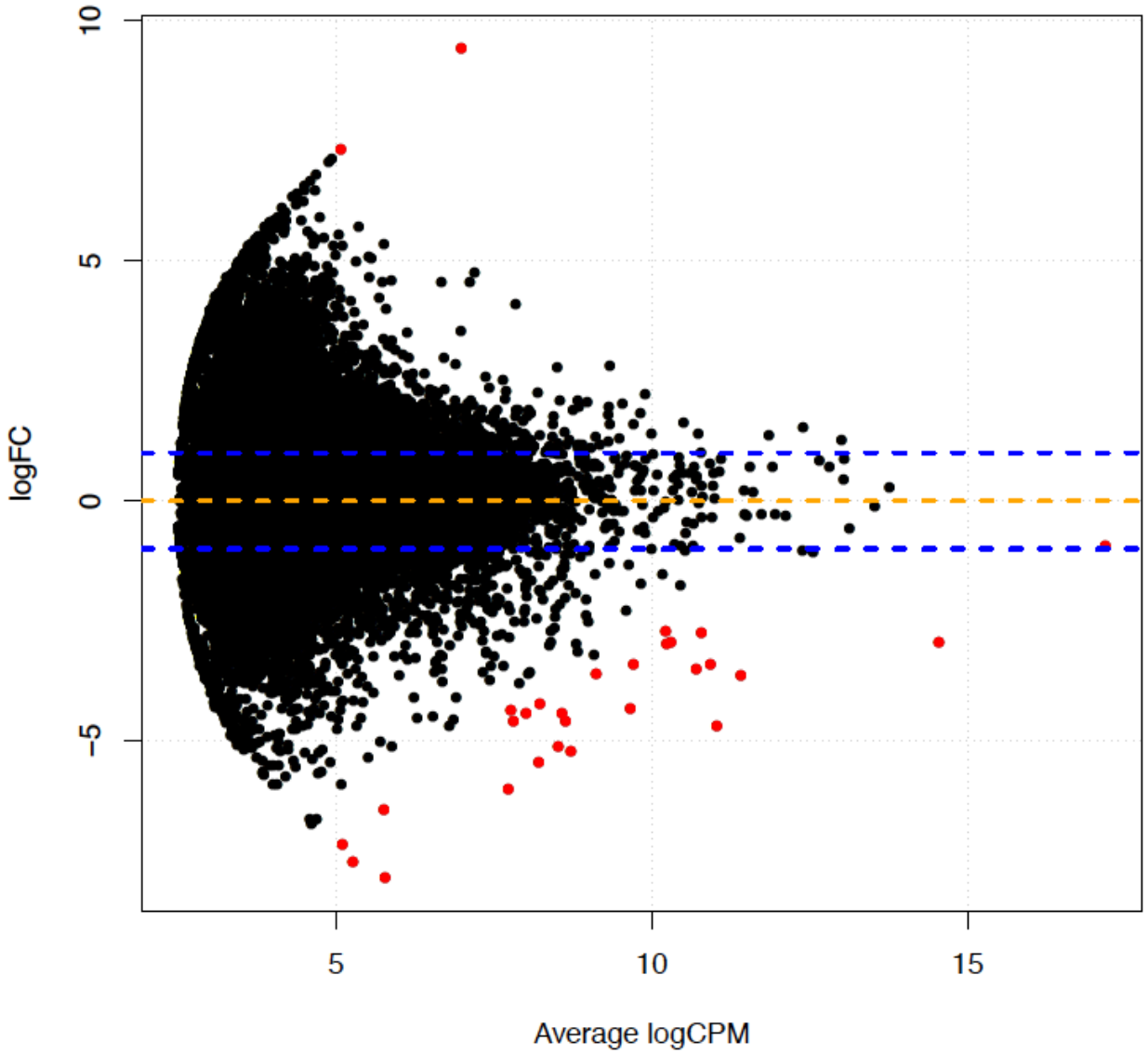

Figure S3. Differentially expressed genes. Log_2-fold change (logFC) in gene expression plotted against the average log of read counts per million (logCPM). No logFC (horizontal orange line) and logFC of -1 and 1 (blue lines) are indicated. Genes that were significantly differentially expressed in cave populations of *P. mexicana* compared to surface populations (FDR < 0.05) are indicated by red points. Twenty-seven genes were downregulated in cave compared to surface populations, whereas two were upregulated.

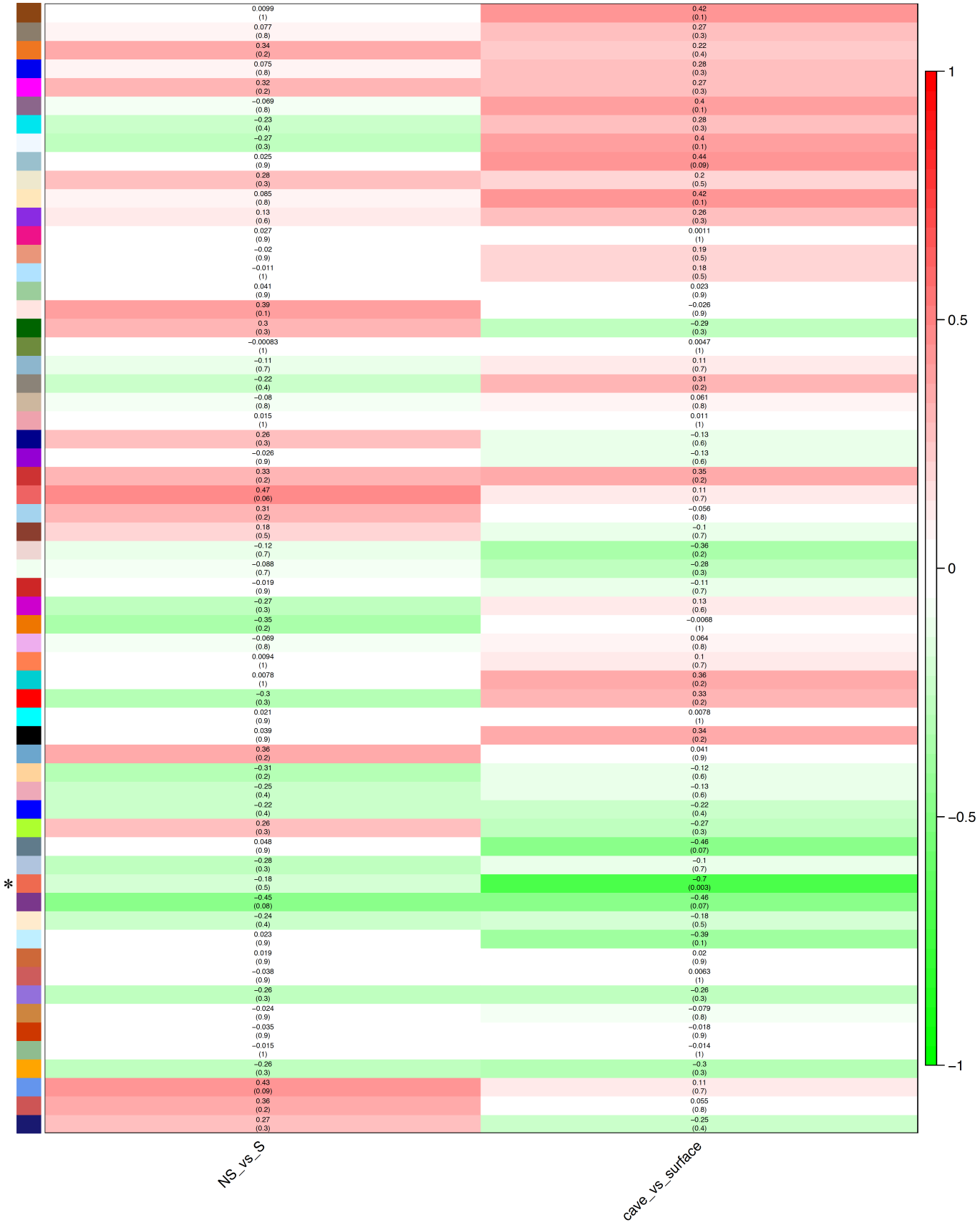
0.0099 (1) 0.42 (0.1)
0.077 (0.8) 0.27 (0.3)
0.34 (0.2) 0.22 (0.4)
0.075 (0.8) 0.28 (0.3)
0.32 (0.2) 0.27 (0.3)
−0.069 (0.8) 0.4 (0.1)
−0.23 (0.4) 0.28 (0.3)
−0.27 (0.3) 0.4 (0.1)
0.025 (0.9) 0.44 (0.09)
0.28 (0.3) 0.2 (0.5)
0.085 (0.8) 0.42 (0.1)
0.13 (0.6) 0.26 (0.3)
0.027 (0.9) 0.0011 (1)
−0.02 (0.9) 0.19 (0.5)
−0.011 (1) 0.18 (0.5)
0.041 (0.9) 0.023 (0.9)
0.39 (0.1) −0.026 (0.9)
0.3 (0.3) −0.29 (0.3)
−0.00083 (1) 0.0047 (1)
−0.11 (0.7) 0.11 (0.7)
−0.22 (0.4) 0.31 (0.2)
−0.08 (0.8) 0.061 (0.8)
0.015 (1) 0.011 (1)
0.26 (0.3) −0.13 (0.6)
−0.026 (0.9) −0.13 (0.6)
0.33 (0.2) 0.35 (0.2)
0.47 (0.06) 0.11 (0.7)
0.31 (0.2) −0.056 (0.8)
0.18 (0.5) −0.1 (0.7)
−0.12 (0.7) −0.36 (0.2)
−0.088 (0.7) −0.28 (0.3)
−0.019 (0.9) −0.11 (0.7)
−0.27 (0.3) 0.13 (0.6)
−0.35 (0.2) −0.0068 (1)
−0.069 (0.8) 0.064 (0.8)
0.0094 (1) 0.1 (0.7)
0.0078 (1) 0.36 (0.2)
−0.3 (0.3) 0.33 (0.2)
0.021 (0.9) 0.0078 (1)
0.039 (0.9) 0.34 (0.2)
0.36 (0.2) 0.041 (0.9)
−0.31 (0.2) −0.12 (0.6)
−0.25 (0.4) −0.13 (0.6)
−0.22 (0.4) −0.22 (0.4)
0.26 (0.3) −0.27 (0.3)
0.048 (0.9) −0.46 (0.07)
−0.28 (0.3) −0.1 (0.7)
−0.18 (0.5) −0.7 (0.003)
−0.45 (0.08) −0.46 (0.07)
−0.24 (0.4) −0.18 (0.5)
0.023 (0.9) −0.39 (0.1)
0.019 (0.9) 0.02 (0.9)
−0.038 (0.9) 0.0063 (1)
−0.26 (0.3) −0.26 (0.3)
−0.024 (0.9) −0.079 (0.8)
−0.035 (0.9) −0.018 (0.9)
−0.015 (1) −0.014 (1)
−0.26 (0.3) −0.3 (0.3)
0.43 (0.09) 0.11 (0.7)
0.36 (0.2) 0.055 (0.8)
0.27 (0.3) −0.25 (0.4)
*
NS_vs_S
cave_vs_surface
1
0.5
0
−0.5
−1

Figure S4. Results of weighted gene co-expression network module-trait relationships. Correlation between module eigenvalues and environment. Each row, indicated by a color on the left, represents a module eigengene. Module-trait relationships are reported with Pearson correlations (top of each cell) and their associated *p*-values (bottom). Cell coloration represents the Pearson correlation value according to the scale bar on the right. The * indicates the module with a significant correlation between the habitat type and gene expression (r = -0.7, p = 0.003). NS_vs_S = non-sulfidic versus sulfidic. cave_vs_surface = light environment.

References

1. Passow CN, Brown AP, Arias-Rodriguez L, Yee MC, Sockell A, Schartl M, et al. Complexities of gene expression patterns in natural populations of an extremophile fish (Poecilia mexicana, Poeciliidae). Mol Ecol. 2017;26(16):4211-25.
2. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. 2012;40(10):4288-97.
3. Gordon MS, Rosen DE. A Cavernicolous Form of the Poeciliid Fish Poecilia sphenops from Tabasco, Mexico. American Society of Ichthyologists and Herpetologists. 1962;1962(2):360-8.