Electronic Supplement for

Local ancestry analysis reveals genomic convergence in extremophile fishes

Anthony P. Brown[1], Kerry L. McGowan[1], Enrique J. Schwarzkopf[1], Ryan Greenway[2], Lenin Arias Rodriguez[3], Michael Tobler[2] & Joanna L. Kelley[1]

[1] School of Biological Sciences, Washington State University, 100 Dairy Road, Pullman, WA 99164, USA

[2] Division of Biology, Kansas State University, 116 Ackert Hall, Manhattan, KS 66506, USA

[3] División Académica de Ciencias Biológicas, Universidad Juárez Autónoma de Tabasco (UJAT), C.P. 86150, Villahermosa, Tabasco, México

Methods for identifying candidate genes

Data source and sample information

Raw RNA-seq reads from Kelley *et al.* 2016 [2] (study accession ID: PRJNA290391) were downloaded. These data were from gill samples collected from individuals of the *Poecilia mexicana* species complex from three drainages in the Río Grijalva basin in southern Mexico (Pichucalco drainage: $N = 6$ [Pich NS2] individuals and $N = 6$ [Pich S1] individuals; Puyacatengo: $N = 5$ [Puya NS] and $N = 5$ [Puya S]; and Tacotalpa drainage: $N = 6$ [Taco NS] and $N = 6$ [Taco S]).

Identifying genes that were consistently differentially expressed between sulfidic and non-sulfidic populations

Raw reads were trimmed using default settings in TrimGalore v. 0.4.2. (https://github.com/FelixKrueger/TrimGalore). The *Poecilia mexicana* reference genome (NCBI accession: GCA_001443325.1) was converted from GFF format to GTF format using the *gffread* function in Cufflinks v. 2.2.1 [3]. Splice sites and exons were extracted from the resulting GTF file using the Python scripts hisat2_extract_splice_sites.py and hisat2_extract_exons.py, respectively, from the HISAT2 v. 2.1.0 package [4]. *hisat2-build* was used to index the reference genome with the slice sites (*--ss* option) and exons (*--exon* option) included. Trimmed reads were mapped to the indexed reference genome using HISAT2 v. 2.1.0 [4] including the *--dta* option to tag spliced read alignments. StringTie v. 1.3.3 [5] was run with the *-BeG* options to include reference annotations. The resulting output was used as input for the Python script prepDE.py provided as part of the StringTie package [5]. The resulting output was a gene counts matrix written as a CSV file.

The gene counts matrix was imported into EdgeR v. 3.22.5 [6] and split by drainage (Pichucalco, Puyacatengo, and Tacotalpa). Analyses were then performed for each drainage independently. Genes with zero counts across all samples were removed before creating a DGElist object. Library sizes

were normalized for each sample. A design matrix grouped non-sulfidic and sulfidic samples separately. Common and tagwise dispersions were estimated using Cox-Reid profile-adjusted likelihood. Quasi-likelihood F-tests were performed to determine which genes were significantly differentially expressed (upregulated and downregulated) comparing individuals from sulfidic springs to individuals from non-sulfidic springs.

General results from differential expression analyses

Out of the 26,817 genes that had counts for at least one sample, 150 genes were consistently upregulated in sulfidic fish compared to non-sulfidic fish, while 149 genes were consistently downregulated in sulfidic fish compared to non-sulfidic fish in all three drainages (Pichucalco, Puyacatengo, and Tacotalpa; see Figure S5 and Table S7).

Supplementary Figures

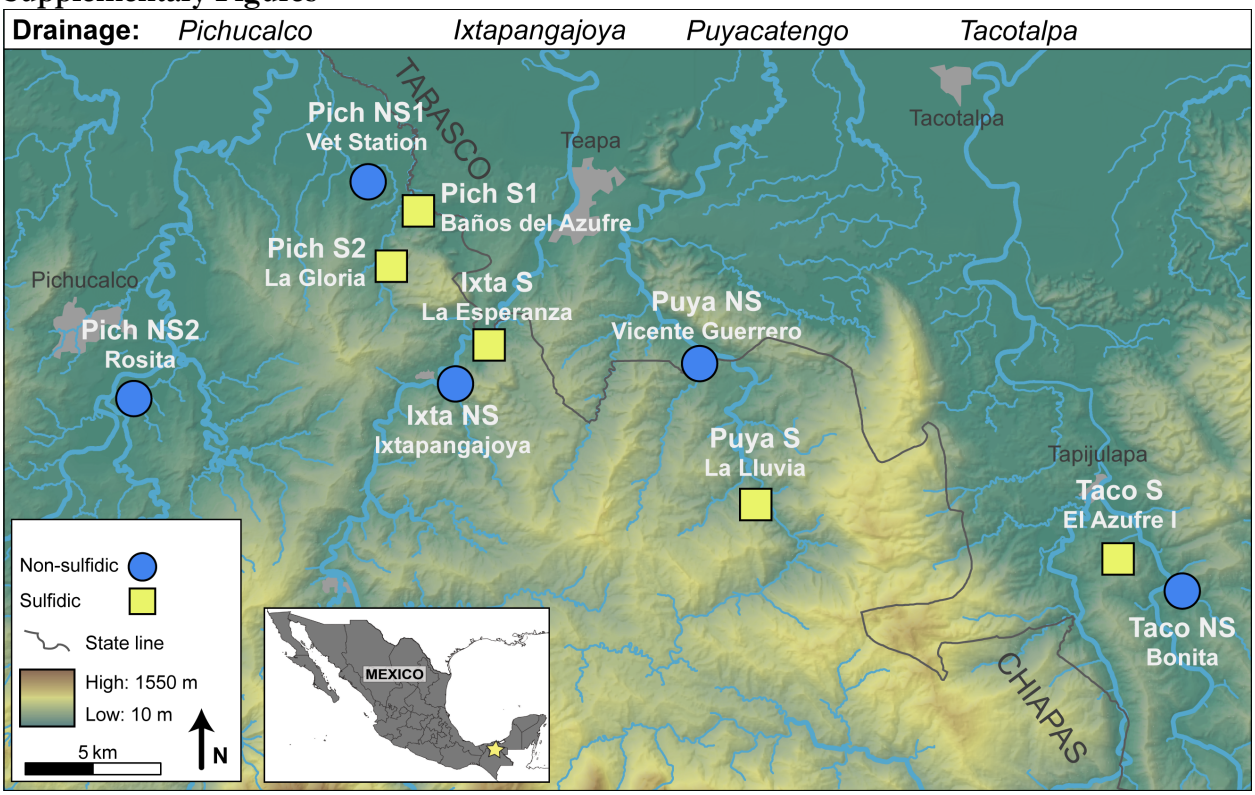
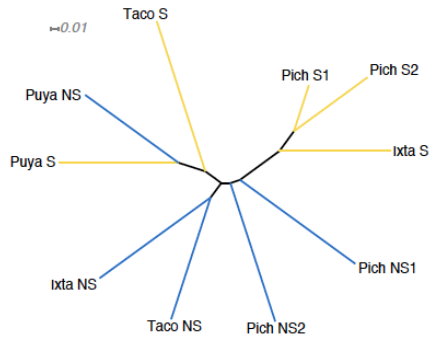
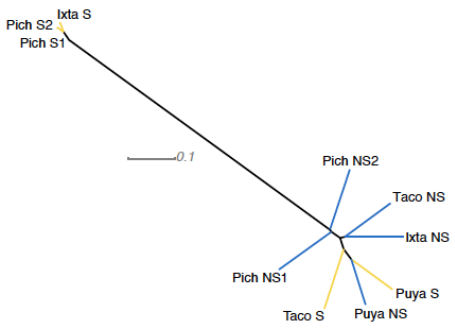


Figure S1. Map of study region and sites. Samples were collected from 10 sites in the Río Grijalva basin, including one individual each from 5 sulfidic and 5 non-sulfidic habitats in four different river drainages. Study area is indicated by a star in the inset map of Mexico. This figure was adapted from [7].

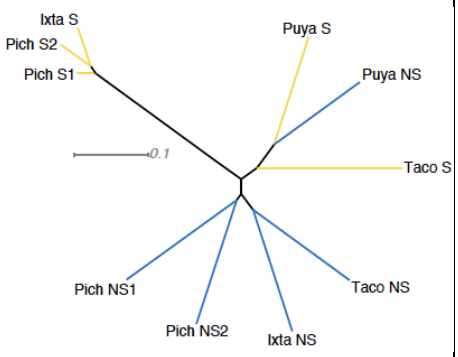
1



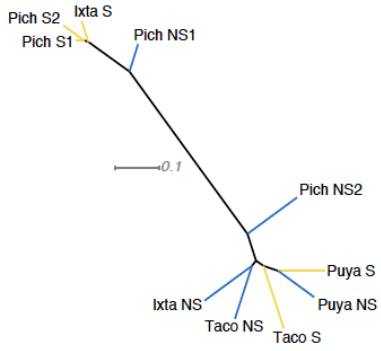
2



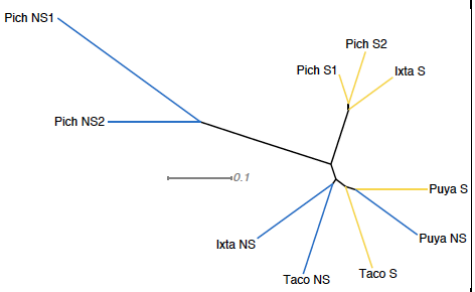
3



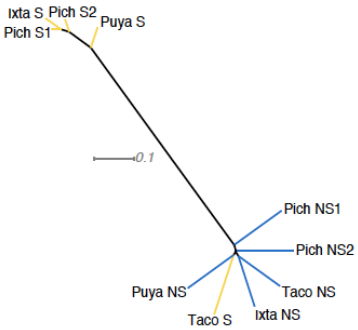
4



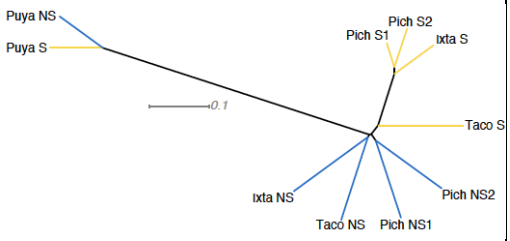
5



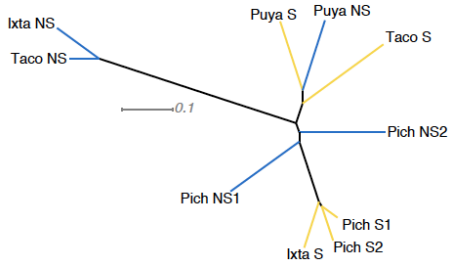
6



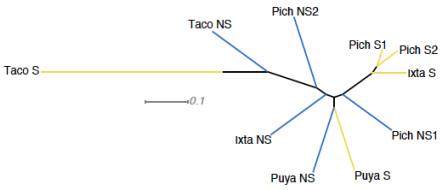
7



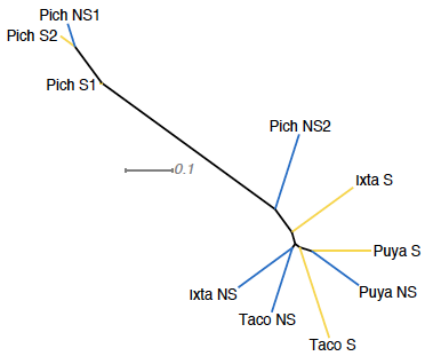
8



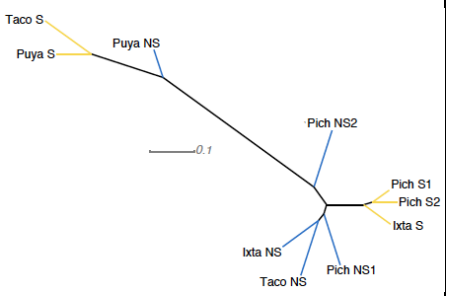
9



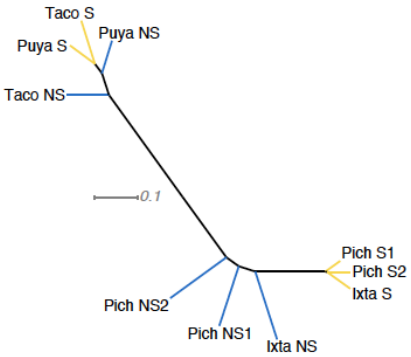
10



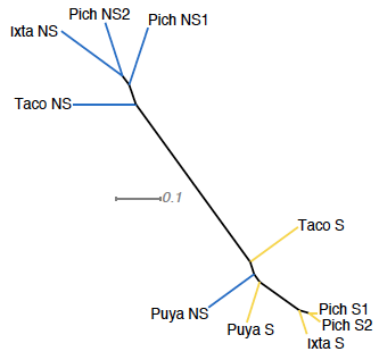
11



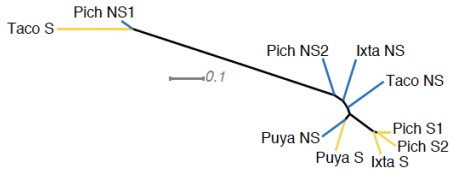
12



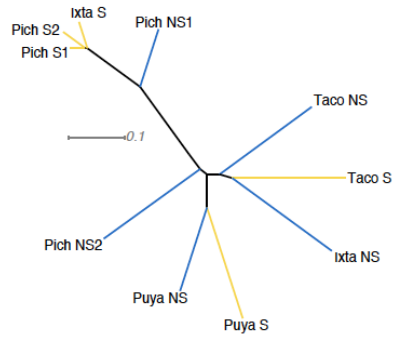
13



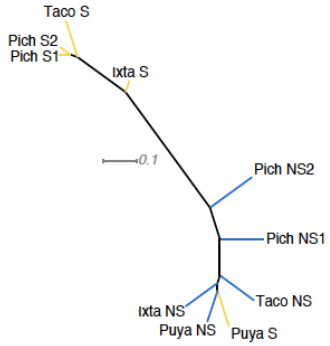
14



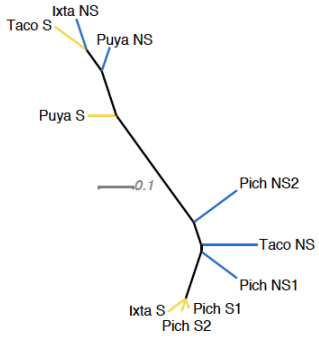
15



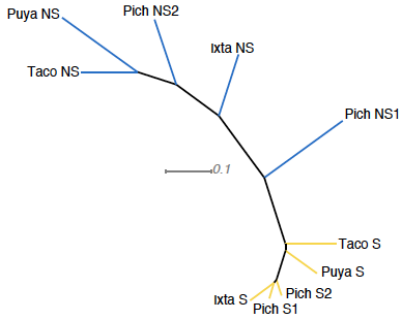
16



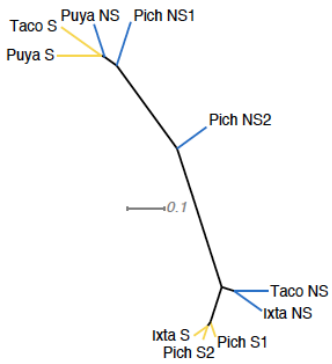
17



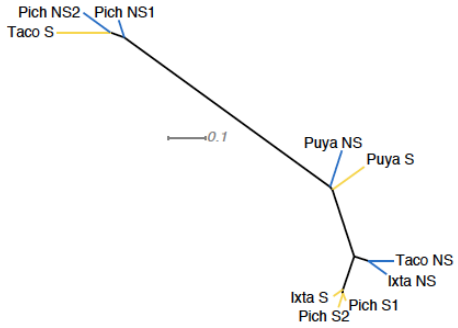
18



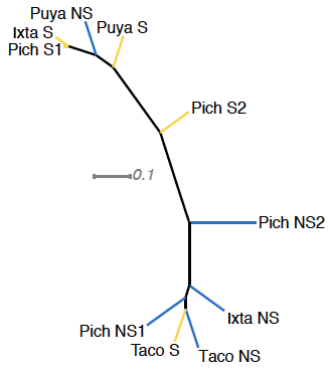
19



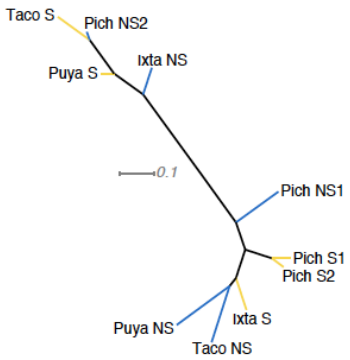
20



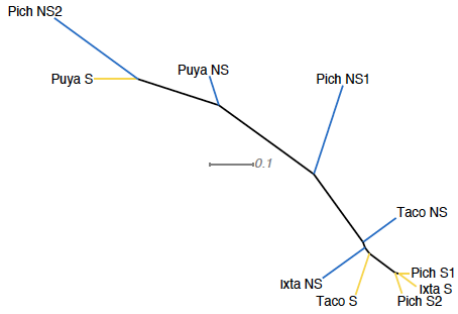
21



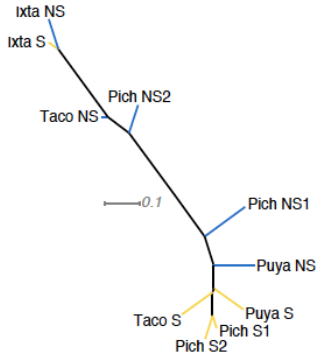
22



23



24



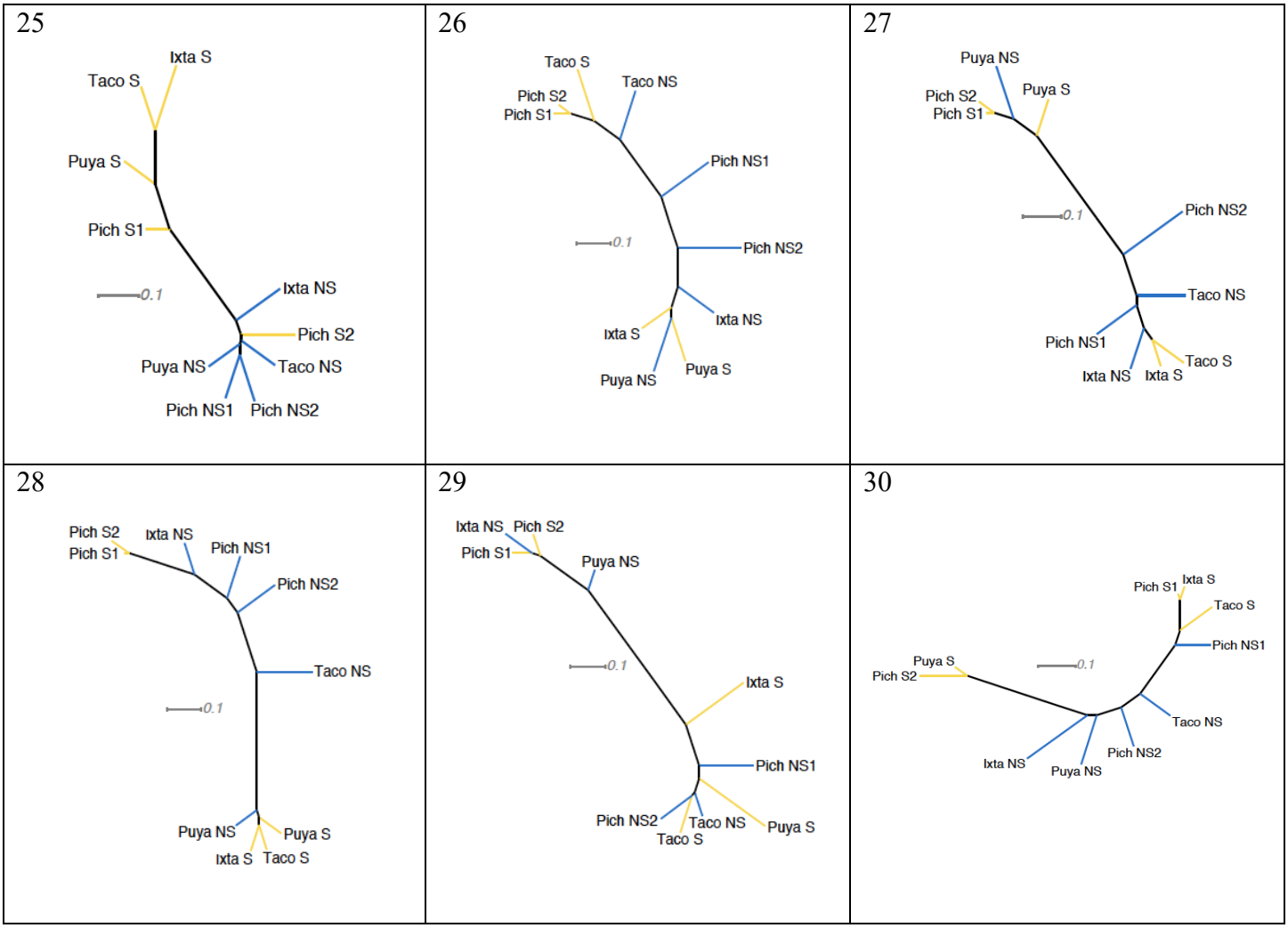
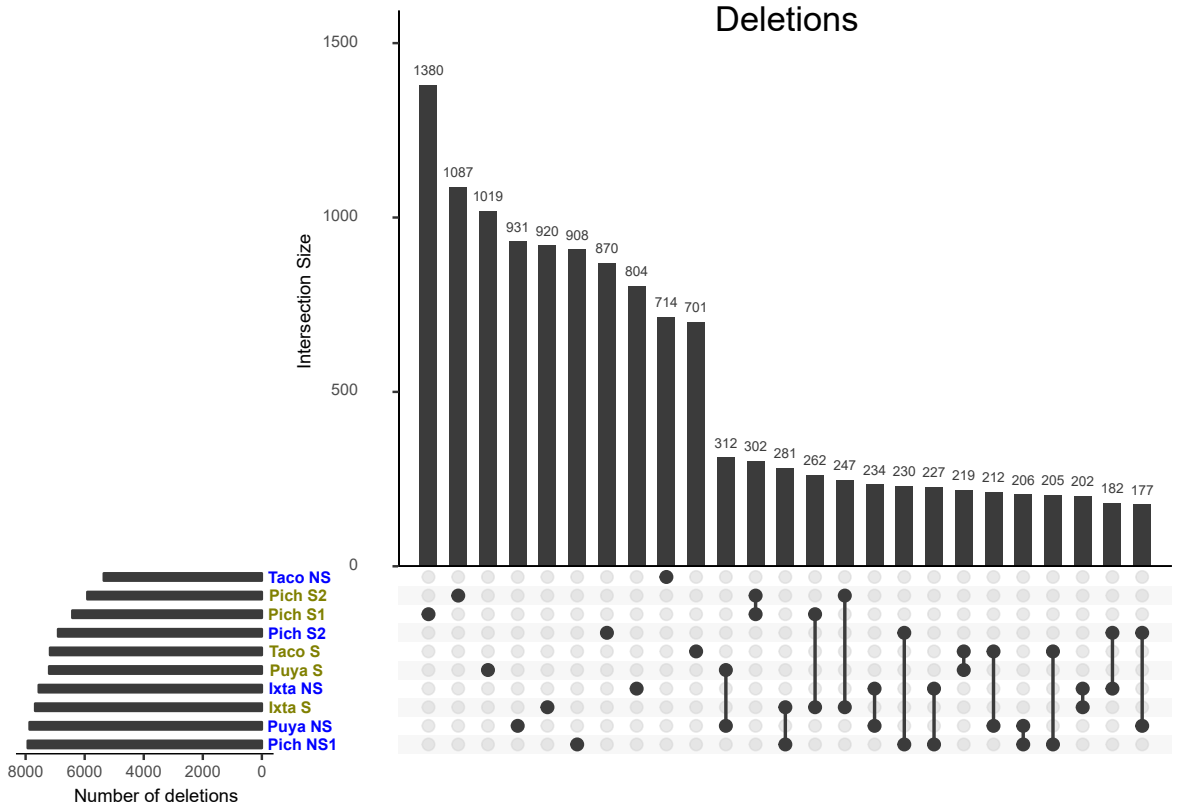
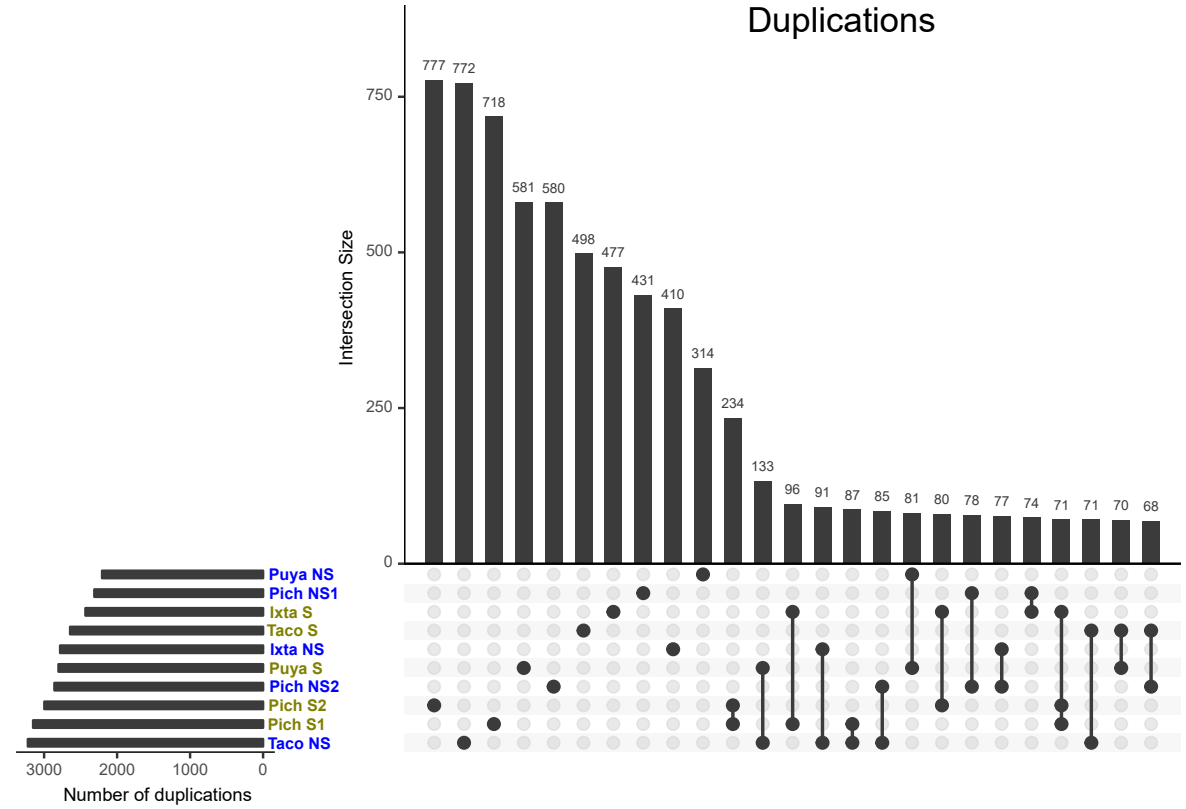
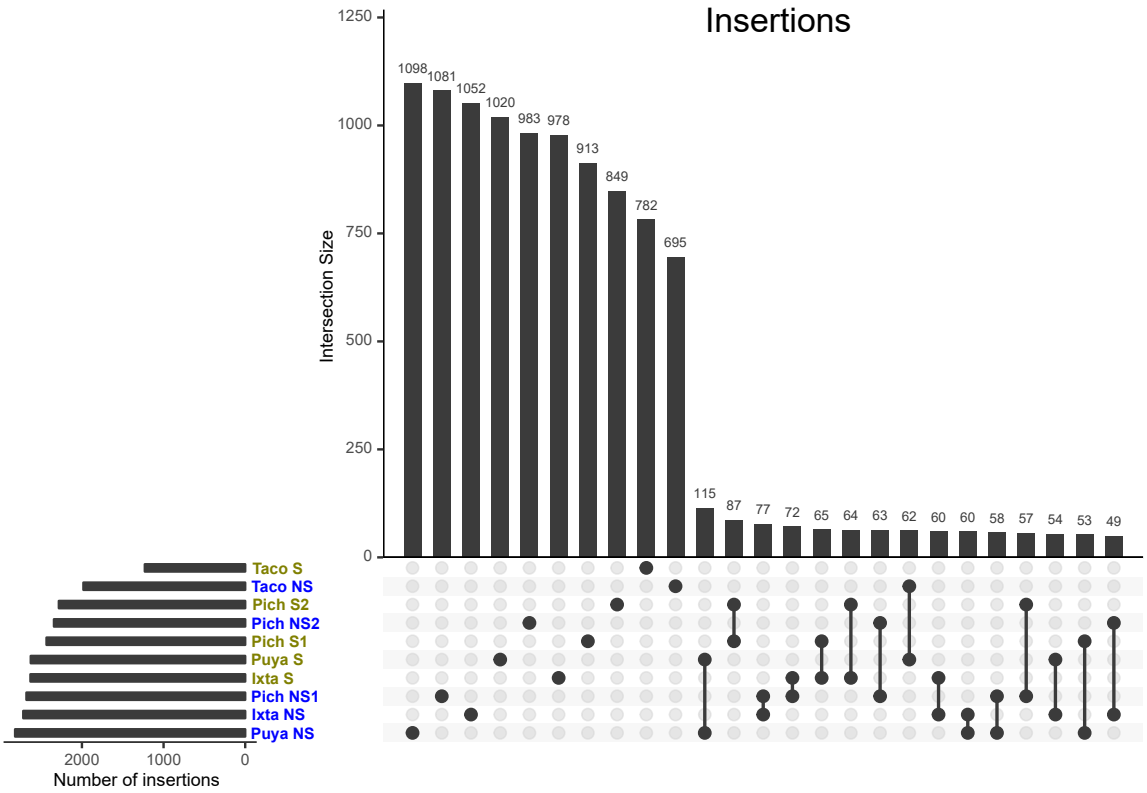


Figure S2. All 30 local topologies (cacti) hypothesized by Saguaro [1] across the genome. Cacti are presented in order of length of the genome covered (see Table 1).


Deletions
Intersection Size
1500
1000
500
0
1380
1087
1019
931
920
908
870
804
714
701
312
302
281
262
247
234
230
227
219
212
206
205
202
182
177
Taco NS
Pich S2
Pich S1
Pich S2
Taco S
Puya S
Ixta NS
Ixta S
Puya NS
Pich NS1
8000
6000
4000
2000
0
Number of deletions


Duplications
Intersection Size
750
500
250
0
777
772
718
581
580
498
477
431
410
314
234
133
96
91
87
85
81
80
78
77
74
71
71
70
68
Puya NS
Pich NS1
Ixta S
Taco S
Ixta NS
Puya S
Pich NS2
Pich S2
Pich S1
Taco NS
3000
2000
1000
0
Number of duplications


Insertions
Intersection Size
1250
1000
750
500
250
0
1098
1081
1052
1020
983
978
913
849
782
695
115
87
77
72
65
64
63
62
60
60
58
57
54
53
49
Taco S
Taco NS
Pich S2
Pich NS2
Pich S1
Puya S
Ixta S
Pich NS1
Ixta NS
Puya NS
2000
1000
0
Number of insertions

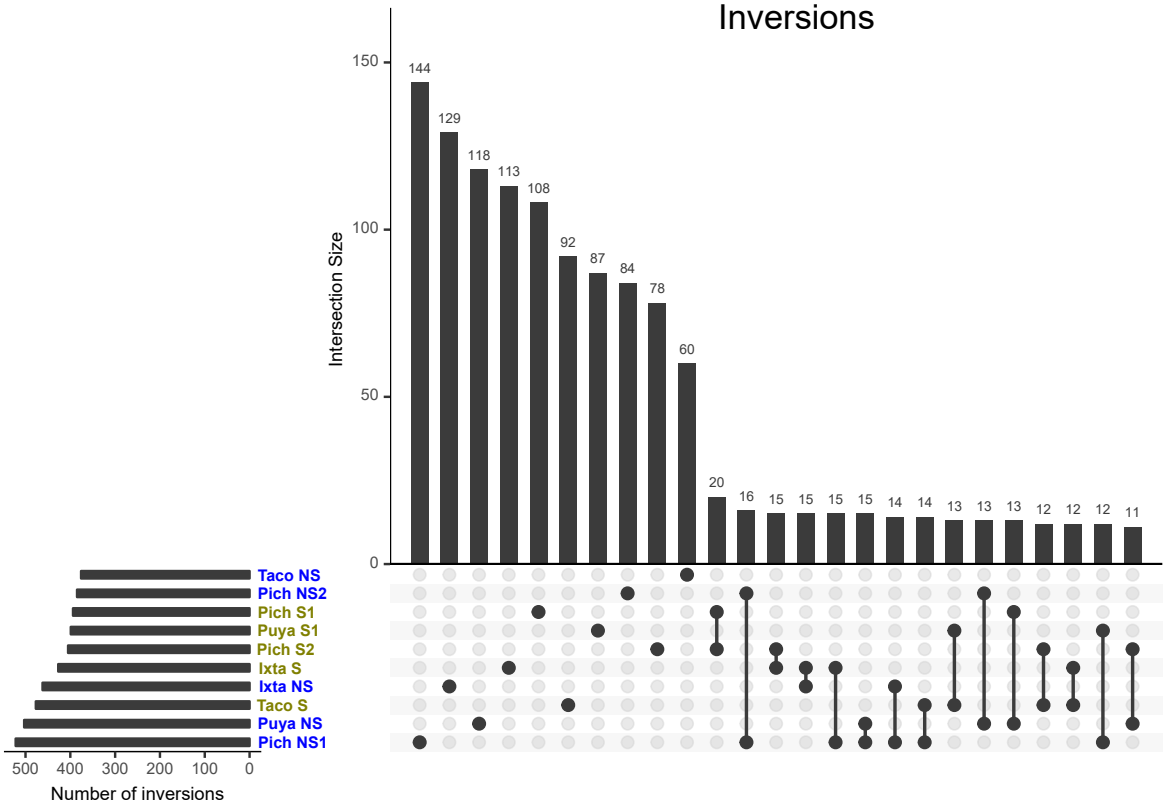
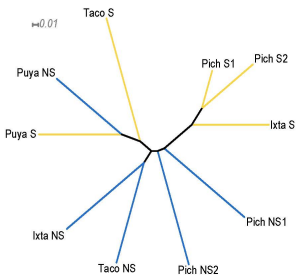
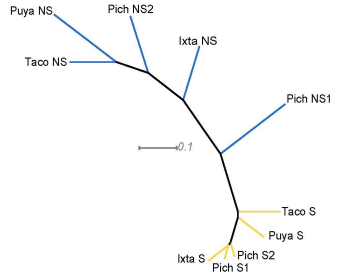
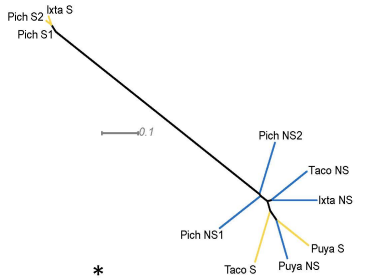

Inversions
Intersection Size
150
100
50
0
144
129
118
113
108
92
87
84
78
60
20
16
15
15
15
15
14
14
13
13
13
12
12
12
11
Taco NS
Pich NS2
Pich S1
Puya S1
Pich S2
Ixta S
Ixta NS
Taco S
Puya NS
Pich NS1
500
400
300
200
100
0
Number of inversions

Figure S3. Sharing of structural variants across individuals, visualized using UpSet [8]. These plots show the total number of each type of structural variant that was identified in each individual (non-sulfidic [blue] and sulfidic [yellow]) on the bottom left (for example, "Number of inversions"), and the top 25 categories of sharing across the ten individuals. The vertical bar represents the number of elements in the set indicated below. The set indicated is either a single filled circle, which represents variants unique to the indicated individual, whereas a line connecting filled circles represents variants shared between the indicated individuals.







Ixta S	C	T	T							C	A	T	C	A	G	G	A	T	C	G	G	G	T	A	G	T	G	A
Pich S1	C	T	T							C	A	T/C	C	A	G	G	A	T	C	G	G	G	T	A	G	T	G	A
Pich S2	T/C	C/T	T							C	A	T	C	A	G	G	A	T	C	G	G	G	T	A	G	T	G	A
Puya S	C	T	T							C	G	C	C	A	G	G	T	T	C	G	G	A	T	G	A	C	G	G
Taco S	C	T	T							T/C	A	C	T	G	G	G	T	T	C	G	G	A	C	G	A	C	G	G
Ixta NS	T/C	T	T							C	A	C	T	G	A	A	T	C	T	G	A	A	C	G	A	C	G	G
Pich NS1	C	T	T							C	A	C	T	G	A	A	T	C	C	G	A	A	C	G	A	C	G	G
Pich NS2	C	T	T							C	A	C	T	G	A	A	T	C	T/C	G	A	A	C	G	A	C	G	G
Puya NS	C	T	T							C	A	C	T	G	A	A	T	C	T	A	A	A	C	G	A	C	A	G
Taco NS	T/C	T	C/T							C	A	C	T	G	A	A	T	C	T	G	A	A	C	G	A	C	G	G

Figure S4. Local relationship patterns across three different loci. A portion of three gene regions are shown: putatively neutral gene (*TRAPPC9*) (left), *sqrdl* (middle), and mitochondrial *COX1* gene (right). Reference (black) and alternate (red) alleles are indicated for each variable site in the 500 bp region (not perfectly to scale). Unless otherwise noted by a slash, individuals are homozygous for the alleles shown. The cacti that were assigned to each locus are shown above. The site marked with an asterisk is a previously identified adaptive substitution that arose via de novo mutations in the Pichucalco and Puyacatengo sulfidic populations independently [9].

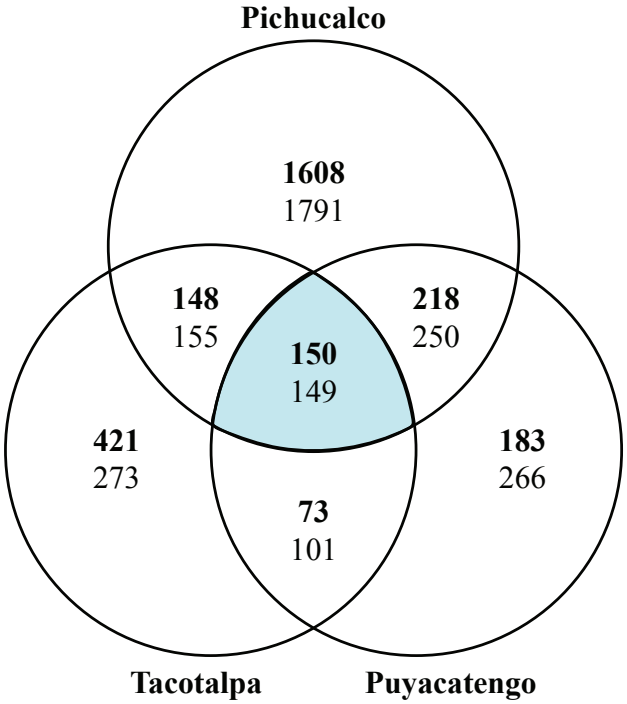


Figure S5. Venn diagram of candidate genes. Upregulated genes in sulfidic environments are in bold font. Downregulated genes are in regular font.

Supplementary Tables

Table S1. Collection sites, presence or absence of hydrogen sulfide (H_2S), number of reads, coverage.

Site	Site Name	Drainage	H_2S	Species	Number of reads, pre-trimming (Millions)	Number of reads, post-trimming (Millions)	Mean coverage per site in the genome
Ixta NS	Ixtapangajoya	Ixtapangajoya	-	*P. mexicana*	409	376	48.89
Ixta S	La Esperanza	Ixtapangajoya	+	*P. thermalis*	482	457	59.70
Pich NS1	Vet Station	Pichucalco	-	*P. mexicana*	505	469	60.60
Pich NS2	Rosita	Pichucalco	-	*P. mexicana*	367	352	45.33
Pich S1	Baños del Azufre	Pichucalco	+	*P. sulphuraria*	318	299	41.81
Pich S2	La Gloria	Pichucalco	+	*P. sulphuraria*	367	345	45.45
Puya NS	Vicente Guerrero	Puyacatengo	-	*P. mexicana*	365	344	45.21
Puya S	La Lluvia	Puyacatengo	+	*P. mexicana*	297	283	36.93
Taco NS	Bonita	Tacotalpa	-	*P. mexicana*	249	233	30.47
Taco S	El Azufre I	Tacotalpa	+	*P. mexicana*	369	342	48.36

Table S2. Genes that were assigned to cacti that cluster by ecotype. *Poecilia mexicana* accessions from GenBank. Human orthologues obtained from a previous BLASTX search [10]. The first column is the gene ID from the *Poecilia mexicana* annotation, the second column is the gene name from the *Poecilia mexicana* annotation, the third column is the name of the gene with the top BLASTX hit from Swissprot, the fourth column is the Swissprot accession for that hit, and the last column describes whether that gene was up or downregulated in the sulfidic populations.

See provided excel spreadsheet

Table S3. The number of basepairs each type of structural variant covered in each individual.

Individual	Inversion	Gain	Insertion	Deletion	Total
Ixta NS	2,092,950	12,023,728	842,095	17,271,706	32,230,479
Ixta S	2,472,611	9,031,058	880,540	19,388,589	31,772,798
Pich NS1	3,360,037	9,542,847	899,765	20,273,975	34,076,624
Pich NS2	1,541,464	12,129,418	772,537	15,557,884	30,001,303
Pich S1	1,763,576	15,445,659	758,442	15,782,360	33,750,037
Pich S2	2,099,332	13,550,713	819,770	15,812,613	32,282,428
Puya NS	1,854,590	8,779,470	846,548	19,541,285	31,021,893
Puya S	1,217,186	9,933,593	759,072	17,808,904	29,718,755
Taco NS	1,535,298	13,527,813	646,403	10,968,545	26,678,059
Taco S	1,979,709	11,900,759	465,557	19,673,818	34,019,843

Table S4. Sharing of structural variants (SVs) between sulfidic individuals (and at most one non-sulfidic individual). The first column describes which individuals share the SVs, the second column is the number of SVs that are shared between only these individuals, the third column is the length (in bp) that these SVs cover, the fourth column shows the percentage of the genome that these SVs cover, the fifth column shows the number of genes these SVs overlap with, the sixth column indicates the number of these SVs that overlap with Saguaro regions that cluster the individuals in a similar way, and the seventh column indicates which Saguaro cactus displayed the similar pattern of sharing.

Individuals sharing a structural variant	Count	Length	% genome covered	Genes	Saguaro regions	Similar Saguaro number
Pich S1, Ixta S, Puya S, Taco S	25	65552	0.008	25	0	25
Pich S1, Pich S2, Ixta S, Puya S	24	65544	0.008	24	5	6
Pich S1, Pich S2, Puya S, Taco S	13	48083	0.006	13	0	24
Pich S1, Pich S2, Ixta S, Puya S, Taco S, Pich NS1	13	63774	0.008	12	N/A	N/A
Pich S1, Pich S2, Ixta S, Puya S, Taco S, Puya NS	11	43191	0.005	10	0	13
Pich S2, Ixta S, Puya S, Taco S	11	33848	0.004	8	N/A	N/A
Pich S1, Pich S2, Ixta S, Puya S, Taco S, Ixta NS	10	39469	0.005	8	N/A	N/A
Pich S1, Pich S2, Ixta S, Puya S, Taco S, Pich NS2	8	40929	0.005	9	N/A	N/A
Pich S1, Pich S2, Ixta S, Puya S, Taco S, Taco NS	7	23293	0.003	5	N/A	N/A
Pich S1, Pich S2, Ixta S, Puya S, Taco S	3	9620	0.001	2	0	18
Pich S1, Pich S2, Ixta S, Taco S	0	0	0	0	0	16

Table S5. Genes that overlapped with structural variants (SVs) that were shared by sulfidic individuals. The first column is the gene ID from the *Poecilia mexicana* annotation, the second column is the gene name from the *Poecilia mexicana* annotation, the third column is the name of the gene with the top BLASTX hit from SwissProt, the fourth column is the SwissProt accession for that hit, the fifth column describes the type of SV (INS = Insertion, DEL = Deletion, GAIN = Duplication, INV = Inversion), the sixth column describes the individuals the SV was identified in, and the last column describes whether that gene was up or downregulated in the sulfidic populations.

See provided excel spreadsheet

Table S6. Enriched Gene Ontology (GO) biological process terms in genes assigned to each cactus that clusters by ecotype, and all genes in any cacti that cluster by ecotype. N = total number of genes in the reference set, B = number of genes in the reference set that were associated with GO term, n = total number of genes in the test set, b = number of genes in the test set that were associated with GO term. Enrichment calculated by comparing test set relative to the reference set $[(b/n)/(B/N)]$.

See provided excel spreadsheet

Table S7. Genes that were consistently differentially expressed between sulfidic and non-sulfidic populations in three drainages (Pichucalco, Puyacatengo, and Tacotalpa). The first column is the gene ID from the *Poecilia mexicana* annotation, the second column is the gene name from the *Poecilia mexicana* annotation, the third column is the name of the gene with the top BLASTX hit from Swissprot, the fourth column is the Swissprot accession for that hit, and the fifth column describes whether the gene was up or downregulated in the sulfidic populations.

See provided excel spreadsheet

References

[1] Zamani, N., Russell, P., Lantz, H., Hoeppner, M.P., Meadows, J.R.S., Vijay, N., Mauceli, E., di Palma, F., Lindblad-Toh, K., Jern, P., et al. 2013 Unsupervised genome-wide recognition of local relationship patterns. BMC Genomics 14, 347. (doi:10.1186/1471-2164-14-347).

[2] Kelley, J.L., Arias-Rodriguez, L., Patacsil Martin, D., Yee, M.C., Bustamante, C.D. & Tobler, M. 2016 Mechanisms Underlying Adaptation to Life in Hydrogen Sulfide-Rich Environments. Molecular biology and evolution 33, 1419-1434. (doi:10.1093/molbev/msw020).

[3] Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L. & Pachter, L. 2013 Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol 31, 46-53. (doi:10.1038/nbt.2450).

[4] Kim, D., Langmead, B. & Salzberg, S.L. 2015 HISAT: a fast spliced aligner with low memory requirements. Nat Methods 12, 357-360. (doi:10.1038/nmeth.3317).

[5] Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T. & Salzberg, S.L. 2015 StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol 33, 290-295. (doi:10.1038/nbt.3122).

[6] Robinson, M.D., McCarthy, D.J. & Smyth, G.K. 2010 edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics (Oxford, England) 26, 139-140. (doi:10.1093/bioinformatics/btp616).

[7] Hotaling, S., Quackenbush, C.R., Bennett-Ponsford, J., New, D.D., Arias-Rodriguez, L., Tobler, M. & Kelley, J.L. 2018 Bacterial Diversity in Replicated Hydrogen Sulfide-Rich Streams. Microbial Ecology. (doi:10.1007/s00248-018-1237-6).

[8] Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R. & Pfister, H. 2014 UpSet: Visualization of Intersecting Sets. IEEE transactions on visualization and computer graphics 20, 1983-1992. (doi:10.1109/TVCG.2014.2346248).

[9] Pfenninger, M., Lerp, H., Tobler, M., Passow, C., Kelley, J.L., Funke, E., Greshake, B., Erkoc, U.K., Berberich, T. & Plath, M. 2014 Parallel evolution of cox genes in H2S-tolerant fish as key adaptation to a toxic environment. Nature communications 5. (doi:10.1038/ncomms4873).

[10] Passow, C.N., Brown, A.P., Arias-Rodriguez, L., Yee, M.C., Sockell, A., Schartl, M., Warren, W.C., Bustamante, C., Kelley, J.L. & Tobler, M. 2017 Complexities of gene expression patterns in natural populations of an extremophile fish (Poecilia mexicana, Poeciliidae). Molecular ecology 26, 4211-4225. (doi:10.1111/mec.14198).