

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)

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.

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

Maxime Policarpo^{1,†}, Laurent Legendre¹, Isabelle Germon¹, Philippe Lafargeas¹, Luis Espinasa², Sylvie Rétaux^{3,*} and Didier Casane^{1,4,*}

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

² School of Science, Marist College, Poughkeepsie, New York, USA.

³ Institut de Neurosciences Paris-Saclay, Université Paris-Saclay and CNRS, 91400 Saclay, France.

⁴ Université Paris Cité, UFR Sciences du Vivant, 75013 Paris, France.

[†]: present address: Zoological Institute, Department of Environmental Sciences, University of Basel, Basel, Switzerland.

* Corresponding authors: e-mails: sylvie.retaux@cnrs.fr; didier.casane@universite-paris-saclay.fr.

Abstract

Background

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

Results

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

Conclusions

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

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

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

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

Background

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

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

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

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

Results

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

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

Genome-wide analysis of the distribution of LoF mutations

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

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

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

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

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

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

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

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

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

We next wished to extend our analysis of the distribution of LoF mutations to additional cavefish populations distributed in the 3 mountain ranges (**Fig. 1**), for which we have collected DNA samples. These included 11 caves located in Sierra de Guatemala [Jineo (n = 6 individuals), Molino (n = 9), Escondido (n = 7), Caballo moro (n = 2)], in Sierra de El Abra [Pachón (n = 55), Sabinos (n = 33), Tinaja (n = 16), Curva (n = 23), Toro (n = 30), Chica (n = 33)] and in Micos area [Subterráneo (n = 26)], together with 234 surface fish sampled in local rivers and ponds (**Fig. 1**). In addition to the two large deletions described above in two non-visual opsins (*opn8* and *rgr2*), we selected 10 LoF mutations in nine genes of particular interest: seven genes related to vision (*rh1.2*, *pde6b*, *cryba2a*, *imp1b*, *prph2b*, *gja8a*, *lim2.1*), one to pigmentation (*tyrp1a*) and one to the circadian clock (*per2*). For these 10 LoF mutations, mutation type and position can be found in **Additional file 1: LoF_list.xlsx**. To note, two independent LoF mutations were found in *per2*, one specific to Molino (called Ψ -*per2*-M, first identified with the genomic approach, see **Table 1** and **Table 2**) and one specific to Tinaja (called Ψ -*per2*-T). As the latter was present in only 1 individual in the genomic data set, it was not identified initially with the genomic approach using the threshold ‘present in at-least two individuals’. Mutation frequencies estimated in Molino, Pachón, Tinaja and Chica populations from our wild-caught samples were similar to those inferred from genomic data, suggesting that both estimates were reliable (**Additional file 2: Fig S3**). The geographic distribution of LoF mutations was very variable, from ‘specific to one cave’ to ‘present in most caves’. Moreover, when found in several caves, LoFs primarily distributed in a geographical cluster of caves instead of being scattered in the whole cavefish distribution area (**Table 2**). Two pseudogenes were specific to one cave: Ψ -*pde6b* in Pachón and Ψ -*per2*-T in Tinaja. Other pseudogenes showed larger geographical areas, while none of the

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

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

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

-0.91, p -value = 4.1×10^{-12}) was found between the relative eye size and the number of LoF alleles (**Fig. 3**).

Discussion

Nature and age of pseudogenes segregating in caves

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

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

Origins of *A. mexicanus* cavefish

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

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

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

Remarkable convergence or common origin of cave traits?

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

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

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

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

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

Conclusion

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

Materials and Methods

Variant calling

Genomic reads from 62 *Astyanax mexicanus* fish (9 Molino, 9 Pachón, 10 Tinaja, 19 Chica, 9 Choy and 6 Rascón), one *Astyanax aeneus* individual and one *Gymnocorymbus ternetzi* individual were retrieved on NCBI (SRA accessions in **Additional file 3: SRA_coverage.xlsx**) and aligned to the *Astyanax mexicanus* surface fish reference genome (GCA_000372685.2) using BWA-MEM2 [42]. Duplicated reads were removed using GATK MarkDuplicates (GATK 4.2.4.1) [43]. Mean read coverage computed with SAMtools v1.15 [44] for each individual can be found in **Additional file 3: SRA_coverage.xlsx**. Variant calling was performed following GATK Best Practices [45] by applying hard filters (QD < 2.0 || FS > 60.0 || MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0 || ExcessHet > 54.69) for SNPs and (QD < 2.0 || FS > 200.0 || ReadPosRankSum < -20.0 || ExcessHet > 54.69) for indels. Invariant sites and sites for which not all *Astyanax mexicanus* individuals were genotyped were removed. The putative impact of each variant was assessed using Ensembl Variant Effect Predictor (VEP) [46] using Ensembl v108 *Astyanax mexicanus* gene annotation. Variants leading to putative loss-of-function (LoF) (*i.e.*, premature stop codons, frameshifts, loss of expected start or stop codons, splice site mutations) were extracted using bcftools +split-vep [44]. Only mutations impacting the longest transcript of each gene were retained, for a total of 11,608 LoF mutations in *Astyanax mexicanus* individuals. To reduce sequencing artifacts, we only retained 7,220 variants present in at least two individuals. We then kept 2,095 cave-specific LoF mutations by removing variants found in surface *Astyanax mexicanus* or outgroup species (*Astyanax aeneus* and *Gymnocorymbus ternetzi*).

Removal of artifactual LoF variants

As inferred LoF mutations heavily depend on the gene annotation quality, we used MAFFT v7.467 [47] to align *A. mexicanus* protein sequences (keeping the longest transcript of each gene) with orthologous sequences of 9 other teleost species retrieved using *Broccoli v1.2* : *Danio rerio* (GCF_000002035.6), *Pygocentrus nattereri* (GCF_015220715.1), *Electrophorus electricus* (GCF_013358815.1), *Bagarius yarrelli* (GCA_005784505.1), *Tachysurus fulvidraco* (GCF_003724035.1), *Ameiurus melas* (GCA_012411365.1), *Triplophysa tibetana* (GCA_008369825.1), *Anabarilius grahami* (GCA_003731715.1), *Labeo rohita* (GCA_004120215.1).

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

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

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

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

GO-term enrichment of cave pseudogenes

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

Characterization of large deletions

In a previous study [9], we identified two large deletions in a Pachón cavefish genome assembly, one leading to a truncated *opn8a* gene and the absence of *opn8b* gene, the other leading to a truncated *rgr2* gene. Here, we took advantage of a new Pachón genome assembly (GCA_019721115.1) to verify and further characterize these deletions.

The genes *opn8a* (LOC111191907) and *opn8b* (LOC111197225) are located in tandem in the surface fish genome (in the genomic region NC_035899.1:44673130-44727205) and are flanked by the genes *runx2b* (LOC103028241, NC_035899.1:44525886-44575275) upstream and *mmut* (LOC103027932, NC_035899.1:44737167-44760205) downstream. We used *runx2b* and *mmut* protein sequences as tblastn [54] queries against the Pachón genome assembly to identify the corresponding region (CM033892.1:72415284-72611085). This

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

The gene *rgr2* (LOC103043895) is located on the genomic region NW_019172852.1:462710-472903 in the surface genome assembly. The corresponding region in the Pachón genome assembly (CM033906.1:17152181-17158392) was extracted by using *rgr2* as a query in a tblastn against this genome. The Surface and Pachón regions were also aligned and visualized using dotter. The frequencies of these two large deletions were inferred for Molino, Tinaja, Pachón and Chica cave populations by computing the read alignment coverage in these regions for each surface and cave individual using bamCoverage [56].

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

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

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

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

Phenotype and genotype scoring of Toro cavefish

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

Declaration

Ethics approval and consent to participate

Not applicable

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

Sampling authorization

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

Supplementary Material

Supplementary data are available at BMC Ecology and Evolution

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

Pseudogenes	Gene sets	LoF mutations	Molino (9)	Caves		
				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-rhl.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			

Notes: sample size indicated between parenthesis.

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

Pseudogenes														
	L	n	$\Delta\text{-trgr2}$	$\Delta\text{-opn8}$	$\Psi\text{-rhl.2}$	$\Psi\text{-cryba2a}$	$\Psi\text{-imp81b}$	$\Psi\text{-prph2b}$	$\Psi\text{-gia8a}$	$\Psi\text{-lim2.1}$	$\Psi\text{-pde6b}$	$\Psi\text{-per2-M}$	$\Psi\text{-per2-T}$	$\Psi\text{-tyrp1a}$
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												

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

References

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

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

independently evolved populations. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* 2020, **334**(7-8):423-437.

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

Figure legends

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

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

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

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

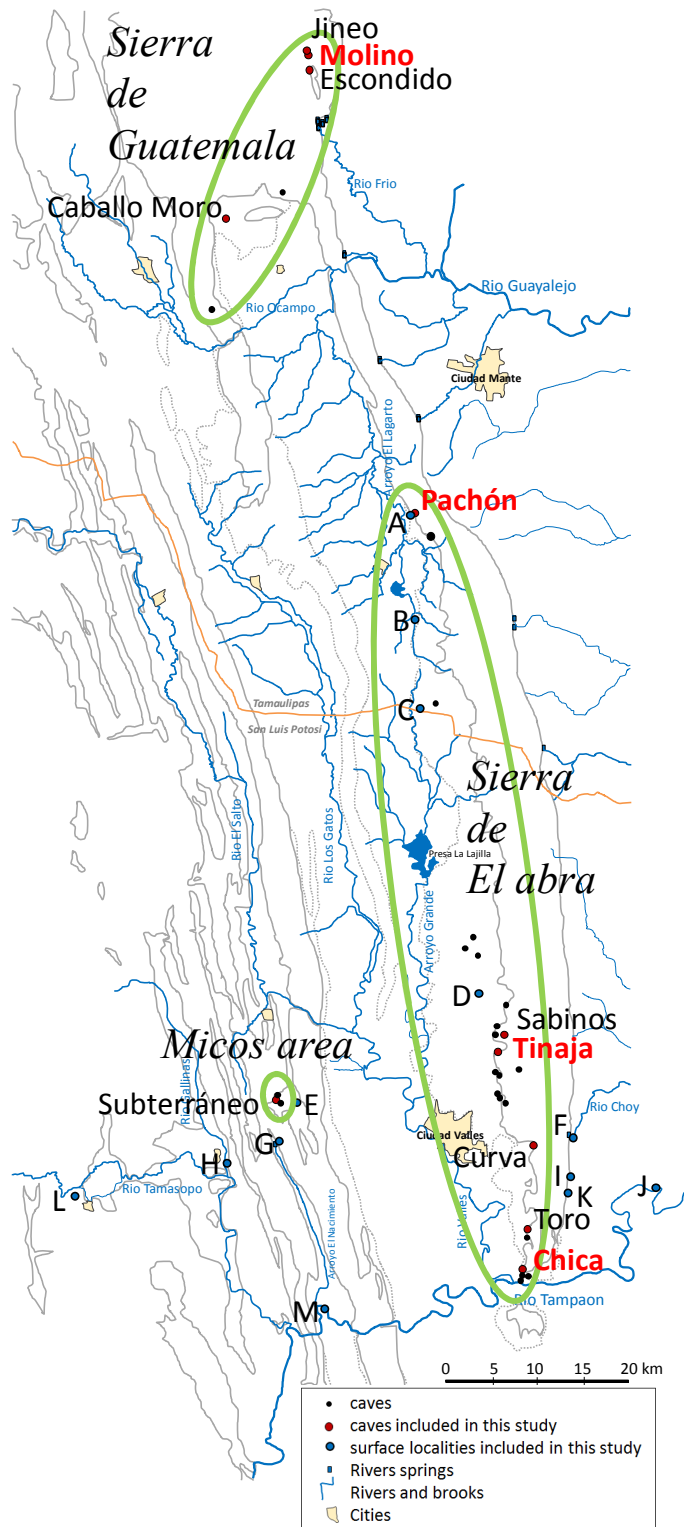
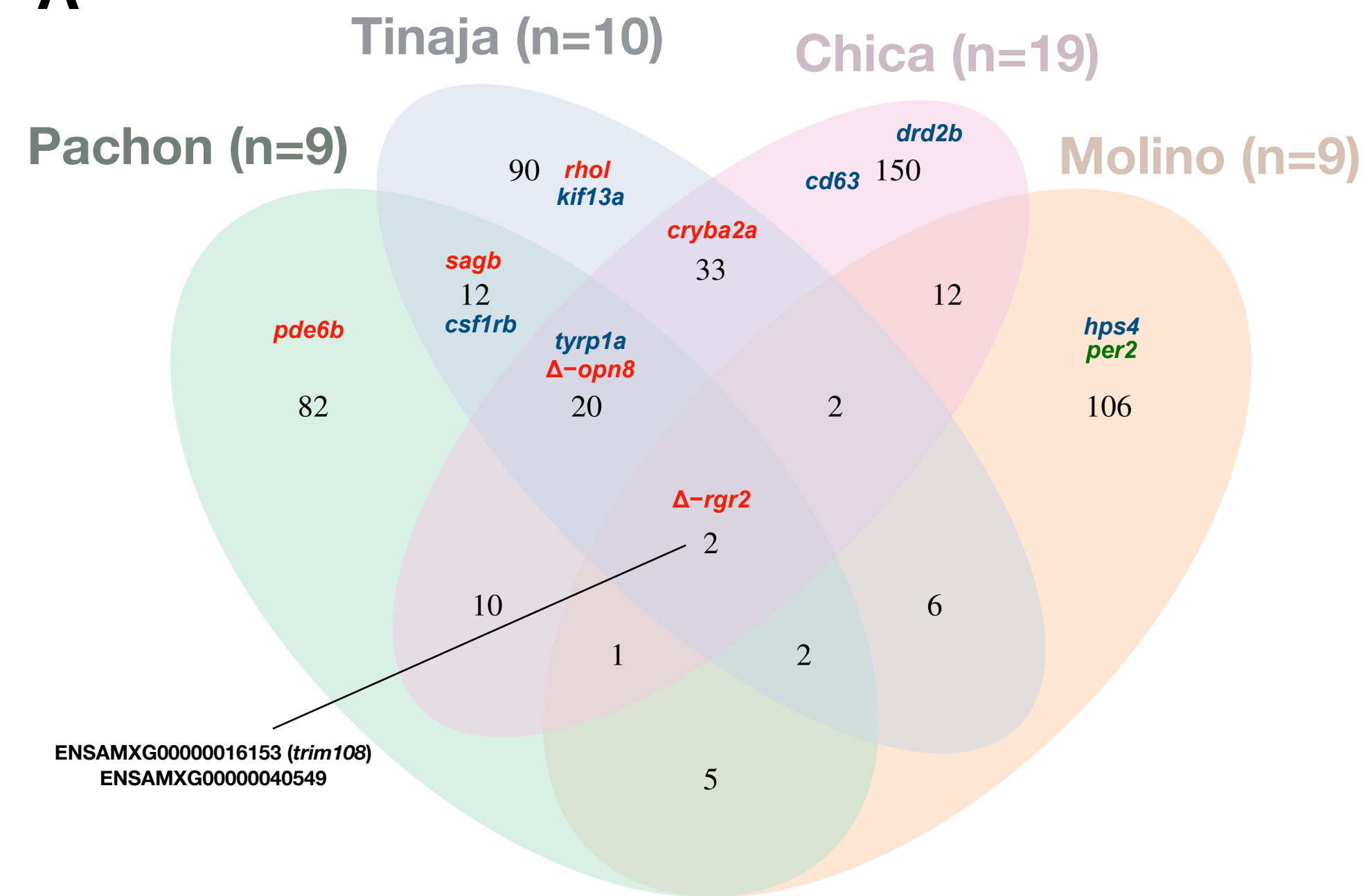
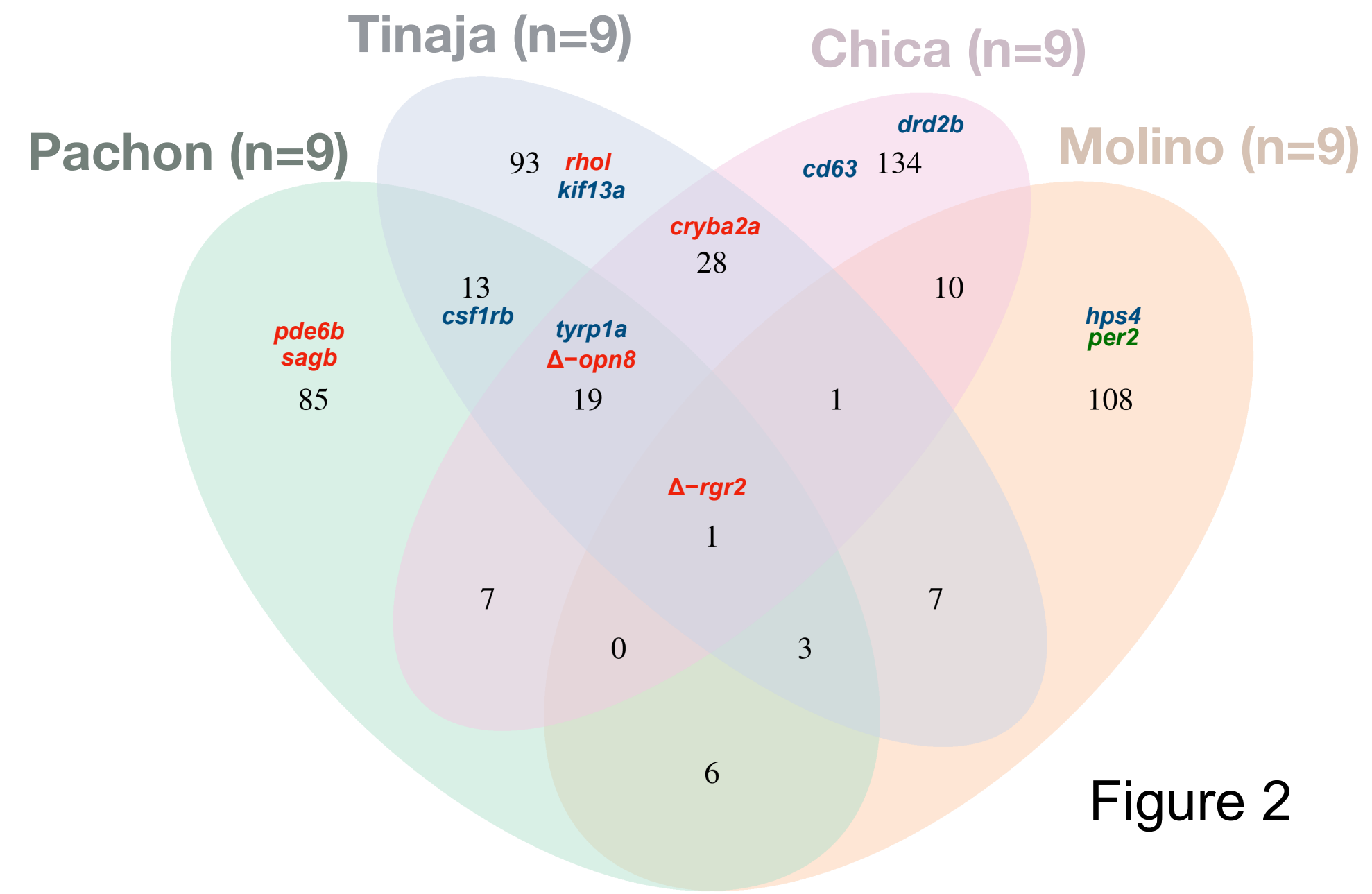


Figure 1

A**B****Figure 2**

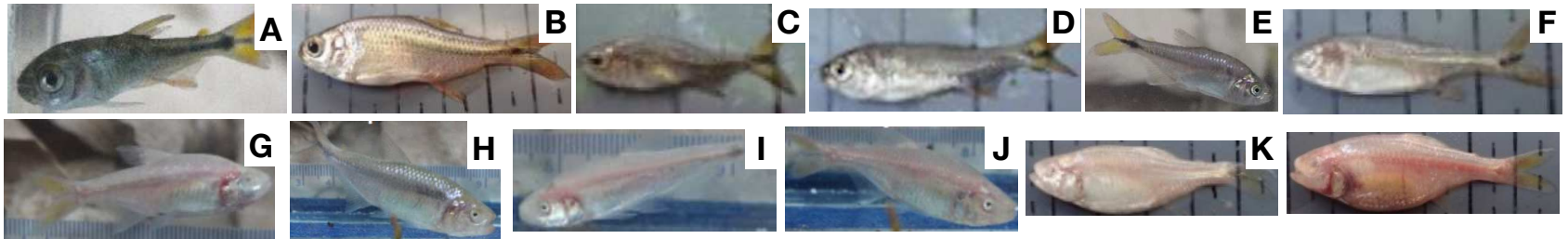
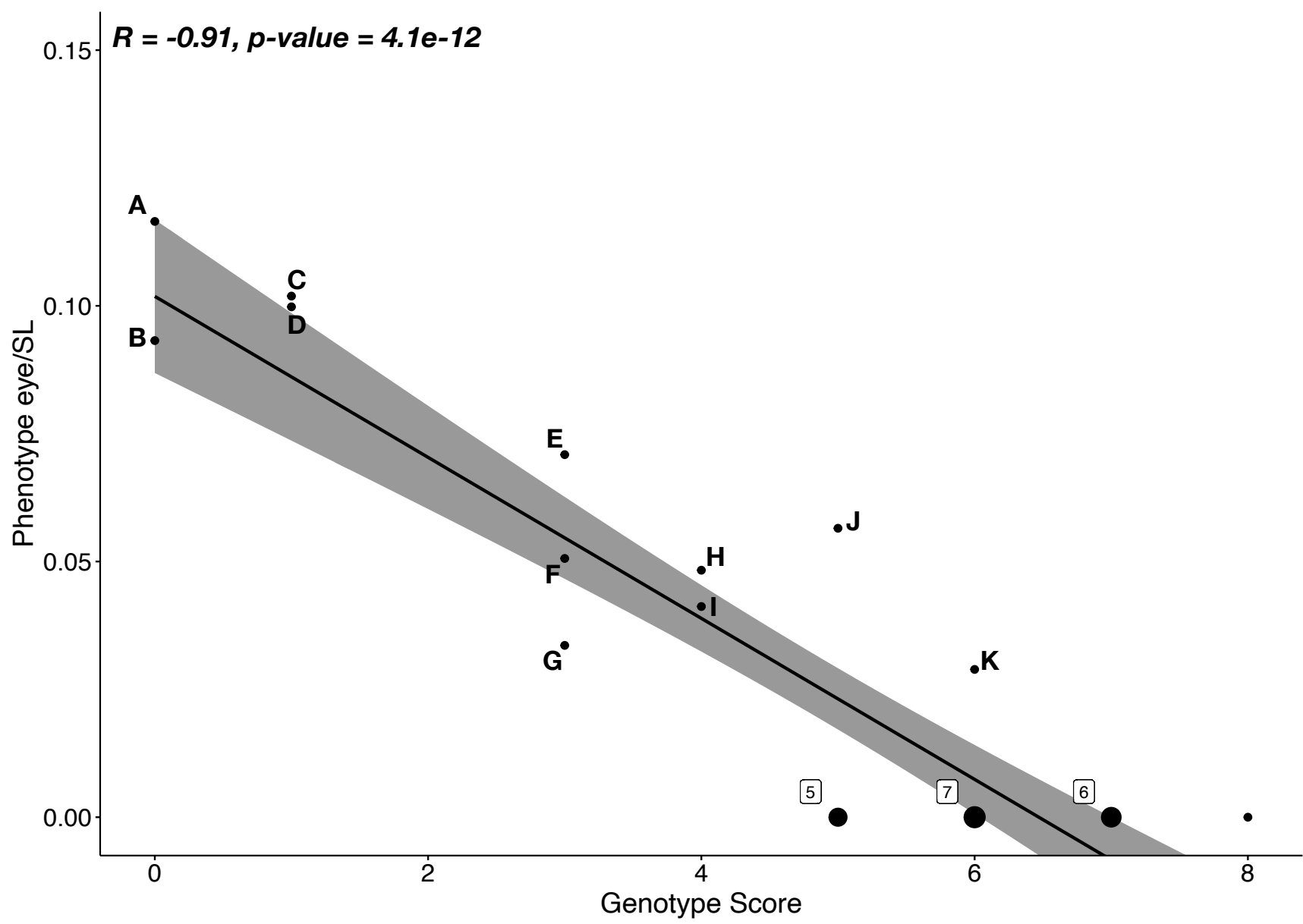


Figure 3

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](#)