



**HAL**  
open science

# The nature and distribution of non-functional alleles suggest only two independent events at the origins of *Astyanax mexicanus* cavefish populations

Maxime Policarpo, Laurent Legendre, Isabelle Germon, Luis Espinasa, Sylvie Rétaux, Didier Casane, Philippe Lafargeas

## ► To cite this version:

Maxime Policarpo, Laurent Legendre, Isabelle Germon, Luis Espinasa, Sylvie Rétaux, et al.. The nature and distribution of non-functional alleles suggest only two independent events at the origins of *Astyanax mexicanus* cavefish populations. 2023. hal-04279274

**HAL Id: hal-04279274**

**<https://hal.science/hal-04279274>**

Preprint submitted on 10 Nov 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

# The nature and distribution of non-functional alleles suggest only two independent events at the origins of *Astyanax mexicanus* cavefish populations

**Maxime Policarpo**

Université Paris-Saclay, CNRS, IRD, UMR Évolution

**Laurent Legendre**

Université Paris-Saclay, CNRS, IRD, UMR Évolution

**Isabelle Germon**

Université Paris-Saclay, CNRS, IRD, UMR Évolution

**Luis Espinasa**

Marist College

**Sylvie Rétaux**

Université Paris-Saclay and CNRS

**Didier Casane** (✉ [didier.casane@universite-paris-saclay.fr](mailto:didier.casane@universite-paris-saclay.fr))

Université Paris-Saclay, CNRS, IRD, UMR Évolution

**Philippe Lafargeas**

Université Paris-Saclay, CNRS, IRD, UMR Évolution

---

## Research Article

**Keywords:** cavefish, relaxed selection, pseudogenes, visual perception

**Posted Date:** July 17th, 2023

**DOI:** <https://doi.org/10.21203/rs.3.rs-3155771/v1>

**License:** © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

**Additional Declarations:** No competing interests reported.

---

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21

**The nature and distribution of non-functional alleles suggest only two independent events at the origins of *Astyanax mexicanus* cavefish populations.**

Maxime Policarpo<sup>1,†</sup>, Laurent Legendre<sup>1</sup>, Isabelle Germon<sup>1</sup>, Philippe Lafargeas<sup>1</sup>, Luis Espinasa<sup>2</sup>, Sylvie Rétaux<sup>3,\*</sup> and Didier Casane<sup>1,4,\*</sup>

<sup>1</sup> Université Paris-Saclay, CNRS, IRD, UMR Évolution, Génomes, Comportement et Écologie, 91190 Gif-sur-Yvette, France.

<sup>2</sup> School of Science, Marist College, Poughkeepsie, New York, USA.

<sup>3</sup> Institut de Neurosciences Paris-Saclay, Université Paris-Saclay and CNRS, 91400 Saclay, France.

<sup>4</sup> Université Paris Cité, UFR Sciences du Vivant, 75013 Paris, France.

<sup>†</sup>: present address: Zoological Institute, Department of Environmental Sciences, University of Basel, Basel, Switzerland.

\* Corresponding authors: e-mails: [sylvie.retaux@cnrs.fr](mailto:sylvie.retaux@cnrs.fr); [didier.casane@universite-paris-saclay.fr](mailto:didier.casane@universite-paris-saclay.fr).

## 22 **Abstract**

23

### 24 **Background**

25 Several studies suggested that cavefish populations of *Astyanax mexicanus* settled during the  
26 Late Pleistocene. This implies that the cavefish's most conspicuous phenotypic changes,  
27 blindness and depigmentation, and more cryptic characters important for cave life, evolved  
28 rapidly.

### 29 **Results**

30 Using the published genomes of 47 *Astyanax* cavefish from la Cueva de El Pachón, El Sótano  
31 de la Tinaja, La Cueva Chica and El Sótano de Molino, we searched for loss-of-function  
32 mutations in previously defined sets of genes, *i.e.*, vision, circadian clock and pigmentation  
33 genes — non-functional alleles for four vision genes were identified. Then, we searched  
34 genome-wide for non-functional alleles in these four cave populations. Among 512 genes  
35 with segregating non-functional alleles, we found an enrichment in visual perception genes.  
36 Among cavefish populations, different levels of shared non-functional alleles were found.  
37 Using a subset of 12 genes for which loss-of-function mutations were found, we extend the  
38 analysis of shared pseudogenes to 11 cave populations. Using a subset of six genes for which  
39 loss-of-function mutations were found in the El Sótano del Toro population, where extensive  
40 hybridization with surface fish occurs, we found a correlation between the level of eye  
41 regression and the amount of non-functional alleles.

### 42 **Conclusions**

43 We can confirm that very few non-functional alleles are present in a large set of vision genes,  
44 in accordance with the recent origin of *Astyanax mexicanus* cavefish. Nevertheless, the  
45 genome-wide analysis indicates an enrichment of loss-of-function alleles in genes with vision-

46 related GO-terms, suggesting that visual perception may be the function chiefly impacted by  
47 gene losses related to the shift from a surface to a cave environment.

48 The geographic distribution of the loss-of-function alleles supports the hypothesis that cave  
49 populations from Sierra de Guatemala and Sierra de El Abra share a common origin, albeit  
50 followed by independent evolution for a long period of time. It also supports that populations  
51 from the Micos area have an independent origin. In El Sótano del Toro, the troglomorphic  
52 phenotype is maintained despite massive introgression of the surface genome.

53

54 **Keywords:** cavefish, relaxed selection, pseudogenes, visual perception.

55

## 56 **Background**

57

58 In the species *Astyanax mexicanus*, a Mexican freshwater fish, cave and surface fish  
59 populations with strikingly different phenotypes coexist [1]. Surface fish have a wide  
60 distribution and are abundant, whereas the known cavefish are found in 33 caves [2-6] in a  
61 small area comprising three mountain ranges in Northeastern Mexico: Sierra de El Abra,  
62 Sierra de Guatemala, and Micos area (**Fig. 1**). The timing and the number of independent  
63 colonization events that gave rise to the cavefish populations is a matter of debate, but most  
64 recent studies point to a relatively recent origin (20,000 to 200,000 years) and few  
65 independent events (2 to 4) [7-9].

66 The possibility of breeding cavefish and surface fish makes *Astyanax mexicanus* an  
67 outstanding model to study genetic mechanisms underlying the evolution of some cave-  
68 associated traits such as eye loss, depigmentation, aggression loss, and sleep loss [10]. QTL  
69 analyses identified loci involved in these traits [10], but only a few studies could pinpoint  
70 changes at the DNA level [11, 12]. Different loci or mutations have been associated with  
71 similar phenotypes in some cases, demonstrating convergent evolution of some traits in  
72 different caves [10, 12, 13]. However, the fact that independent mutations are involved in the  
73 evolution of convergent traits in different cave populations does not invalidate the hypothesis  
74 of a common origin. One could imagine an ancestral population that evolved several cave-  
75 specific traits before dispersing to other caves, in which these traits continued to evolve  
76 independently in parallel with the emergence of new additional specific traits. Whereas high  
77 migration flow between distant caves is unlikely, sporadic migration over relatively long  
78 distances has been documented in other cavefishes, allowing range extension and independent  
79 evolution [14-17].

80 Population isolation is often studied using genetic markers assumed to be neutral, such as  
81 SNPs and microsatellites. Depending on the model assumed, if there is or is not an  
82 equilibrium between genetic drift and gene flow, it allows the estimation of demographic  
83 parameters such as effective population size and gene flow (at genetic drift / gene flow  
84 equilibrium) and isolation time (if no equilibrium is assumed). However, it is difficult to  
85 evaluate the relative effect of recent isolation and gene flow on shared polymorphism when  
86 comparing recently isolated populations. Surface and cave populations of *A. mexicanus* may  
87 share many alleles at many loci because cave populations are recent. This hypothesis is  
88 supported by an analysis of microsatellite and SNP polymorphism [8] and the presence of  
89 only one pseudogene among a set of 85 vision genes in the *A. mexicanus* El Pachón cavefish  
90 reference genome; as opposed to the numerous pseudogenes found in other, more ancient  
91 cavefish species [9, 18, 19].

92 Here, we analyse the nature and geographic distribution of cave-specific loss-of-function  
93 (LoF) mutations, alleles that most likely appeared after cave colonisation and could not easily  
94 spread through surface fish to other caves. Using 47 published genomes from four different  
95 *Astyanax* cave populations, we searched for LoF mutations in previously defined sets of genes  
96 (vision, circadian clock and pigmentation genes) and genome-wide. We identified a few LoF  
97 in vision, circadian clock and pigmentation genes, and 533 LoF mutations in 512 genes at the  
98 whole genome scale. Using our DNA library of wild fish from 11 different caves, we further  
99 examined the geographic distribution of 12 of these LoF mutations. These data suggest  
100 different levels of common *versus* independent evolution among caves in the Micos area,  
101 Sierra de El Abra, and Sierra de Guatemala.

102

## 103 **Results**

104

## 105 Distribution of LoF mutations in vision, circadian clock and pigmentation genes

106 In a previous study [9], we defined 63 zebrafish eye-specific genes expressed only in the eyes,  
107 together with 32 genes coding for non-visual opsins that we collectively called vision genes.  
108 We also defined two other gene sets: 42 genes involved in the circadian clock and 257 genes  
109 involved in pigmentation. In the reference genome of *Astyanax mexicanus* surface fish, we  
110 found 56, 30, 38 and 249 genes homologous to these zebrafish genes, respectively [9]. Here,  
111 using published genomic reads of 47 *Astyanax mexicanus* cavefish [7, 20], from four cave  
112 populations (9 Molino, 9 Pachón, 10 Tinaja and 19 Chica), we identified LoF mutations  
113 present in at least two individuals in these three sets of genes (**Table 1**). First, we confirmed  
114 the LoF mutation in the vision gene *pde6b*, found previously in the Pachón cavefish reference  
115 genome [9]. This mutation is not fixed in the Pachón population (present in 8 out of 9 Pachón  
116 individuals, 4 homozygotes and 4 heterozygotes, **Additional file 1**). It is absent from the 3  
117 other cave populations examined. Among vision genes, we found LoF mutations in three  
118 additional genes (*sagb*, *cryba2a* and *rho1*), which are not present in the Pachón cavefish  
119 reference genome, and none is fixed. While the *sagb* non-functional allele was found in  
120 Pachón and Tinaja, the *cryba2a* non-functional allele was found in Tinaja and Chica, and the  
121 *rho1* non-functional allele was only found in Tinaja (**Additional file 1**). We also retrieved a  
122 recently identified LoF mutation in the circadian clock gene *per2* [21], which is present only,  
123 and fixed, in Molino. We further detected LoF mutations in six pigmentation genes (*csfr1b*,  
124 *kif13a*, *cd63*, *hps4*, *drd2b*, *tyrp1a*), none being fixed except *hps4*, found only in Molino  
125 (**Table 1**). The LoF mutation in *tyrp1a* is present in the Pachón cavefish reference genome  
126 [9]. Thus, using our predefined gene sets, differences in the distribution of LoF mutations  
127 emerged between Sierra de El Abra caves on the one hand (Pachón, Tinaja, and Chica) and  
128 Molino cave located in Sierra de Guatemala on the other hand.

129



## 130 Genome-wide analysis of the distribution of LoF mutations

131 We then extended the search for LoF mutations genome-wide. Using genomic reads from the  
132 47 cavefish described above, genomics reads of 15 surface fish from two surface locations (9  
133 from Nacimiento del Rio Choy and 6 from Rio Gallinas, near Rascón), as well as one  
134 *Astyanax aeneus* and one *Gymnocorymbus ternetzi* individual as outgroups, we identified a  
135 total of 533 cave specific LoF mutations distributed in 512 genes (491 genes with 1 LoF  
136 mutation and 21 genes with 2 LoF mutations) (**Additional file 1**). Among them, there were  
137 368 frameshifts (240 deletions and 128 insertions), 140 premature stop codons, 13 losses of  
138 the start codons, 2 losses of stop codons and 10 splice site mutations (**Additional file 2: Fig.**  
139 **S1A**). The distribution of frameshift sizes followed the expected pattern, with most indels  
140 involving one or two base pairs (**Additional file 2: Fig. S1B**).

141 There were 82, 90, 150 and 106 LoF mutations that were specific to Pachón, Tinaja, Chica  
142 and Molino populations, respectively (**Fig. 2A**). Chica had the highest count of LoF  
143 mutations, which is not due to the higher sample size of this population, as it was still the case  
144 when sampling at random nine fish in each population (**Fig. 2B**). While 78 LoF mutations  
145 were shared between cave pairs, only 25 were shared between 3 caves and only 2 between the  
146 4 caves (**Fig. 2A**). Pairwise comparisons showed that more LoF mutations were shared  
147 between pairs of Sierra de El Abra caves (12, 10 and 33) than between a Sierra de El Abra  
148 cave and Molino cave (Sierra de Guatemala region; 5, 6 and 12) (**Fig. 2A**). Nevertheless, 30  
149 LoF mutations were shared between Molino and at least one Sierra de El Abra cave (**Fig. 2A**).  
150 The same trend was observed when considering 9 fish per population (**Fig. 2B**).

151

## 152 Enrichment of LoF in genes with vision-related GO-terms

153 In the set of 512 pseudogenes identified above, we searched for enrichment in some GO-  
154 terms, using human and zebrafish orthologous gene names. Using orthologous zebrafish

155 genes, we found significant enrichments in biological processes linked to eye and vision. The  
156 three most significant terms were “eye photoreceptor cell differentiation”, “peripheral nervous  
157 system neuron development” and “cellular response to UV” (**Additional file 2: Table S1A**).  
158 Significantly enriched terms, almost all related to eye and vision, were also found with the  
159 Phenotype GeneRIF Predicted Z-score database (**Additional file 2: Table S1B**).  
160 We found that the most enriched biological process was “visual perception” with human  
161 orthologous genes using the DAVID database, but the FDR-adjusted p-value was not  
162 significant. Furthermore, there was a significant enrichment of genes involved in the disease  
163 “*retinitis pigmentosa*” using both OMIM and ClinVar 2019 databases (**Additional file 2:**  
164 **Table S1C**).

165

### 166 **Characterisation of large deletions in *opn8* and *rgr2* in cave populations**

167 To further identify LoF mutations due to large deletions that may have been overlooked with  
168 the above method, we used a combination of read-depth analysis and genome assembly  
169 comparison between a surface and cave fish. We identified a deletion of ~50kb encompassing  
170 the 3 last exons of the non-visual opsin gene *opn8a* (gene with 4 exons) and the whole *opn8b*  
171 gene (gene with 4 exons), called  $\Delta$ -*opn8*. These two genes are located in a genomic interval  
172 associated with differences in locomotor activity patterns between cave and surface fishes [22,  
173 23]. The deletion was found at high frequency in Pachón (0.43), Tinaja (0.8) and Chica (0.66),  
174 but was absent in Molino (**Additional file 2: Fig S2A,B**).

175 Another deletion of ~3kb genomic region containing the exons 4 and 5 of the non-visual opsin  
176 gene *rgr2* (gene with 7 exons), called  $\Delta$ -*rgr2*, was shared at high frequency in the four  
177 populations (Molino = 1.00, Pachón = 0.94, Tinaja = 1.00 and Chica = 0.82) (**Additional file**  
178 **2: Fig S2C, D**). Mutations in the human orthologous gene *RGR* are responsible for *retinitis*  
179 *pigmentosa* [24].

180

## 181 Distribution of a subset of 12 LoF mutations in 11 cave populations

182 We next wished to extend our analysis of the distribution of LoF mutations to additional  
183 cavefish populations distributed in the 3 mountain ranges (**Fig. 1**), for which we have  
184 collected DNA samples. These included 11 caves located in Sierra de Guatemala [Jineo (n = 6  
185 individuals), Molino (n = 9), Escondido (n = 7), Caballo moro (n = 2)], in Sierra de El Abra  
186 [Pachón (n = 55), Sabinos (n = 33), Tinaja (n = 16), Curva (n = 23), Toro (n = 30), Chica (n =  
187 33)] and in Micos area [Subterráneo (n = 26)], together with 234 surface fish sampled in local  
188 rivers and ponds (**Fig. 1**). In addition to the two large deletions described above in two non-  
189 visual opsins ( *opn8*  and  *rgr2* ), we selected 10 LoF mutations in nine genes of particular  
190 interest: seven genes related to vision ( *rh1.2* ,  *pde6b* ,  *cryba2a* ,  *impg1b* ,  *prph2b* ,  *gja8a* ,  *lim2.1* ),  
191 one to pigmentation ( *tyrp1a* ) and one to the circadian clock ( *per2* ). For these 10 LoF  
192 mutations, mutation type and position can be found in **Additional file 1: LoF\_list.xlsx**. To  
193 note, two independent LoF mutations were found in  *per2* , one specific to Molino (called  
194  $\Psi$ - *per2* -M, first identified with the genomic approach, see **Table 1** and **Table 2**) and one  
195 specific to Tinaja (called  $\Psi$ - *per2* -T). As the latter was present in only 1 individual in the  
196 genomic data set, it was not identified initially with the genomic approach using the threshold  
197 ‘present in at-least two individuals’. Mutation frequencies estimated in Molino, Pachón,  
198 Tinaja and Chica populations from our wild-caught samples were similar to those inferred  
199 from genomic data, suggesting that both estimates were reliable (**Additional file 2: Fig S3**).  
200 The geographic distribution of LoF mutations was very variable, from ‘specific to one cave’  
201 to ‘present in most caves’. Moreover, when found in several caves, LoFs primarily distributed  
202 in a geographical cluster of caves instead of being scattered in the whole cavefish distribution  
203 area (**Table 2**). Two pseudogenes were specific to one cave:  $\Psi$ - *pde6b*  in Pachón and  
204  $\Psi$ - *per2* -T in Tinaja. Other pseudogenes showed larger geographical areas, while none of the

205 tested LoF mutations was found in all 11 caves. The most widely distributed pseudogene  
206 ( $\Delta$ -*rgr2*) was present at high frequencies in all caves except Subterráneo, in which it was  
207 absent. All other mutations were restricted to one mountain range only. The  $\Delta$ -*opn8* and the  
208  $\Psi$ -*tyrp1a* alleles were found in all caves in Sierra de El Abra, while  $\Psi$ -*rh1.2*,  $\Psi$ -*imp1b*,  
209  $\Psi$ -*prph2b* and  $\Psi$ -*gja8a* were found in more restricted regions of Sierra de El Abra. On the  
210 other hand,  $\Psi$ -*per2-M* and  $\Psi$ -*lim2.1* alleles were only found in caves in the North of Sierra  
211 de Guatemala (Jineo + Molino + Escondido). Furthermore, LoF mutations were fixed in  
212 Sierra de Guatemala, while most LoF mutations were at high frequency (> 0.5) but not fixed  
213 in most Sierra de El Abra caves - except in Curva. None of these LoF mutations were found in  
214 234 surface fish or the 26 Subterráneo cavefish.

215

## 216 Eye size polymorphism and surface genome introgression in the Sótano del Toro

217 Finally, we sought to investigate the relationship between the level of troglomorphy,  
218 measured as the relative eye size (*i.e.*, eye diameter divided by the standard length of the fish,  
219 [25]) and the number of LoF alleles (**Additional file 1 and Additional file 2: Fig S4**). This  
220 was done with samples from the Sótano del Toro (simply Toro hereafter), in which  
221 hybridization between cavefish and surface fish is frequent and gives rise to a large variability  
222 of eye sizes [5]. Although the two large deletions described above were segregating in this  
223 cave, they were not included in the analysis because our genotyping method did not allow us  
224 to differentiate between heterozygous and homozygous fish. Thus, Toro cavefish were  
225 genotyped at five loci for which segregating LoF alleles were identified (**Table 1**). To have  
226 more information, we included an additional locus, the vision-associated gene *zgc:153441*  
227 (orthologous to the human gene *retinol dehydrogenase 13*), in which a 2 bp insertion was  
228 identified (**Additional file 1: LoF\_list.xlsx**). A significant negative correlation (Pearson's  $r =$

229 -0.91, p-value = 4.1e-12) was found between the relative eye size and the number of LoF  
230 alleles (**Fig. 3**).

231

## 232 **Discussion**

233

### 234 **Nature and age of pseudogenes segregating in caves**

235 *Astyanax mexicanus* cavefish are a good model to study evolution in caves because they show  
236 many morphological, physiological and behavioral traits corresponding to a cave syndrome,  
237 which is often observed in animals living in this environment. The time needed to evolve  
238 many cave traits, some being adaptive and others resulting from genetic drift under relaxed  
239 selection, is a matter of debate. Several studies of the genetic polymorphism in *A. mexicanus*  
240 indicated a recent origin of cave populations (in a range of 10,000 to 200,000 years),  
241 suggesting that in this species, the cave syndrome could have evolved quickly [7, 8]. To  
242 examine further the consequences of a recent origin, we examined gene decay in several  
243 caves. First in three sets of genes involved in vision, circadian clock and pigmentation, and  
244 then at the whole genome scale. Under the hypothesis of a recent origin, we would expect that  
245 cave-specific pseudogenes would accumulate few LoF mutations. Using 47 cavefish genomes  
246 from four different caves, our results supported this idea, we only discovered three new vision  
247 genes with segregating non-functional alleles, each with only one LoF mutation, and that were  
248 present in one or two caves. In addition, one of these LoF mutations was previously identified  
249 in the Pachón genome assembly [9] and is only found in this population. While many vision  
250 pseudogenes can be found in other cavefish species, *A. mexicanus* data show that few vision  
251 genes are pseudogenized, thus supporting the recent evolution of blindness in this species [9,  
252 18, 19]. Moreover, at a genome-wide scale and without a priori selection of genes involved in  
253 vision, we found an enrichment of vision-related GO-terms among 512 pseudogenes,

254 corroborating previous studies comparing the accumulation of conservative versus radical  
255 mutations in transcripts [26]. In accordance with a recent origin of *A. mexicanus* cavefish,  
256 most pseudogenes contain only one LoF mutation, a few of them two LoF mutations, but  
257 none have more than two LoF mutations [9].

258

### 259 **Origins of *A. mexicanus* cavefish**

260 Another point of discussion concerns the number of independent origins of *A. mexicanus* cave  
261 populations. When they were first discovered, a different species name was given to each of  
262 them (*Astyanax jordani* Hubbs and Innes 1936 from La Cueva Chica, *A. antrobius* Alvarez  
263 1946 from La Cueva de El Pachón, and *A. hubbsi* Alvarez 1947 from La Cueva de Los  
264 Sabinos). However, hybridization in the wild and the lab between surface and cavefish, and  
265 between cavefish from different caves suggested they belong to the same species. More  
266 recently, population genetics studies suggest gene flow and similar ages for cave populations,  
267 casting doubt on numerous independent origins of cave populations [7, 8]. Our analysis of  
268 shared pseudogenes helps to refine recent claims of a reduced number of independent origins.  
269 First, the analysis of the distribution of pseudogenes in three cave populations, at relatively  
270 large geographic distances within the Sierra de El Abra identified many pseudogenes shared  
271 by all caves or cave pairs. This result can be best explained by a common origin resulting  
272 from a single initial colonization event, a hypothesis already proposed based on the analysis  
273 of RAPD markers [27]. Moreover, a long period of subsequent secondary independent  
274 evolution is well supported by the existence of a large number of cave-specific pseudogenes.  
275 While there are fewer pseudogenes in both Sierra de El Abra and Sierra de Guatemala, it is  
276 significant that there are still several shared ones. Globally, we interpret the pattern of shared  
277 pseudogenes as evidence of a common origin of the four populations (Molino, Pachón,  
278 Tinaja, and Chica) from a single colonization event, with early isolation of Sierra de

279 Guatemala caves from Sierra de El Abra caves. Moreover, this pattern of shared pseudogenes  
280 also suggests that the period of independent evolution of the caves is longer than the period of  
281 common evolution. In terms of biogeography, we can assume that one cavefish population  
282 evolved first, spread rapidly from one Sierra to the other, and more recently spread further  
283 within each Sierra.

284 To refine this scenario, we examined twelve LoF mutations in eleven genes involved in  
285 photoreception, circadian clock and pigmentation in our library of DNA samples from 11  
286 different *A. mexicanus* caves. The large deletion in the gene *rgr2* is shared by all cave  
287 populations in Sierra de Guatemala and Sierra de El Abra, but not with Subterráneo in the  
288 Micos area. No LoF mutation identified in Sierra de Guatemala or Sierra de El Abra could be  
289 found in Subterráneo, strongly suggesting that the cave populations in the Micos area could  
290 have a completely distinct origin. Populations that are geohydrologically close to each other  
291 tend to share more markers. For example, in the Sierra de Guatemala, the three populations  
292 geographically very close in the North share two pseudogenes, which are absent in the South  
293 (Caballo Moro).

294 In Sierra de El Abra, while shared by several populations, most pseudogenes are not fixed,  
295 suggesting sufficient gene flow from the surface to limit their fixation in most caves.

296 However, this is not the case in Curva, in which all pseudogenes were found fixed, which is  
297 evidence for the complete isolation of this cave [28].

298 The distribution of cave-specific pseudogenes suggests a separate origin of cavefish in the  
299 Micos area on the one hand and in Sierra de Guatemala + Sierra de El Abra on the other hand.

300 In the latter region, cavefish spread rapidly from a yet unidentified cave in the whole region,  
301 leading to two well-differentiated groups: Sierra de Guatemala caves and Sierra de El Abra  
302 caves. Inside each cluster, particularly within the Sierra de El Abra, the differentiation is low.

303 The location of the first cavity that was colonized before spreading through the underground  
304 aquifer remains unknown.

305 Finally, in Toro, where important gene flow from the surface is present, we observed an  
306 expected correlation between the level of phenotypic troglomorphy and the number of LoF  
307 alleles in the genome. The continuous presence of LoF alleles in Toro despite important  
308 surface gene flow suggests that these LoF alleles are not easily wiped out from the cavefish  
309 genomes under underground specific environmental conditions. Regarding eye loss, the  
310 gradient of eye sizes observed in Toro individuals could be used in future studies to decipher  
311 between genes and alleles involved in controlling eye size (in individuals with small but  
312 existing eyes) and those involved in the degeneration process (in eyeless individuals).

313

#### 314 Remarkable convergence or common origin of cave traits?

315 Many studies have assumed that *Astyanax* cavefish have multiple origins with many  
316 independent colonization and convergent evolution events: 1) Micos area, 2) Sierra de  
317 Guatemala, 3) Central Sierra de El Abra (Sabinos, Tinaja, Piedras and Curva), which belong  
318 to a mitochondrial haplogroup, 4) the other caves within Sierra de El Abra which belong to  
319 another haplogroup, sometimes differentiating Pachón from Chica as yet another separate  
320 evolutionary event. Using this assumption, several studies inferred evolutionary convergences  
321 at the morphological, physiological and behavioral levels. Some were particularly remarkable.  
322 For example, in a review of the data available on events and processes associated with eye  
323 degeneration in five caves (Sabinos, Tinaja, Curva, Pachón and Chica), no difference was found  
324 [29]. In all cavefish, when information was available, the following observations were made:  
325 smaller eye primordium, loss of ventral optic cup, lens apoptosis, continued cell division in the  
326 ciliary marginal zone, eye restoration by lens transplantation, *Hsp90α* activation, *Pax6* down-  
327 regulation in optic fields, and *Hh* expansion at the embryonic midline. These similar phenotypes



328 were assumed to result from convergence and not from a common origin. More recently, it has  
329 been proposed that the repeated evolution of eye loss in Mexican cavefish is evidence of the  
330 convergent evolution of developmental mechanisms in independently evolved Pachón and  
331 Molino populations [30]. Similarly, sleep loss in Molino, Pachón and Tinaja cavefish  
332 populations [31] and the functional and anatomical brain similarities among cavefish from  
333 Molino, Pachón and Tinaja have been interpreted as convergences [32]. Also, evolutionary  
334 convergence was proposed regarding a neural mechanism in Molino, Pachón and Tinaja  
335 cavefish lateral line system [33]. The evolution of similar regressive and constructive  
336 craniofacial traits across distinct Pachón, Tinaja and Chica cavefish populations [34] and the  
337 unique transcriptional signatures of sleep loss across Molino, Pachón and Tinaja cavefish  
338 populations were similarly interpreted. However the authors acknowledged the possibility that  
339 the traits could have evolved in their common ancestors [35]. At a genome scale, the chromatin  
340 architecture of the two cavefish morphotypes, Pachón and Tinaja, were more similar to each  
341 other than to the surface fish. These data have been interpreted as a phenotypic convergence in  
342 independently derived cavefish populations and congruent with loss of eyes and pigment,  
343 accumulation of excess fat, and insulin resistance [36].

344 We think that a common origin more easily explains the similarity of the cavefish syndrome in  
345 different *Astyanax* caves. Indeed, at the geographic scale of the Sierra de El Abra, this  
346 hypothesis was supported by several well-studied cave specific mutations: a *P106L* mutation in  
347 the MAO protein affecting behaviour is present in all seven Sierra de El Abra populations tested  
348 [28]; a *P211L* mutation in *INSRA* protein contributing to hyperglycaemia [37] is found in  
349 Pachón, Yerbaniz, Japonés and Tinaja populations; an *R263Q* deleterious mutation in *CRY1A*  
350 is found in Pachón, Tinaja and Chica [20]. Considering the numerous non-functional alleles  
351 described in the present study, there is little doubt that cavefish in the whole extent of Sierra de  
352 El Abra have a common origin that explains their similar cave syndrome. A common origin

353 does not exclude periods of independent evolution between sub-clusters of caves and within  
354 each cave, nor re-introgessions or mixing of cavefish populations after some time of  
355 independent evolution. Indeed, independent loss-of-function mutations in *oca2* are present in  
356 Molino and Pachón [12], and Pachón and Tinaja share a loss-of-function mutation in *per2* while  
357 another mutation was found in Molino [21]. Such partially independent evolution is also  
358 supported by the many cave-specific, or cave-cluster-specific, non-functional alleles found in  
359 the present study.

360 At the phenotypic level, the partial complementation restoring sight in blind cavefish larvae  
361 after crosses between different cave populations [13], the loss of schooling behaviour that is  
362 very similar between Pachón and Tinaja but shows some differences in Molino [38], or the  
363 complete loss of aggressiveness in Sierra de El Abra populations but less so in Sierra de  
364 Guatemala populations [39, 40] are also in favour of a common origin followed by  
365 independent evolution. Likewise, the observation that, among genes whose rhythmic  
366 expression is lost in at least one of three caves, 22% were arrhythmic in the three caves, was  
367 interpreted as evidence of repeated evolution of circadian clock dysregulation in independent  
368 populations [41]. It could instead suggest a common origin, explaining a large part of shared  
369 gene dysregulation, followed by independent evolution allowing both specific and  
370 “convergent” gene expression dysregulation.

371

## 372 Conclusion

373 Altogether, the distribution of loss-of-function mutations in cavefish populations supports an  
374 independent origin of cavefish in the Micos area on one hand, and Sierra de Guatemala +  
375 Sierra de El Abra on the other hand. Populations in Sierra de Guatemala and Sierra de El Abra  
376 probably diverged early, while populations within Sierra de El Abra diverged later.

377

## 378 **Materials and Methods**

379

### 380 **Variant calling**

381 Genomic reads from 62 *Astyanax mexicanus* fish (9 Molino, 9 Pachón, 10 Tinaja, 19 Chica, 9  
382 Choy and 6 Rascón), one *Astyanax aeneus* individual and one *Gymnocorymbus ternetzi*  
383 individual were retrieved on NCBI (SRA accessions in **Additional file 3:**

384 **SRA\_coverage.xlsx**) and aligned to the *Astyanax mexicanus* surface fish reference genome  
385 (GCA\_000372685.2) using BWA-MEM2 [42]. Duplicated reads were removed using GATK  
386 MarkDuplicates (GATK 4.2.4.1) [43]. Mean read coverage computed with SAMtools v1.15  
387 [44] for each individual can be found in **Additional file 3: SRA\_coverage.xlsx**.

388 Variant calling was performed following GATK Best Practices [45] by applying hard filters  
389 (QD < 2.0 || FS > 60.0 || MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0 ||  
390 ExcessHet > 54.69) for SNPs and (QD < 2.0 || FS > 200.0 || ReadPosRankSum < -20.0 ||  
391 ExcessHet > 54.69) for indels. Invariant sites and sites for which not all *Astyanax mexicanus*  
392 individuals were genotyped were removed. The putative impact of each variant was assessed  
393 using Ensembl Variant Effect Predictor (VEP) [46] using Ensembl v108 *Astyanax mexicanus*  
394 gene annotation.

395 Variants leading to putative loss-of-function (LoF) (*i.e.*, premature stop codons, frameshifts,  
396 loss of expected start or stop codons, splice site mutations) were extracted using bcftools  
397 +split-vep [44]. Only mutations impacting the longest transcript of each gene were retained,  
398 for a total of 11,608 LoF mutations in *Astyanax mexicanus* individuals. To reduce sequencing  
399 artifacts, we only retained 7,220 variants present in at least two individuals. We then kept  
400 2,095 cave-specific LoF mutations by removing variants found in surface *Astyanax mexicanus*  
401 or outgroup species (*Astyanax aeneus* and *Gymnocorymbus ternetzi*).

402

## 403 Removal of artifactual LoF variants

404 As inferred LoF mutations heavily depend on the gene annotation quality, we used MAFFT  
405 v7.467 [47] to align *A. mexicanus* protein sequences (keeping the longest transcript of each  
406 gene) with orthologous sequences of 9 other teleost species retrieved using *Broccoli v1.2* :  
407 *Danio rerio* (GCF\_000002035.6), *Pygocentrus nattereri* (GCF\_015220715.1), *Electrophorus*  
408 *electricus* (GCF\_013358815.1), *Bagarius yarelli* (GCA\_005784505.1), *Tachysurus fulvidraco*  
409 (GCF\_003724035.1), *Ameiurus melas* (GCA\_012411365.1), *Triplophysa tibetana*  
410 (GCA\_008369825.1), *Anabarilius grahami* (GCA\_003731715.1), *Labeo rohita*  
411 (GCA\_004120215.1).

412 Conserved regions of these alignments, *i.e.*, gene parts that are consistently annotated in the  
413 different genome assemblies [48, 49], were defined using trimAl v1.4.1 [50]. We kept 933  
414 LoF variants located in these conserved regions in *A. mexicanus* and removed genes for which  
415 the ensemble annotation lacked a methionine start codon.  
416 We also removed premature stop codons and frameshifts located in the last 5% of the genes  
417 and consecutive frameshifts restoring the coding frame, for a final dataset of 533 cave-  
418 specific LoF variants located on 512 genes.

419

## 420 Distribution of LoF mutations in vision, circadian clock and pigmentation genes

421 Coding sequences and genomic locations of 86 vision, 38 circadian clock and 249  
422 pigmentation genes in the surface *Astyanax mexicanus* genome were retrieved from a  
423 previous study [9]. For each of these genes, the corresponding Ensembl gene ID was retrieved  
424 on Ensembl (release 109) using a combination of blastn and the Ensembl genome browser  
425 (**Additional file 1**). Five genes were not retrieved on the Ensembl annotation: one vision gene  
426 (*cryaa*) and four pigmentation genes (*mc1r2*, *mgn1a*, *shroom2a*, *ap3d1*). Furthermore, one  
427 pigmentation gene was annotated as a pseudogene on Ensembl (*mc1r1*) (**Additional file 1**).

428 Vision, circadian clock and pigmentation pseudogenes were then defined using the  
429 intersection between this Ensembl gene ID list and the Ensembl gene ID of the 512 genome-  
430 wide identified pseudogenes.

431

### 432 GO-term enrichment of cave pseudogenes

433 For each pseudogene, we retrieved *Danio rerio* and *Homo sapiens* orthologous ensembl gene  
434 name and Ensembl gene ID using the R package “biomaRt” v3.15 [51]. A total of 243  
435 orthologous genes found in humans were used as input in the DAVID database  
436 (<https://david.ncifcrf.gov/>, [52]) and in EnrichR [53] to investigate diseases enrichment in the  
437 OMIM and ClinVar 2019 databases. In zebrafish, we found 335 orthologous genes, and we  
438 looked for enriched terms in the “GO Biological Process GeneRIF Predicted Z-score” and  
439 “Phenotype GeneRIF Predicted Z-score” databases implemented in FishEnrichR [53].

440

### 441 Characterization of large deletions

442 In a previous study [9], we identified two large deletions in a Pachón cavefish genome  
443 assembly, one leading to a truncated *opn8a* gene and the absence of *opn8b* gene, the other  
444 leading to a truncated *rgr2* gene. Here, we took advantage of a new Pachón genome assembly  
445 (GCA\_019721115.1) to verify and further characterize these deletions.

446 The genes *opn8a* (LOC111191907) and *opn8b* (LOC111197225) are located in tandem in the  
447 surface fish genome (in the genomic region NC\_035899.1:44673130-44727205) and are  
448 flanked by the genes *runx2b* (LOC103028241, NC\_035899.1:44525886-44575275) upstream  
449 and *mmut* (LOC103027932, NC\_035899.1:44737167-44760205) downstream. We used  
450 *runx2b* and *mmut* protein sequences as tblastn [54] queries against the Pachón genome  
451 assembly to identify the corresponding region (CM033892.1:72415284-72611085). This

452 region was extracted using samtools faidx [44], and its alignment with the surface fish region  
453 was visualized using dotter [55].

454 The gene *rgr2* (LOC103043895) is located on the genomic region NW\_019172852.1:462710-  
455 472903 in the surface genome assembly. The corresponding region in the Pachón genome  
456 assembly (CM033906.1:17152181-17158392) was extracted by using *rgr2* as a query in a  
457 tblastn against this genome. The Surface and Pachón regions were also aligned and visualized  
458 using dotter. The frequencies of these two large deletions were inferred for Molino, Tinaja,  
459 Pachón and Chica cave populations by computing the read alignment coverage in these  
460 regions for each surface and cave individual using bamCoverage [56].

461

#### 462 **Frequencies of 10 LoF mutations and 2 large deletions in 11 cave populations**

463 For each cavefish or surface fish genotyped, DNA fragments containing small indels (< 7 bp)  
464 leading to frameshifts or point mutations leading to premature stop codons were amplified by  
465 PCR (primer list in **Additional file 4: primer\_list.xlsx**). Amplicon sequencing allowed us to  
466 determine if a LoF mutation was absent, or, if present, if the fish was heterozygous or  
467 homozygous for the variant (**Additional file 2: Fig S5**). For larger indels, amplification by  
468 PCR followed by a migration of amplicons on a 3% agarose gel electrophoresis allowed us to  
469 determine if an indel was absent, or, if present, if the fish was heterozygous or homozygous  
470 (**Additional file 2: Fig S5**).

471 We also genotyped fish for the two large deletions by performing two PCR for each deletion.  
472 First, a control PCR was made using primers in a non-deleted exon of *opn8a* (exon 1) or *rgr2*  
473 (exon 3) (**Additional file 4: primer\_list.xlsx**, **Additional file 2: Fig S6**). A second PCR was  
474 made using a primer in putative deleted exons (exon 4 of *opn8a* or exon 5 of *rgr2*). If the  
475 second PCR led to an amplification, we inferred that the individual had at least one non-  
476 deleted allele, while if there was no amplification, the individual was considered homozygous

477 for the deletion. This strategy did not allow us to differentiate between non-deleted  
478 homozygous and heterozygous individuals. In such case, when only one homozygous state is  
479 identifiable (here homozygosity for a deletion), we have to assume random mating to estimate  
480 the frequency of this allele in a population, which is the square root of the frequency of  
481 homozygotes.

482

### 483 Phenotype and genotype scoring of Toro cavefish

484 A troglomorphism score (eye diameter normalized by the standard length of the fish,  
485 measured in pixels using *ImageJ* [57]) was assigned to 30 wild Toro cavefish individuals  
486 photographed in the field. In this cave, cavefish cohabit with surface fish and hybrids. Eyeless  
487 specimens thus had a troglomorphism score of 0. In parallel, a genomic score was assigned to  
488 each fish: for each of the five LoF mutations found in this population, the score was set to 2 if  
489 the fish was homozygous for the mutation, 1 if it was heterozygous, 0 if it did not carry this  
490 the mutation. To increase statistical power, fish were genotyped at a sixth locus, a vision-  
491 related gene (*zgc:15531*) for which we identified a segregating LoF variant in the Toro  
492 population (**Additional file 1, Additional file 4: primer\_list.xlsx, Additional Fig S5**).  
493 Scores at the six loci were summed to get a total score between 0 and 12, which was used as a  
494 proxy of the proportion of the cavefish genome (0 for surface fish and 12 for cavefish).

495

### 496 Declaration

497

### 498 Ethics approval and consent to participate

499 Not applicable

500

### 501 Consent for publication

502 Not applicable

503

### 504 **Availability of data and materials**

505 A bash script describing read alignment steps and variant calling, as well as a Rscript used for  
506 data analysis and related files are provided in GitHub

507 (<https://github.com/MaximePolicarpo/non-functional-alleles-Astyanax-mexicanus>)

508

### 509 **Competing interests**

510 The authors declare that they have no competing interests.

511

### 512 **Funding**

513 This work was supported by collaborative grants from Agence Nationale de la Recherche  
514 (BLINDTEST to S.R. and D.C.), Institut Diversité Écologie et Évolution du Vivant (to S.R.  
515 and D.C.) and an Ecos-Nord Franco-Mexican exchange program (to S.R.).

516

### 517 **Authors' contributions**

518 M.P., S.R. and D.C. conceived and designed the research; M.P., L.L., L.E. and S.R. conducted  
519 field surveys and collected samples; M.P., P.L. and I.G. performed molecular work; M.P.,  
520 S.R. and D.C. analysed data; M.P., L.L., L.E., S.R. and D.C. wrote the manuscript. All authors  
521 read and approved the final version of the manuscript.

522

### 523 **Acknowledgments**

524 We thank all Rétaux's group members for their collaborative spirit during fieldwork.

525



526 **Sampling authorization**

527 *Astyanax mexicanus*: small fin clips from surface and cave morphs of *A. mexicanus* were  
528 sampled during 4 field expeditions between March 2011 and March 2017, under the auspices  
529 of the collecting permits 02241/13 (delivered to Ernesto Maldonado and S.R.), 02438/16 and  
530 05389/17 (delivered to Patricia Ornelas-Garcia and S.R.) by the SEMARNAT (Secretaria de  
531 medio ambiente y recursos naturales) of Mexico.

532

533 **Supplementary Material**

534 Supplementary data are available at BMC Ecology and Evolution

535

536

537

538

539 **Table 1.** Pseudogene frequencies in four caves among vision, circadian clock and  
 540 pigmentation genes

Pseudogenes	Gene sets	LoF mutations	Caves			
			Molino (9)	Pachón (9)	Tinaja (10)	Chica (19)
<i>Ψ-sagb</i>	vision	Deletion		0.39	0.05	
<i>Ψ-cryba2a</i>	vision	premature stop			0.40	0.39
<i>Ψ-pde6b</i>	vision	deletion		0.67		
<i>Ψ-rh1.2</i>	vision	premature stop			0.35	
<i>Ψ-per2</i>	circadian clock	insertion	1.00			
<i>Ψ-kif13a</i>	pigmentation	insertion			0.25	
<i>Ψ-tyrp1a</i>	pigmentation	deletion		0.94	0.85	0.58
<i>Ψ-csf1rb</i>	pigmentation	start loss		0.44	0.10	
<i>Ψ-cd63</i>	pigmentation	premature stop				0.26
<i>Ψ-drd2b</i>	pigmentation	deletion				0.11
<i>Ψ-hps4</i>	pigmentation	premature stop	1.00			

541 Notes: sample size indicated between parenthesis.

542

543

544 **Table 2.** Frequencies of 12 pseudogenes in 11 cave populations and surface fish

		<i>Pseudogenes</i>												
	L	n	<i>Δ-trgr2</i>	<i>Δ-opn8</i>	<i>Ψ-rh1.2</i>	<i>Ψ-cryba2a</i>	<i>Ψ-impglb</i>	<i>Ψ-prph2b</i>	<i>Ψ-gia8a</i>	<i>Ψ-lim2.1</i>	<i>Ψ-pde6b</i>	<i>Ψ-per2-M</i>	<i>Ψ-per2-T</i>	<i>Ψ-tyrp1a</i>
Jineo	G	6	1.00							1.00		1.00		
Molino	G	9	1.00							1.00		1.00		
Escondido	G	7	1.00							1.00		1.00		
Caballo Moro	G	2	1.00											
Pachón	A	55	0.95	0.66	0.04		0.24	0.15			0.76			0.69
Sabinos	A	33	0.83	1.00	0.24	0.50	0.53	1.00	0.88					1.00
Tinaja	A	16	1.00	0.87	0.44	0.16	0.56	0.94	0.69				0.13	0.88
Curva	A	23	1.00	1.00	1.00		1.00	1.00	1.00					1.00
Toro	A	30	0.75	0.73		0.65	0.18	0.75	0.65					0.05
Chica	A	33	0.70	0.52		0.53								0.47
Subterráneo	M	26												
Surface		234												

545 Notes: L = location, G = Sierra de Guatemala, A = Sierra de El Abra, M = Micos area; n = sample size.

546

547

548 **References**

549

- 550 1. Jeffery WR: **Astyanax surface and cave fish morphs**. *EvoDevo* 2020, **11**(1):14.
- 551 2. Elliott W: **The Astyanax caves of Mexico. Cavefishes of Tamaulipas, San Luis**  
552 **Potosi, and Guerrero**. *Association for Mexican Cave Studies Bulletin* 26 2018:1-325.
- 553 3. Mitchell RW, Russell WH, Elliott WR: **Mexican eyeless characin fishes, genus**  
554 **Astyanax: environment, distribution and evolution**. *Spec Publ Mus Texas Techn*  
555 *University* 1977, **12**:1-89.
- 556 4. Espinasa L, Legendre L, Fumey J, Blin M, Rétaux S, Espinasa M: **A new cave**  
557 **locality for Astyanax cavefish in Sierra de El Abra, Mexico**. *Subterranean Biology*  
558 2018, **26**:39-53.
- 559 5. Espinasa L, Ornelas-García CP, Legendre L, Rétaux S, Best A, Gamboa-Miranda R,  
560 Espinosa-Pérez H, Sprouse P: **Discovery of Two New Astyanax Cavefish Localities**  
561 **Leads to Further Understanding of the Species Biogeography**. *Diversity* 2020,  
562 **12**(10):368.
- 563 6. Miranda-Gamboa R, Espinasa L, Verde-Ramírez MdlA, Hernández-Lozano J, Lacaille  
564 JL, Espinasa M, Ornelas-García CP: **A new cave population of Astyanax mexicanus**  
565 **from Northern Sierra de El Abra, Tamaulipas, Mexico**. *Subterranean Biology*  
566 2023, **45**:95-117.
- 567 7. Herman A, Brandvain Y, Weagley J, Jeffery WR, Keene AC, Kono TJY, Bilandžija  
568 H, Borowsky R, Espinasa L, O'Quin K *et al*: **The role of gene flow in rapid and**  
569 **repeated evolution of cave-related traits in Mexican tetra, Astyanax mexicanus**.  
570 *Molecular ecology* 2018, **27**(22):4397-4416.

- 571 8. Fumey J, Hinaux H, Noirot C, Thermes C, Rétaux S, Casane D: **Evidence for late**  
572 **Pleistocene origin of *Astyanax mexicanus* cavefish.** *BMC Evolutionary Biology*  
573 2018, **18**(1):43.
- 574 9. Policarpo M, Fumey J, Lafargeas P, Naquin D, Thermes C, Naville M, Dechaud C,  
575 Volff J-N, Cabau C, Klopp C *et al*: **Contrasting gene decay in subterranean**  
576 **vertebrates: insights from cavefishes and fossorial mammals.** *Molecular Biology*  
577 *and Evolution* 2021, **38**(2):589-605.
- 578 10. Casane D, Rétaux S: **Evolutionary Genetics of the Cavefish *Astyanax mexicanus*.**  
579 *Advances in Genetics* 2016, **95**:117-159.
- 580 11. Ma L, Gore AV, Castranova D, Shi J, Ng M, Tomins KA, van der Weele CM,  
581 Weinstein BM, Jeffery WR: **A hypomorphic cystathionine  $\beta$ -synthase gene**  
582 **contributes to cavefish eye loss by disrupting optic vasculature.** *Nature*  
583 *Communications* 2020, **11**(1):2772.
- 584 12. Protas ME, Hersey C, Kochanek D, Zhou Y, Wilkens H, Jeffery WR, Zon LI,  
585 Borowsky R, Tabin CJ: **Genetic analysis of cavefish reveals molecular convergence**  
586 **in the evolution of albinism.** *Nat Genet* 2006, **38**(1):107-111.
- 587 13. Borowsky R: **Restoring sight in blind cavefish.** *Curr Biol* 2008, **18**(1):R23-24.
- 588 14. García-Machado E, Hernandez D, Garcia-Debras A, Chevalier-Monteagudo P,  
589 Metcalfe C, Bernatchez L, Casane D: **Molecular phylogeny and phylogeography of**  
590 **the Cuban cave-fishes of the genus *Lucifuga*: evidence for cryptic allopatric**  
591 **diversity.** *Mol Phylogenet Evol* 2011, **61**(2):470-483.
- 592 15. Hernández D, Casane D, Chevalier-Monteagudo P, Bernatchez L, García-Machado E:  
593 **Go West: A One Way Stepping-Stone Dispersion Model for the Cavefish**  
594 ***Lucifuga dentata* in Western Cuba.** *PLOS ONE* 2016, **11**(4):e0153545.

- 595 16. Hernández D, Møller PR, Casane D, García-Machado E: **A new species of the cave-**  
596 **fish genus *Lucifuga* (Ophidiiformes, Bythitidae), from eastern Cuba.** *ZooKeys*  
597 2020, **946**:17-35.
- 598 17. Malard F, Machado EG, Casane D, Cooper S, Fišer C, Eme D: **Dispersal and**  
599 **geographic range size in groundwater.** In: *Groundwater Ecology and Evolution*  
600 *(Second Edition)*. Edited by Malard F, Griebler C, Rétaux S. San Diego: Academic  
601 Press; 2023: 185-207.
- 602 18. Policarpo M, Laurenti P, García-Machado E, Metcalfe CJ, Rétaux S, Casane D:  
603 **Genomic evidence that blind cavefishes are not wrecks of ancient life.** *BioRxiv*  
604 2021.
- 605 19. Casane D, Saclier N, Policarpo M, François C, Lefébure T: **Evolutionary genomics**  
606 **and transcriptomics in groundwater animals.** In: *Groundwater Ecology and*  
607 *Evolution (Second Edition)*. Edited by Malard F, Griebler C, Rétaux S. San Diego:  
608 Academic Press; 2023: 393-412.
- 609 20. Moran RL, Jaggard JB, Roback EY, Kenzior A, Rohner N, Kowalko JE, Ornelas-  
610 García CP, McGaugh SE, Keene AC: **Hybridization underlies localized trait**  
611 **evolution in cavefish.** *iScience* 2022, **25**(2):103778.
- 612 21. Xiong S, Wang W, Kenzior A, Olsen L, Krishnan J, Persons J, Medley K, Peuß R,  
613 Wang Y, Chen S *et al*: **Enhanced lipogenesis through Ppar $\gamma$  helps cavefish adapt**  
614 **to food scarcity.** *Current Biology* 2022, **32**(10):2272-2280.e2276.
- 615 22. Carlson BM, Klingler IB, Meyer BJ, Gross JB: **Genetic analysis reveals candidate**  
616 **genes for activity QTL in the blind Mexican tetra, *Astyanax mexicanus*.** *PeerJ*  
617 2018, **6**:e5189.
- 618 23. Warren WC, Boggs TE, Borowsky R, Carlson BM, Ferrufino E, Gross JB, Hillier L,  
619 Hu Z, Keene AC, Kenzior A *et al*: **A chromosome-level genome of *Astyanax***

- 620 ***mexicanus* surface fish for comparing population-specific genetic differences**  
621 **contributing to trait evolution.** *Nature Communications* 2021, **12**(1):1447.
- 622 24. Morimura H, Saindelle-Ribeaudeau F, Berson EL, Dryja TP: **Mutations in RGR,**  
623 **encoding a light-sensitive opsin homologue, in patients with retinitis pigmentosa.**  
624 *Nature Genetics* 1999, **23**(4):393-394.
- 625 25. Hinaux H, Pottin K, Chalhoub H, Pere S, Elipot Y, Legendre L, Retaux S: **A**  
626 **developmental staging table for *Astyanax mexicanus* surface fish and Pachon**  
627 **cavefish.** *Zebrafish* 2011, **8**(4):155-165.
- 628 26. Hinaux H, Poulain J, Da Silva C, Noirot C, Jeffery WR, Casane D, Retaux S: **De novo**  
629 **sequencing of *Astyanax mexicanus* surface fish and Pachon cavefish**  
630 **transcriptomes reveals enrichment of mutations in cavefish putative eye genes.**  
631 *PLoS One* 2013, **8**(1):e53553.
- 632 27. Espinasa L, Borowsky RB: **Origins and relationship of cave populations of the**  
633 **blind Mexican tetra, *Astyanax fasciatus*, in the Sierra de El Abra.** *Environmental*  
634 *Biology of Fishes* 2001, **62**(1-3):233-237.
- 635 28. Pierre C, Pradère N, Froc C, Ornelas-García P, Callebert J, Rétaux S: **A mutation in**  
636 **monoamine oxidase (MAO) affects the evolution of stress behavior in the blind**  
637 **cavefish *Astyanax mexicanus*.** *Journal of Experimental Biology* 2020,  
638 **223**(18):jeb226092.
- 639 29. Jeffery WR: **Adaptive evolution of eye degeneration in the Mexican blind cavefish.**  
640 *J Hered* 2005, **96**(3):185-196.
- 641 30. Sifuentes-Romero I, Ferrufino E, Thakur S, Laboissonniere LA, Solomon M, Smith  
642 CL, Keene AC, Trimarchi JM, Kowalko JE: **Repeated evolution of eye loss in**  
643 **Mexican cavefish: Evidence of similar developmental mechanisms in**

- 644 **independently evolved populations.** *Journal of Experimental Zoology Part B:*  
645 *Molecular and Developmental Evolution* 2020, **334**(7-8):423-437.
- 646 31. Duboué ER, Keene AC, Borowsky RL: **Evolutionary convergence on sleep loss in**  
647 **cavefish populations.** *Curr Biol* 2011, **21**(8):671-676.
- 648 32. Jaggard JB, Lloyd E, Yuiska A, Patch A, Fily Y, Kowalko JE, Appelbaum L, Duboue  
649 ER, Keene AC: **Cavefish brain atlases reveal functional and anatomical**  
650 **convergence across independently evolved populations.** *Science Advances* 2020,  
651 **6**(38):eaba3126.
- 652 33. Lunsford ET, Paz A, Keene AC, Liao JC: **Evolutionary convergence of a neural**  
653 **mechanism in the cavefish lateral line system.** *eLife* 2022, **11**:e77387.
- 654 34. Powers AK, Berning DJ, Gross JB: **Parallel evolution of regressive and**  
655 **constructive craniofacial traits across distinct populations of *Astyanax mexicanus***  
656 **cavefish.** *Journal of Experimental Zoology Part B: Molecular and Developmental*  
657 *Evolution* 2020, **334**(7-8):450-462.
- 658 35. McGaugh SE, Passow CN, Jaggard JB, Stahl BA, Keene AC: **Unique transcriptional**  
659 **signatures of sleep loss across independently evolved cavefish populations.**  
660 *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*  
661 2020, **n/a**(n/a):1-14.
- 662 36. Krishnan J, Seidel CW, Zhang N, Singh NP, VanCampen J, Peuß R, Xiong S, Kenzior  
663 A, Li H, Conaway JW *et al*: **Genome-wide analysis of cis-regulatory changes**  
664 **underlying metabolic adaptation of cavefish.** *Nature Genetics* 2022, **54**(5):684-693.
- 665 37. Riddle MR, Aspiras AC, Gaudenz K, Peuß R, Sung JY, Martineau B, Peavey M, Box  
666 AC, Tabin JA, McGaugh S *et al*: **Insulin resistance in cavefish as an adaptation to**  
667 **a nutrient-limited environment.** *Nature* 2018, **555**:647.

- 668 38. Patch A, Paz A, Holt KJ, Duboué ER, Keene AC, Kowalko JE, Fily Y: **Kinematic**  
669 **analysis of social interactions deconstructs the evolved loss of schooling behavior**  
670 **in cavefish.** *PLOS ONE* 2022, **17**(4):e0265894.
- 671 39. Elipot Y, Hinaux H, Callebert J, Rétaux S: **Evolutionary shift from fighting to**  
672 **foraging in blind cavefish through changes in the serotonin network.** *Curr Biol*  
673 2013, **23**(1):1-10.
- 674 40. Espinasa L, Collins E, Ornelas García CP, Rétaux S, Rohner N, Rutkowski J:  
675 **Divergent evolutionary pathways for aggression and territoriality in *Astyanax***  
676 **cavefish.** *Subterranean Biology* 2022, **43**:169-183.
- 677 41. Mack KL, Jaggard JB, Persons JL, Roback EY, Passow CN, Stanhope BA, Ferrufino  
678 E, Tsuchiya D, Smith SE, Slaughter BD *et al*: **Repeated evolution of circadian clock**  
679 **dysregulation in cavefish populations.** *PLOS Genetics* 2021, **17**(7):e1009642.
- 680 42. Md V, Misra S, Li H, Aluru S: **Efficient Architecture-Aware Acceleration of BWA-**  
681 **MEM for Multicore Systems.** In: *2019 IEEE International Parallel and Distributed*  
682 *Processing Symposium (IPDPS): 20-24 May 2019* 2019. 314-324.
- 683 43. Van der Auwera GA, O'Connor BD: **Genomics in the cloud: using Docker, GATK,**  
684 **and WDL in Terra:** O'Reilly Media; 2020.
- 685 44. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A,  
686 Keane T, McCarthy SA, Davies RM *et al*: **Twelve years of SAMtools and BCFtools.**  
687 *GigaScience* 2021, **10**(2):giab008.
- 688 45. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis  
689 AA, del Angel G, Rivas MA, Hanna M *et al*: **A framework for variation discovery**  
690 **and genotyping using next-generation DNA sequencing data.** *Nature Genetics*  
691 2011, **43**(5):491-498.



- 692 46. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P,  
693 Cunningham F: **The Ensembl Variant Effect Predictor**. *Genome Biology* 2016,  
694 **17(1):122**.
- 695 47. Katoh K, Standley DM: **MAFFT Multiple Sequence Alignment Software Version**  
696 **7: Improvements in Performance and Usability**. *Molecular Biology and Evolution*  
697 2013, **30(4):772-780**.
- 698 48. Khenoussi W, Vanhoutrève R, Poch O, Thompson JD: **SIBIS: a Bayesian model for**  
699 **inconsistent protein sequence estimation**. *Bioinformatics* 2014, **30(17):2432-2439**.
- 700 49. Meyer C, Scalzitti N, Jeannin-Girardon A, Collet P, Poch O, Thompson JD:  
701 **Understanding the causes of errors in eukaryotic protein-coding gene prediction:**  
702 **a case study of primate proteomes**. *BMC Bioinformatics* 2020, **21(1):513**.
- 703 50. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T: **trimAl: a tool for automated**  
704 **alignment trimming in large-scale phylogenetic analyses**. *Bioinformatics* 2009,  
705 **25(15):1972-1973**.
- 706 51. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W:  
707 **BioMart and Bioconductor: a powerful link between biological databases and**  
708 **microarray data analysis**. *Bioinformatics* 2005, **21(16):3439-3440**.
- 709 52. Sherman BT, Hao M, Qiu J, Jiao X, Baseler MW, Lane HC, Imamichi T, Chang W:  
710 **DAVID: a web server for functional enrichment analysis and functional**  
711 **annotation of gene lists (2021 update)**. *Nucleic Acids Research* 2022,  
712 **50(W1):W216-W221**.
- 713 53. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A:  
714 **Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool**.  
715 *BMC Bioinformatics* 2013, **14(1):128**.

- 716 54. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL:  
717 **BLAST+: architecture and applications.** *BMC Bioinformatics* 2009, **10**(1):421.
- 718 55. Sonnhammer ELL, Durbin R: **A dot-matrix program with dynamic threshold**  
719 **control suited for genomic DNA and protein sequence analysis.** *Gene* 1995,  
720 **167**(1):GC1-GC10.
- 721 56. Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S,  
722 Dündar F, Manke T: **deepTools2: a next generation web server for deep-**  
723 **sequencing data analysis.** *Nucleic Acids Research* 2016, **44**(W1):W160-W165.
- 724 57. Schneider CA, Rasband WS, Eliceiri KW: **NIH Image to ImageJ: 25 years of image**  
725 **analysis.** *Nature Methods* 2012, **9**(7):671-675.

726

727

728 **Figure legends**

729

730 **Fig. 1** Sampling localities. Cave population names are indicated on the map. Surface fish  
731 localities: (A) Pozo de Pachón Praxedis Guerrero, (B) Arroyo Tampemole El Barranco, (C)  
732 Arroyo El Salvador El Salvador, (D) Arroyo El Limoncito Puente C85, (E) Arroyo La Pagua  
733 Otates, (F) Nacimiento del Río Choy, (G) Nacimiento del Arroyo El Nacimiento, (H) Río  
734 Gallinas Puente Sur Rascón, (I) Estanque Entrada Camino Taninul, (J) Río Tampaon La  
735 Fortaleza, (K) Estanque del Cocodrilo Taninul, (L) Arroyo Puente de Dios Tamasopo, (M)  
736 Río Tampaon Puente Tanchachin. In red the name of cavefish populations for which genomic  
737 data were used.

738

739 **Fig. 2** Distribution of non-functional alleles in four cave populations.

740 A - Venn diagram using all available genomes (sample size in parenthesis). B - Using 9  
741 genomes per population.

742

743 **Fig. 3** Correlation between the phenotype and genotypic score in Sótano del Toro.



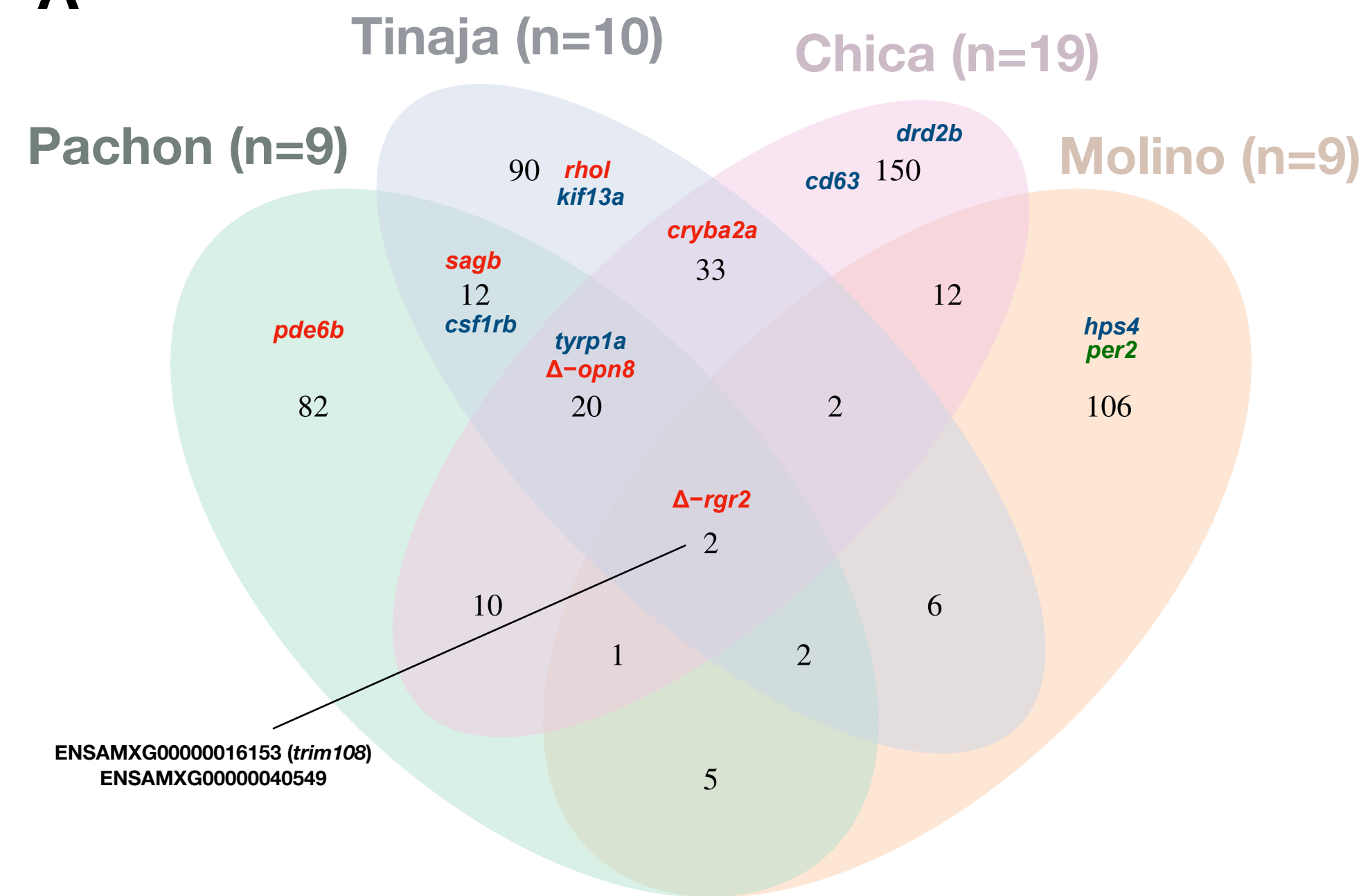
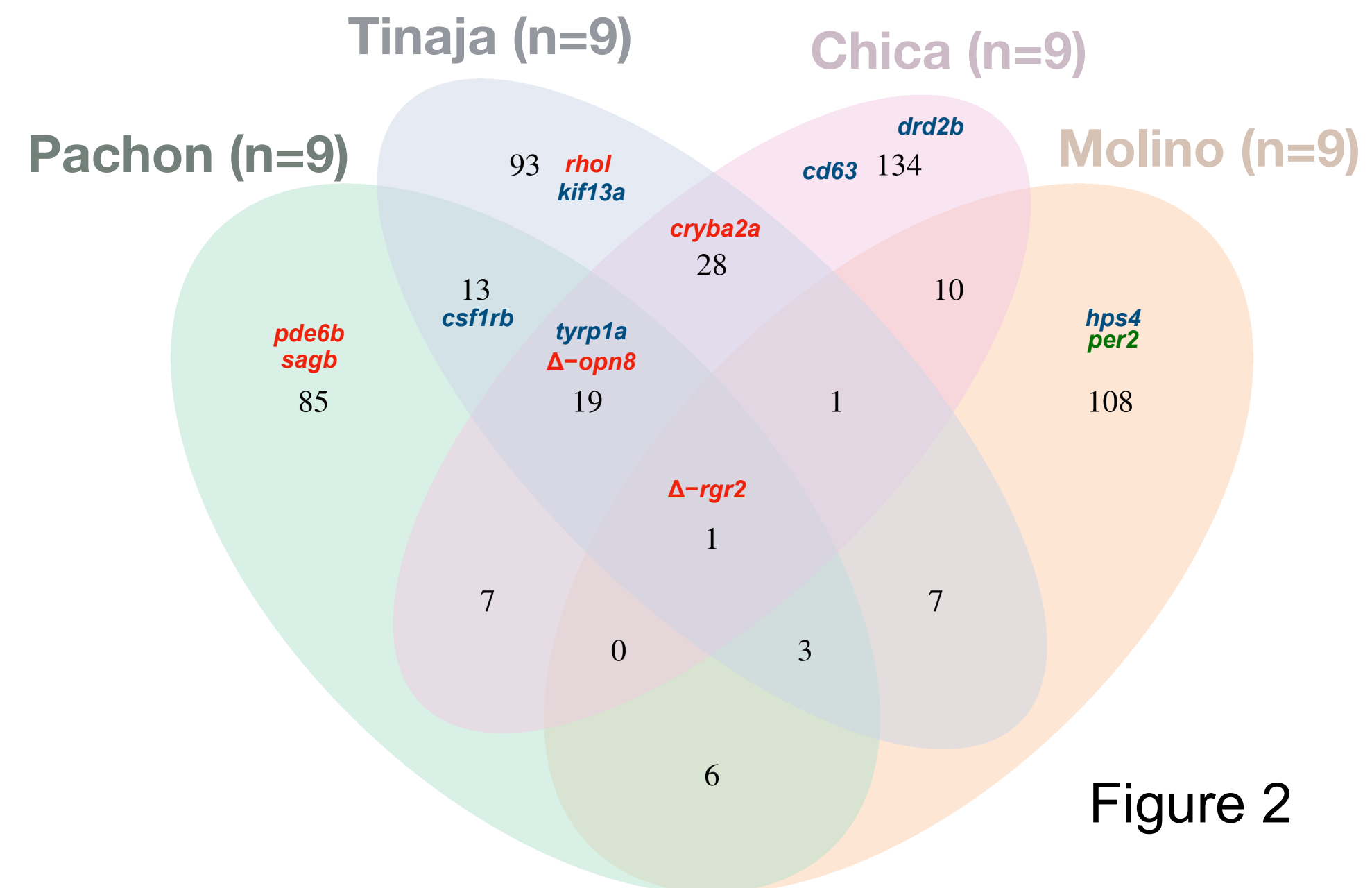
**A****B**

Figure 2

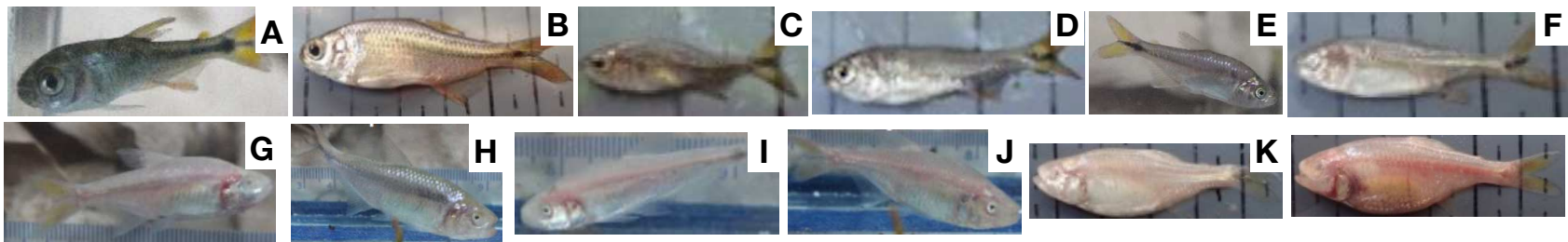
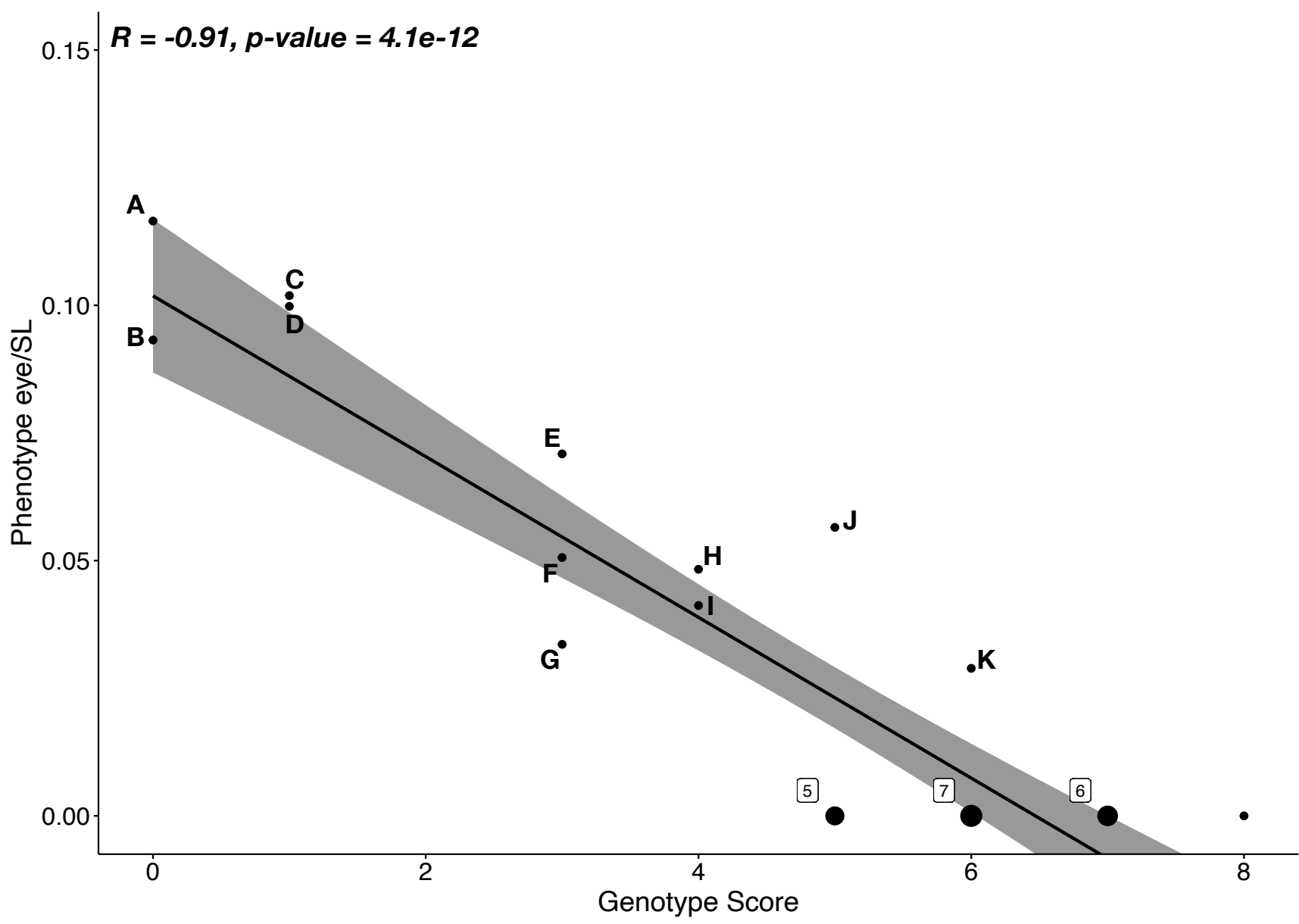


Figure 3

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Additionalfile1LoFlist.xlsx](#)
- [Additionalfile2.pdf](#)
- [Additionalfile3SRACoverage.xlsx](#)
- [Additionalfile4primerlist.xlsx](#)
- [Additionalfilelegends.pdf](#)