Evidence of Late Pleistocene origin of *Astyanax mexicanus* cavefish

Julien Fumey[1], Hélène Hinaux[2], Céline Noirot[3], Sylvie Rétaux[2] and Didier Casane[1,4,*]

[1] Évolution, Génomes, Comportement, Écologie. CNRS, IRD, Univ Paris-Sud. Université Paris-Saclay. F-91198 Gif-sur-Yvette, France.

[2] DECA group, Paris-Saclay Institute of Neuroscience, UMR 9197, CNRS, Gif sur Yvette, France.

[3] Plateforme Bioinformatique Toulouse, Midi-Pyrénées, UBIA, INRA, Auzeville Castanet-Tolosan, France

[4] Université Paris Diderot, Sorbonne Paris Cité, France.

* Corresponding author:

Didier Casane

Laboratoire Évolution, Génomes, Comportement, Écologie, UMR 9191 CNRS, 1 avenue de la Terrasse, 91198 Gif sur Yvette, France.

Tel: +33169823759

Email: Didier.Casane@egce.cnrs-gif.fr

Abstract

Background: Cavefish populations belonging to the Mexican tetra species *Astyanax mexicanus* are outstanding models to study the tempo and mode of adaptation to a radical environmental change. They share similar phenotypic changes such as blindness and depigmentation that are the result of independent and convergent evolution. In particular they allow to examine whether their evolution involved the fixation of standing genetic variation and/or *de novo* mutations. Cavefish populations are currently assigned to two main groups, the so-called “old” and “new” lineages, which would have populated several caves independently and at different times. However, we do not have yet accurate estimations of the time frames of evolution of these populations.

Results: First, we reanalyzed published mitochondrial DNA and microsatellite polymorphism and we found that these data do not unambiguously support an ancient origin of the old lineage. Second, we identified a large number of single-nucleotide polymorphisms (SNPs) in transcript sequences of two pools of embryos (Pool-seq) belonging to the “old” Pachón cave population and a surface population of Texas. Based on the summary statistics that could be computed with these data, we developed a method in order to 1) detect a recently isolated small population and 2) estimate its age. This approach is based on the detection of a transient increase of the neutral substitution rate in such a population. Indeed Pachón cave population showed more neutral substitutions than the surface population, which could be a signature of its recent origin. Third, when we applied this method to estimate the age of the Pachón cave population which is considered one of the oldest and most isolated cavefish populations we found that it has been isolated less than 30,000 years, that is during the Late Pleistocene.

Conclusions: Although it is often assumed that Pachón cavefish population has a very ancient origin, within the range of the late Miocene to the middle Pleistocene, a recent origin of this population is well supported by our analyses of DNA polymorphism as well as by other sources of evidence. It suggests that the many phenotypic changes observed in these cavefish would have mainly involved the fixation of genetic variants present in surface fish populations and within a short period of time.

Keywords: cavefish, adaptation, high-throughput sequencing, SNPs, molecular dating

Background

Two well-differentiated morphotypes, surface fish and cavefish, are found in the species *Astyanax mexicanus*. Twenty-nine cavefish populations have been discovered so far in limestone caves in the El Abra region of northeastern Mexico [1, 2]. Cavefish differ from their surface counterparts in numerous morphological, physiological and behavioral traits, the most striking being that most cavefish lack functional eyes and are depigmented [3]. Most caves inhabited by cavefish share a number of abiotic and biotic characteristics such as constant darkness and absence of predators, and most cavefish show evolution of a number of characters [4], either because they are dispensable - regressive traits - such as loss of eyes and pigmentation [5], or because they are involved in the adaptation - constructive traits - to this environment which is inhospitable for most fishes. For example, cavefish have a lower metabolic rate [6-8], produce larger eggs [9], have more and larger superficial neuromasts involved in vibration attraction behavior [10-12], sleep very little [13, 14], have shifted from fighting to foraging behavior [15], have larger numbers of taste buds [16, 17], have enhanced chemosensory capabilities [18] and have enhanced prey capture skill at both the larval and adult stages [11, 19, 20].

Very significant advances have been made in identifying proximal mechanisms [21], which are the mutations that have changed physiological, developmental, and behavior traits of cavefish and new molecular tools available today will allow us to identify such mutations at an ever increasing pace [22-26]. However it is much more tricky to disentangle distal mechanisms [21], *i.e.* evolutionary mechanisms. Were these mutations already present at low frequency in surface fish standing variation or did they appear after settlement? Are there pleiotropic effects and epistatic interactions? What is the impact of recombination, genetic drift, selection and migration in cavefish evolution? These questions have fueled discussions on the relative importance of these different evolutionary mechanisms [12, 17, 27-31].

In order to analyze several of these issues such as the relative weight of selection, migration and genetic drift, it would be very useful to have accurate estimations of some parameters to describe the dynamic of cavefish evolution. Gene flow from the surface populations has been estimated to be from very low, if any, to very high, depending on the cave population examined. Some studies have also found significant and higher gene flow from cave to surface populations than in the opposite direction [32-37]. Moreover, some caves are very close to each other and fish migrations within some cave clusters are likely.

Among other processes that have to be studied, two are particularly important: 1) when did cave settlements occur and 2) how long did it take for different groups of surface fish to adapt to the cave environment. Currently, no reliable datings are available but *Astyanax mexicanus* cave populations have been assigned to two main groups, the so-called “old” and “new” lineages, which would have populated several caves independently and at different times [37-39], reviewed in [2]. However, and putting aside early estimations of the age of cavefish populations [40] that were not based on reliable data and method, the age of cavefish settlement has been estimated for two populations only, inhabiting the Pachón and Los Sabinos caves, which both belong to the “old” lineage. On the basis of allozyme polymorphism [32] and a population genetic method specifically designed to estimate the time after divergence between incompletely isolated populations of unequal sizes (such as cave and surface populations), these populations were estimated to be 710,000 and 525,000 years old, respectively, suggesting that they could be ancient [41]. However, the small number of loci studied (17 allozyme loci scored), the absence of polymorphism in Pachón and very low polymorphism in Los Sabinos did not allow accurate estimations and the standard error (SE) was very large, 460,000 and 330,000 years, respectively. Taking into account that the

95% confidence interval is ± 1.96 x SE, it implies that these populations could be either very recent or very ancient.

The hypothesis of an ancient origin of the old lineage currently only relies on analyses of mitochondrial DNA (mtDNA) phylogenies of surface fish and cavefish [37, 39, 42]. However, as we will show below, these phylogenies do not necessarily imply an ancient origin of some cavefish populations, an hypothesis which is based on biased lectures of phylogenetic trees and which implies *ad hoc* hypotheses to explain the pattern of population differentiation found at the nuclear level using microsatellite markers. In addition, no dating has ever been performed with these nuclear markers, only estimation of the population differentiation [32-34, 38] and estimation of migration rates among populations [33, 35]. Here, we found that the distribution of the microsatellite polymorphism within and between surface and cave populations could be explained by a recent origin of cave populations. It could also explain the unlikely higher gene flow from several caves to surface populations than from surface to cave populations.

Some comparative analyses of gene sequences also point towards such a hypothesis of a recent origin of the Pachón cavefish population. For example, no obvious loss-of-function mutation, such as frameshifts and stop codons, has been found in eye-specific crystallin genes [26] and opsin genes [43-45], an unexpected result if this population was established at least several hundred of thousand years ago, and very unlikely if it was established more than one million years ago [46]. Indeed, other fish that are confined into caves for millions of years have fixed loss-of-function mutations in several opsins and crystallins genes [47-49].

Using a population genetic approach, we developed a method to estimate the age of a small population recently isolated from a large population, which is based on the detection of a transient higher number of neutral substitutions in the small population than in the large population. The rationale and a detailed description of this method is given in **Additional File**

1. When we analyzed the single-nucleotide polymorphisms (SNPs) in transcript sequences of two pools of embryos (Pool-seq) from the Pachón cave and the Texas surface-dwelling populations, we found that the cavefish population has probably not been isolated millions of years ago but more likely during the last 30,000 years, *i.e*. during the Late Pleistocene or even later. This new time frame together with other evidence indicate that the many phenotypic changes observed in these cavefish may have mainly involved the fixation of genetic variants present in surface fish populations, and within a short period of time.

Results

SNPs and substitution rates in surface and cave populations

We defined eight classes of polymorphic sites according to the presence of an ancestral and/or a derived allele in surface fish (SF) and Pachón cavefish (CF) populations, using the Buenos Aires tetra (*Hyphessobrycon anisitsi*) as an outgroup (**Figure 1**).

Using transcriptome sequence datasets from pooled embryos (**Additional File 2; Figure S1**) we estimated the frequencies of these eight SNP classes at synonymous, non-coding and non-synonymous sites (**Table 1**). The frequencies of SNPs in the eight classes were robust according to the Pool-seq approach [50] and the different parameter thresholds used to include SNPs in the analysis (**Materials and methods**, **Table 1**, **Additional File 2; Figure S2, Table S1a and Table S1b)**. The ratio (SF/CF) of synonymous, non-coding and non-synonymous polymorphism was 3.08, 2.71 and 2.34, respectively, and the ratio (CF/SF) of derived fixed alleles was 2.34, 1.45 and 1.52, respectively. This indicates that the level of polymorphism was higher in the SF population, but the number of fixed derived alleles was higher in the CF population. Using the distances between amino acids as defined by Grantham on the basis of three physical and chemical properties (composition, polarity and molecular volume)[51] and taking the largest distance divided by 2 (215 / 2 = 107.5) as a threshold (**Additional File 2; Figure S3**), non-synonymous mutations were classified as conservative (d < 107.5) or radical (d > 107.5). For conservative and radical mutations, the ratio of SF/CF polymorphism was 2.31 and 2.52, respectively, and the ratio of CF/SF derived fixed alleles was 1.48 and 1.85 **(Table 1)**.

We found also polymorphisms for several derived STOP codons (1 and 2 polymorphic codons in SF and CF populations, respectively), and 4 fixed STOP codons in the cavefish populations (**Table 1**). We did not observe any loss of ancestral STOP codons (**Table 1**).

Estimation of Pro106Leu polymorphism at amino acid position 106 in the MAO protein

We next wished to control and validate our experimental design and dataset using the MAO (monoamine oxidase) case. A mutation of the codon CCG to CTG that replaces a Proline by a Leucine at amino acid position 106 in the MAO protein has been found in cavefish [52]. The C/T polymorphism at this position in three batches of embryos (CF, SF and *H. anisitsi*) and the genotype of five fish (two CF, two SF and one H. *H. anisitsi*) were estimated using the Illumina reads covering this site. On the one hand, we counted 92 C and 217 T with pooled CF embryos, 285 C and 0 T with pooled SF embryos, 149 C and 0 T with pooled *H. anisitsi* embryos, therefore showing that the C to T mutation is not fixed in the Pachón population. On the other hand, we counted 663 C and 0 T in transcript sequences of adult brain and olfactory epithelium tissues of two CF, two SF and one *H. anisitsi* individuals. These five fish were thus obviously homozygous for the C allele and it showed that the rate of sequencing error was very low and it did not generated artefactual polymorphism. The low rate of sequencing error is also suggested by the absence of reads with the T allele in pooled SF embryos and pooled *H. anisitsi* embryos that are expected to be homozygous for the C allele.

Using a PCR approach, among thirty one lab stock Pachón cavefish genotyped, 8 C/C, 16 C/T and 7 T/T were identified and among twenty wild-caught Pachón cavefish genotyped, 1 C/C, 9 C/T and 10 T/T were identified. The frequency of the derived allele, T, in Pachón cavefish population was thus 0.7 when estimated with pooled embryos of the lab stock, 0.48 with a sample of thirty one lab stock fish and 0.73 with a sample of twenty wild-caught fish. The genotype frequencies did not deviate from the Hardy-Weinberg equilibrium in both populations (Chi-square test p-value > 0.05), but allele frequencies were different (Chi-square test p-value < 0.05). Estimations of allele frequencies using pooled embryos or different individuals were also different (Chi-square test p-value < 0.05). Globally, these results showed that: 1) genetic drift occurred in the lab but has been limited, 2) population polymorphism can be identified using pooled embryos and 3) artefactual polymorphism is very low in our dataset.

Estimation of the age of the Pachón cave population

In order to estimate the age of the Pachón cave population, we compared the observed summary statistics of synonymous polymorphism with the summary statistics of neutral polymorphism in simulated populations. We could identify sets of parameters (population sizes, migration rates, generation times, and delay before settlement) that allowed a good fit between the summary statistics of the observed and simulated polymorphism. As an example, we are considering now the simulation that gave the best fit. In this simulation the ancestral population size was set to 10,000 and was at mutation/drift equilibrium; after the separation of the surface and cave populations, the Pachón cave population size was set to 625 and the

Texas surface population size was set to 10,000; there was no delay before cave settlement (t_1 = 0 in **Figure 1**); the probability of migration per year from surface to cave was 0.001 and the number of migrants was 0.1% of the surface population size (*i.e.* 10 fish); the generation time of the cavefish was set to 5 years and the generation time of the surface fish was set to 2 years (**Additional File 2; Figure S4**). Every 100 years (i.e. 50 SF generations, or 20 CF generations), 10 fish were sampled in each population to simulate the sampling process when the lab populations were established. Each lab population was then set with a constant effective population size of 10 over 10 generations. Then we compared the frequency of each SNP class in the simulated lab populations with the observed frequency. In this simulation, the best fit with the data occurred when the age of the cave population was 25,500 years (**Figure 2A**). All SNP class frequencies in the simulated lab populations fit very well (goodness of fit score = 0.15) with the observed frequencies (**Figure 2B**). Then, the older was the divergence of the populations and worse was the fit (see **Additional File 2; Figure S5A** for evolution over one million years).

In this simulation, as well as in all other simulations, the mutation rate per generation (u), that is the probability of appearance of a new allele at a new locus in one haploid genome at a given generation, was set to 2.10^{-2}. The number of new SNPs that appeared per generation in a population of size N was 2Nu, each with a frequency of 1/2N. This means that in the surface population there is $2 \times 10{,}000 \times 2 \times 10^{-2}$ = 400 new SNPs at each generation, and that these 400 new SNPs appear with an initial frequency of $1 / (2 \times 10{,}000) = 5 \times 10^{-5}$. In parallel, 25 new SNPs appear with initial frequency of 8×10^{-4} in the cave population at each generation. All loci were independent. It is noteworthy that the fit of the actual and simulated polymorphism did not depend on the mutation rate because we compared the relative frequencies of SNP classes rather their absolute numbers. Indeed when the mutation rate is higher, the number of SNPs in each class is higher, but the relative frequency of each class remains the same. Thus the score of goodness of fit did not depend on the mutation rate. The mutation rate we used was a trade-off between the accuracy of the SNP class frequency estimations in the simulated populations and the time to run a simulation (the higher the mutation rate, the higher the number of polymorphic sites for which allele frequency evolution was simulated). The estimation of the age of the cave population depends on the generation time in each population.

We searched for other sets of parameters that also fit best the distribution of polymorphism within and between populations. We tested the effect of the Texas population size for which we used two values that correspond respectively to the largest effective population size estimated for a surface population (*i.e.* 5,000) and twice this value (10,000) to take into account the possibility that this effective population size is unexpectedly large [33, 35]. The cavefish population size was set to between 313 and 10,000 individuals. We also took into account migration from the surface to the cave: the probability of migration varied between 0.01 and 0.00001 per year and the percentage of surface fish that migrated into the cave varied between 10 % and 0.01 %. We considered that the migration rate and the number of migrants at each migration from the cave to surface was negligible. The other parameters were the same as in the simulation described above. Additional simulations were run, with no migration but allowing a delay (t_1 = 0 to 80,000 years) before settlement in the cave. For each simulation, we recorded the score of the best fit and the age of the cave population that corresponded to this score (**Additional File 2 Table S2a and table S2b**). We also estimated the age of the cave population with the sets of parameters described above, except that generation time was equal to two years in both the cave and surface fish (**Additional File 2 Table S3a and Table S3b**) or generation time was two years in the cavefish and five years in the surface fish (**Additional File 2 Table S4a and Table S4b**). Yet it is very unlikely that the generation time is smaller in caves, it allowed to simulate a higher mutation rate per year in caves.

We also analyzed the effect of a bottleneck at the time of settlement in the cave, with the other parameters being identical to the first model described above. As expected, setting a long and narrow bottleneck reduced the age of the Pachón cave population a lot, but in some cases the fit to the data was also lost (**Additional File 2 Table S5**).

Globally, even if with most simulations we could not find a good fit between the observed and the simulated frequencies of the seven SNP classes, in a few cases the fit was very good. In all these cases, the cavefish population (t_3) was recent, *i.e.* between 1,500 and 30,000 years old.

No evidence of relaxed selection at the whole exome scale

Using the codon frequencies in coding sequences and the observed ratio of transition/transversion (~3) at synonymous SNPs, we calculated the expected proportion of synonymous (26%) and non-synonymous (74%) sites in SF and Pachón CF populations. In these populations, we observed that 67% of the polymorphic sites are synonymous and 62% of the fixed derived alleles are synonymous, which is very significantly different from the expected 26% (chi-square test, $p < 2.2 \times 10^{-16}$), and in accordance with a much stronger negative selection on non-synonymous mutations than on synonymous mutations.

In order to test the possibility of a relaxed selective pressure on amino acids changes in the Pachón cave population, we split the non-synonymous SNPs into two categories. A mutation was classified as conservative if the Grantham's distance [51] between the ancestral and the derived allele was lower than 107.5, and classified as radical otherwise (**Additional File 2; Figure S3**). We also identified seven mutations responsible for the gain of STOP codons (**Table 1**). Using the codon frequencies in coding sequences and the same ratio of transition/tranversion (~3) at synonymous SNPs, we estimated the expected proportion of conservative (78%) and radical (22%) non-synonymous mutations. Using the total number of non-synonymous mutations that reached fixation in cavefish and surface fish (460 + 302 = 762), we calculated the expected number of conservative (762 x 0.78 = 594.4) and radical substitutions (762 x 0.22 = 167.6) if the selective pressure on conservative and radical mutations were the same. These numbers were compared with the observed numbers of conservative substitutions (399 + 269 = 668) and the observed number of radical substitutions (61 + 33 = 94). The excess of conservative substitutions and deficit of radical substitutions is highly significant (chi-square test, $p = 8.04 \times 10^{-7}$), in accordance with a higher negative selection on radical mutations (**Additional File 2; Figure S6**).

When we compared the relative numbers of conservative and radical substitutions in cave and surface populations (399 and 61 *vs* 269 and 33) a non-significant (chi-square test, p = 0.4) excess of radical substitutions in the cavefish population was found **(Figure 3)**. It suggests that there is non-significant relaxed selection at the whole exome scale in Pachón cavefish population.

Discussion

Estimation of genetic drift in the laboratory stock of Pachón cavefish

A mutation from CCG to CTG that replaces a Proline by a Leucine at amino acid position 106 in MAO protein has been involved in the *Astyanax* cavefish behavioural syndrome [52]. In this previously published study 4 wild and 5 lab-raised Pachón cavefish were found homozygous for the CTG codon, *i.e.* 100% of the fish genotyped. In the present study, looking for this same polymorphism in pooled embryos obtained from our laboratory stock of

Pachón CF we counted 217 reads with the T allele among 309 reads, *i.e*. about 70% of T. Thus we looked for the reason of this discrepancy. Either this allele is not fixed in the Pachón population (but the low number of fish tested previously gave a misleading evidence of fixation), or our estimation using pooled embryos of a lab stock gave a poor estimation of the true frequency because a lot of spurious polymorphism was generated. We thus sequenced a sample of twenty wild-caught Pachón cavefish and thirty one lab-raised cavefish. We found that both frequencies estimated with these independent samples, 73% and 48% respectively, were similar to the frequency found with embryo pooled-seq. We concluded that 1) this allele is actually not fixed in the Pachón cave population, like in the 2 other El Abra populations previously tested [52], 2) the genetic drift in the lab stock is limited and 3) pooled-seq of lab stock embryos allow the identification of polymorphic sites. In addition, the pooled-seq of surface fish embryos and RNA-seq of tissues of fish that are homozygous showed that the level of artefactual polymorphism is very low or zero.

In addition, using simulations of the sampling process we could show that a small number of embryos and a large variance of the number of reads per embryos do not allow an accurate estimation of the allele frequencies in a population but the estimation of the summary statistics are nonetheless very accurate because there are not based on the estimation of allele frequencies but solely on the detection of polymorphism (**Additional File 1; Figure S3**).

A reexamination of previous analyses taken as evidence of an ancient origin of the “old lineage” of *Astyanax* cavefish.

First, we re-examine mitochondrial DNA evidence. The hypothesis that cavefish originated from at least two surface fish stock was first formulated on the basis of a NADH dehydrogenase 2 (ND2) phylogeny of cave and surface fish [39]. On the one hand all surface fish from the Sierra de El Abra belonged to a haplogroup named “lineage A”, as well as two surface fish from Texas and a surface fish from the Coahuila state, in northeastern México. Pachón and Chica cavefish also belonged to this haplogroup A. On the other hand Curva, Tinaja and Sabinos cavefish, found in caves geographically close, belonged to another and well differentiated haplogroup named “lineage B”. The authors concluded that Pachón cavefish could nevertheless have the same origin as the haplogroup B cavefish , *i.e.* an old stock of haplogroup B surface fish, but now extinct and replaced by surface fish with haplotypes belonging to haplogroup A. It implies that the mtDNA haplotype A1 found in Pachón cavefish would be the result of a mtDNA introgression involving at least one migration into the cave of a surface female of haplotype A1 and the fixation of this haplotype in the whole cavefish population. It is worth noting that the authors also proposed another and simpler explanation: Pachón cavefish have evolved independently, more recently than haplogroup B cavefish, and they are undergoing troglomorphic evolution more rapidly than other cavefish populations.

This mtDNA phylogeography was confirmed with a partial sequence of the cytochrome b gene [36]. In this study a third haplogroup was identified in Yucatan. Using a more comprehensive sample and the same mtDNA marker [37], up to seven divergent haplogroups were found in Mexico (A to G, the haplogroup G for cytb corresponding to haplogroup B with ND2) with allopatric distribution reflecting a past fragmentation and/or a strong isolation by distance of the species distribution. In this study, haplogroup G was still cave specific and haplogroup A Northern Gulf coast and cave specific. However a more recent analysis [42], expanding further the sampled populations, allowed the identification of surface fish belonging to the haplogroup G (named Clade II lineage Ie) and haplogroup A (named Clade I Ia) in sympatry in the same water bodies, *i.e.* Mezquital and Aganaval, in Northwestern Mexico. This finding invalidates the hypothesis that haplogroup G evolved in El Abra region a long time ago and was replaced by haplogroup A. Indeed haplotypes belonging to haplogroup G are still found in extant surface fish in Northwestern Mexico. Nevertheless, the haplogroups A and G are highly divergent, supporting a model in which they accumulated mutations in different populations isolated during a long period of time and mixed recently, at the time of a secondary contact. Taking into account the current distribution of the main mitochondrial haplogroups, haplogroup G could have evolved in the northwestern region of Mexico and the haplogroup A in the northeastern region of Mexico, where they could have been isolated during a long period of time. During the last glaciation, these populations in north Mexico might have moved south and mixed there. After this glaciation they might have moved north again, now sharing haplotypes belonging to haplogroup A and G (this haplotype mixture is actually observed in the northwestern region, *i.e.* Mezquital and Aganaval water bodies). In the northeastern region, haplotypes belonging to the haplogroup G have up to now been found only in several caves in a restricted geographic area suggesting that these haplotypes were at low frequency and finally disappeared everywhere excepted in several caves where they could reach fixation and they were conserved thanks to cave isolation. Such recent secondary contact of divergent haplogroups were observed at several places in south Mexico [34, 42] suggesting that several populations of *Astyanax mexicanus* were isolated for a long time in different regions in Mexico and Central America and they have recently been in secondary contact.

Second, we re-examine nuclear DNA evidence. As mentioned above, despite mtDNA evidence it has early been proposed that Pachón and Yerbaniz cavefish share a common ancestry with Sabinos, Tinaja, Piedras and Curva cavefish, *i.e.* their ancestors would be a population of surface fish that has been replaced by another population of surface fish after cavefish settlement. This hypothesis is supported by several analyses of RAPD and microsatellite polymorphism. A parsimony analysis of RAPD data gave an unresolved phylogeny with a low support for a unique origin of the cave populations [53]. In a neighbor-joining tree, based on Nei's DA distances estimated using six microsatellite loci, Pachón, Sabinos and Tinaja cavefish were more closely related with each other than with Chica cavefish and surface fish populations [36]. In particular, at three loci Pachón, Sabinos and Tinaja cavefish showed low polymorphism and the same highly frequent allele despite the large number of alleles identified in surface populations. It was also suggested that these cavefish would have been isolated of the surface populations. This result was confirmed using another approach [34, 38] and 26 microsatellite loci [33].

In these studies, Yerbaniz cavefish appeared related to the "old stock" cavefish despite the mtDNA evidence and it would imply, as for Pachón cavefish, the replacement of the original mtDNA haplotype belonging to surface haplogroup G by a mtDNA haplotype belonging to surface haplogroup A, but without detectable introgression at the nuclear level.

We looked at allele frequency distributions found in this study and we came to the conclusion that they do not support unambiguously an ancient origin of cave populations. These distributions are shown in **Additional file 3; Figure S1 to S26**. Before we examine these distributions, we have to describe the expected distributions under the current evolutionary hypothesis about the "old" cavefish populations studied (Pachón, Yerbaniz, Sabinos and Tinaja). If they are actually several hundreds of thousand years old (it is often claimed several million years old) and if the gene flow is low between Pachón and the other caves and with surface fish as several studies suggested, we expect that the distribution of the allele frequencies would be very different between the most isolated caves and between caves and surface populations. Under the stepwise mutation model (mutation by addition or subtraction of one repeat unit) which is the most conservative model in that way that the ancestral distribution diverges slowly in isolated populations, the expected allele frequency distribution in each population is centered on one high frequency allele flanked by alleles with lower numbers and higher numbers of repeats present at low frequencies. In a large population the distribution would be wide but in a small population the distribution would be narrow [54]. This is what is observed in large surface populations and in small cave populations (**Additional file 3; Figure S1 to S26)**. Moreover it is expected that these allele frequency distributions are wandering [55], that is the size of the most frequent allele changes through time and the difference between the mean repeat numbers at a locus in two populations increases with the time of divergence. More precisely $E(m_x - m_y)^2 = 2\mu t$ (where m_x and m_y are the mean repeat numbers at a locus in populations x and y, μ is the mutation rate and t the number of generation since the two populations are separated). For example, if the mutation rate is 5×10^{-4} (mutation rate used in previous *A. mexicanus* population genetic analyses), two populations separated for several tens of thousands years should have allele frequency distributions in which the most frequent allele would be of different size and the mean repeat numbers would be also different. This is not observed for most cavefish population that show at most loci very similar distribution with the same most frequent allele. In addition for several loci two alleles are present with a high frequency, a situation that should be transitory (only one allele should have a high frequency most of the time) and thus this state could not have maintained for hundreds of thousands years independently in several caves. On the contrary when a different "most frequent allele" is found in different caves it is not a signature of an ancient divergence. Indeed, for most loci, the distribution of the allele frequencies is wide and flat in surface populations in accordance with their large population sizes and high mutation rate at the loci analyzed (loci used in population genetic studies are selected for their polymorphism, thus they have a high mutation rate). In such case, it is expected that random changes in allele frequencies led to the fixation of different alleles in independent cave populations. Nevertheless when only one allele (or one high frequency allele and a couple of very low frequency alleles) is shared between caves, this could be the result of a low mutation rate at that locus. However most often many different alleles are found in the surface populations suggesting that a very low mutation rate is not the most likely explanation of such very similar allele frequency distributions in caves (**Additional file 3; Figure S1 to S26)**. It is worth noting that there is no private allele in cavefish populations when there are compared with surface fish populations in the same area, in contrast with several private alleles found in two surface populations from Yucatan [36].

We estimated the distances $(m_x – m_y)^2$ also known as $(\delta\mu)^2$ between populations [56](**Additional file 3; Table S1)**. The distances (excepted for two northern cavefish populations, Molino and Caballo Moro, that are highly divergent) are of the order of magnitude of the distances between African and non-African human populations (6.47) which correspond to about 6,000 generations [56, 57]. Interestingly, the distance between Pachón cavefish (O1) and the closest surface fish population (S3) is 6.7. Assuming that the mutation rate per generation in human and fish are similar [58], that is about 5 x 10^{-4} and taking into account that the generation time is two years for surface fish and five years for cavefish, the age of the cavefish population $t = (\delta\mu)^2 / \mu \text{ x } [(g_{CF} \text{ x } g_{SF})/(g_{CF} + g_{SF})]$, where g_{CF} and g_{SF} are the generation time of cavefish and surface fish respectively. Replacing the parameters by their estimations, we obtained t = 19,142 years. Interestingly this estimation is close to the estimation we obtained with a very different approach (see results).

A recent origin of cavefish populations could also explain several odd results about the migration rates between cave and surface populations: several cases of a higher migration rate from cave to surface than from surface to cave that could appear biologically unrealistic [33, 35]. Indeed if the shared polymorphism observed between cave and surface fish is due to the recent origin of cave populations and it is not an equilibrium between mutation, drift and migration, it is expected that using a software such as MIGRATE [59], an artefactual high gene flow would be found (http://popgen.sc.fsu.edu/Migrate/Blog/Entries/2010/8/15_Violation_of_assumptions%2C_or_are_your_migration_estimates_wrong_when_the_populations_split_in_the_recent_past.html). In addition, as the alleles present in caves are very often a subset of the alleles present in surface populations, it is expected that the artefactual gene flow inferred is from cave to surface.

In summary we came to the conclusion that previous analyses of mitochondrial and microsatellite polymorphism did not unambiguously demonstrate that the old cave populations are actually that old. We thus looked for other evidence that could support a recent origin of Pachón cavefish population.

Dynamic of substitution rates in two recently and incompletely isolated populations of unequal size

When a population splits into two populations, genetic variation continues to be shared by the daughter populations for a period of time thereafter, even in the absence of gene exchange. As divergence proceeds, loci that were polymorphic in the ancestral population experience fixation of alleles in the descendant populations, and this sorting of alleles is part of the way the populations become different. It is thus challenging to estimate if shared polymorphism is due to a recent split, high gene flow or both [60].

We propose a method for dating a recently isolated small population which is based on a transient acceleration of the neutral substitution pace in such a population that would not be observed in an ancient population. Indeed, we found that a higher number of derived alleles (either synonymous, non-synonymous and non-coding mutations) reached fixation in cavefish than in surface fish (**Table 1**) and we seek for an explanation for these observations that were unexpected, in particular for synonymous mutations that are for most of them neutral or nearly neutral mutation in metazoans [61, 62]. Indeed and in such case the substitution rate should be independent of the population size [63]. A simple explanation relies on the fact that when an ancestral population is divided into a large (surface) and a small (cave) population, the probability of fixation of a neutral allele is the same in both populations if its frequency is the same in both populations. However if this neutral allele reaches fixation, the process is faster in the small than in the large population. The consequence is a transient acceleration of the substitution pace in the small population that is not anymore observed, as expected [63], after a long period of time (**Additional file 1; Figure 1C)**. We thought that this information, together with information about the distribution of polymorphism within and between populations, could be used for divergence dating. We thus aimed to find a method based on the simultaneous analysis of polymorphism and divergence at unlinked loci such as SNPs scattered along the genome. First, we define summary statistics describing the polymorphism and the divergence of two populations (**Figure 1)** that could be accurately estimated using pooled RNA-seq [50](**Additional file 1; Figure S3)**. Then we ran simulations of the divergence of two populations according to different sets of demographic and evolutionary parameters (*i.e*. population sizes, migration rates and divergence time) and we looked for simulated populations showing similar summary statistics to those found with the true populations in order to get estimations of the divergence time compatible with the summary statistics.

Evidence for a recent origin of an “old” population

We identified 3.08 times more synonymous polymorphisms in surface fish than in cavefish. If the populations are at mutation and genetic drift equilibrium, this results suggests that the effective population size of surface fish is about three times larger than that of the cavefish. This is in accordance with previous estimations suggesting that *Astyanax* cavefish effective population sizes are often several times smaller than surface fish population sizes [33, 35]. We also observed 2.34 times more derived allele substitutions at synonymous polymorphic sites in the cavefish population than in the surface fish population. This result was unexpected because synonymous mutations are essentially neutral, and we would expect that new neutral alleles would accumulate at the same rate in both populations, *i.e.* independently of the population size [63]. Nevertheless, such a ratio may be observed if most of the derived alleles that are fixed in both populations were already present in the ancestral population as standing variation. In this case, the time for an allele to reach fixation depends on its initial frequency and the population size [64]. We would thus expect that during a transitory period more derived alleles would reach fixation in the smallest population (**Additional file 1; Figure S1E)**. Noteworthy even if the mutation rate is lower in the small population, as it could be expected in a cavefish population that probably experienced an extended generation time, this signal is not erased because most mutations that reached fixation occurred in the common ancestral population. In our simulations of polymorphism evolution in populations, we set the generation time to two and five years for the surface and cave populations, respectively. This surface fish generation time is twice the estimations obtained for other *Astyanax* species [65] and the cavefish generation time is the value estimated by P. Sadoglu, unpublished but reported as a personal communication [41]. This estimation is based on the hypothesis that cavefish may live and remain fertile for a long time, about 15 years. It is unlikely that these generation times are underestimates and they could actually be overestimates of true generation times. As the estimation of the age of the Pachón cavefish population directly depends on these generation times, the ages we discuss below are more likely overestimates than underestimates. We ran the simulations with two other sets of generation times (*i.e.* 2

years for both populations and a more unlikely scenario: 5 years for the SF and 2 years for the CF that implies that the mutation rate per time unit is higher in CF than in SF). The estimations of the age of the Pachón cave population were similar. We also took into account: 1) migrations between the populations, 2) a delay which is the time between the divergence of two surface populations and the settlement of fish belonging to one population into a cave, 3) a bottleneck at the time of settlement, and 4) genetic drift in the lab populations (**Figure1** and **Additional file 2; Figure S4**). In order to search for the upper limit of the age of the cave population, we first ran the simulation without a population bottleneck at settlement and no delay before settlement in the cave (i.e. t_1 and $t_2 = 0$; **Figure1** and **Additional file 2; Figure S4**). The latter hypothesis is actually supported by the fact that most surface populations in El Abra region show almost no differentiation and can be considered as a single large panmictic population [33] that may include the Texas population we studied. Without migration, shared polymorphisms were quickly lost and the best fit of the model to the data was obtained when the cavefish population size (625) was smaller than the surface fish population (5,000) and the age of the cavefish population was 9,490 years (**Additional file 2; Table S2b**). When migration was included, good fit with the data also implied large differences in population sizes, a low migration rate and low numbers of migrants. The very best fit was observed for a SF population size of 10,000, a CF population size of 625, and a CF population age of 25,500 years (**Figure 2** and **Additional file 2; Table S2a**). With the same population sizes and without migration the goodness of fit was not that good. If the cave population is old, *i.e.* more than 100,000 years old, the goodness of fit with observed data was very poor in both cases (**Additional file 2; Figure S5a and S5b**). If we consider that the SF population size may actually be smaller (5,000), the origin of the CF population may be even more recent (~10,000 years) (**Additional file 2; Table S2a**).

There are several reasons to think that the surface fish effective population size is indeed not very large and in the order of magnitude of 10^4. First, previous estimations were all inferior to 10^4 [33, 35]. Second for a fish species such as *A. mexicanus* in which a female can lay thousands of eggs, the variance of the numbers of descendants can be large and thus the effective population size several order of magnitude smaller than the census population size [66, 67].

Third, if the surface fish effective population size is actually much larger, let say 10^6 or more, the cavefish effective population size, which has never been estimated much smaller than the surface fish effective population size, would be about 10^5 or more, which if very unlikely. Good fit between the simulations and observed polymorphisms was also observed with a low migration rate and a large number of migrants at each migration event. In these cases, the system was cyclic, *i.e*. a good fit was observed repeatedly a few thousand years after each massive migration shifting the system far from a mutation/drift/migration equilibrium. As the number of migrants was sometimes larger than the number of cavefish it re-homogenized the two populations. In these simulations we could find several ages for the cavefish population for which the score of goodness of fit was good (**Additional file 2; Figure S7**). Noteworthy, we did not find a stable mutation/drift/migration equilibrium that fitted well with the data and this would imply that the Pachón population could actually be ancient. Other parameters, such as a bottleneck for several generations (t_2 in **Figure 1**) at the time of cave population settlement and the period of time after the separation of two surface populations and before settlement in the cave (t_1 in **Figure 1**), were set to zero in the simulations discussed above. If these parameters were not set to zero, the age of the cavefish population was further reduced. During time t_1 and t_2, differentiation of the populations was already taking place and the observed differentiation could thus be reached within a shorter time (t_3 **in Figure 1**). We examined the consequences of a period of time (t_1) at the surface before settlement in the cave. In this case the age of the cavefish population was also reduced (**Additional file 2; Table S2b, S3b and S4b**). If a population bottleneck during t_2 years is taken into account, the age of the cavefish population was also reduced (**Additional file 2; Table S5**). In conclusion, there is no good fit between the data and a simulation for the Pachón cavefish population being older than 30,000 years. In any case, it may be even more recent.

Other evidence for a recent origin of the Pachón cavefish

First, we found very low mtDNA divergence between the Pachón cavefish and Texas surface populations. In the 602 bp long cytb gene fragment previously used in population genetic studies [37], we found only two substitutions between surface fish and cavefish in our dataset of pooled embryos. To check this result, we sequenced the mtDNA of two fish from both populations and we found these and only these two substitutions. The phylogenetic distance is 2 / 602 = 0.003, which suggests a coalescence time of 200,000 years if we use a substitution rate of 1.5% / million year, as in most phylogenetic studies [37]. However the standard error (SE) on the estimated divergence time is very large (0.002). Taking into account that the 95% confidence interval is ± 1.96 x SE, it implies that the divergence of the mtDNA could be very recent. Indeed, among extant fish in sympatry in a given surface population, mtDNA sequences with this level of divergence have been found [37]. Such a low divergence of mtDNA is thus compatible with a very recent origin of the Pachón cavefish (**Additional file 2; Figure S8**).

Second, in a recent analysis of the expression of 14 crystallin genes in the Pachón cavefish, 4 genes are not expressed or at a very low level, but no stop codon or frameshift could be identified [26]. This result is in accordance with a recent origin of this population, as several loss-of-function mutations should have reached fixation after several hundred thousand years of evolution of genes that would no longer be under selection, as they are not necessary in the dark [46]. Indeed, other fish species that are likely confined into caves for millions of years have fixed loss-of-function mutations in several opsins and crystallins genes [47-49].

Third, a recent study has shown that the heat shock protein 90 (HSP90) phenotypically masks standing eye-size variation in surface populations [68]. This variation is exposed by HSP90 inhibition and can be selected for, ultimately yielding a reduced-eye phenotype even in the presence of full HSP90 activity. This result suggests that standing variation in extant surface populations could have played a role in the evolution of eye loss in cavefish. This is also compatible with a recent origin of the cave population.

Non-equilibrium model of Pachón cave population genetics

The recent origin of the so-called “old” Pachón population can solve two conundrums put forward by previous and the present analyses. First, at the SNP and microsatellite level, the diversity is not that low in Pachón cave when compared with surface populations, *i.e.* about one third. If the populations are at migration/drift equilibrium, it means that the effective population size of Pachón cavefish is about one third of the surface populations, and this is at odds with the huge difference in census population sizes [1, 33]. Of course, we can propose *ad hoc* hypotheses to explain this discrepancy. Cavefish may have a much lower reproductive success variance than surface fish, or surface fish could have larger population size fluctuations through time than cavefish. In such cases, the effective population sizes could be much closer to one another than to the census population size because it is well established that large variance in reproductive success and large population size fluctuations hugely reduce the effective population size [69]. An alternative explanation is that the genetic diversity in the Pachón cave is actually higher than expected at mutation/drift/migration equilibrium. Our results suggest that the effective population size of the surface fish is at least one order of magnitude larger than the effective population size of cavefish, a ratio that is more in accordance with the unknown but certainly very different long term census population sizes.

The new time frame we propose for the evolution of the Pachón cave population would not allow enough time for the fixation of many *de novo* mutations and most would be derived alleles that were already present in the ancestral population (**Figure 2D**). This may imply that the cave phenotype evolved mainly by changes in the frequencies of alleles that were rare in the ancestral surface population. In particular, some of these alleles would have been loss-of-function or deleterious mutations that cannot reach high frequency in surface populations but they could reached high frequency or fixation quickly in a small cave population where they are neutral or even advantageous. It is likely that all *Astyanax* cave populations are recent and evolved in this way, and it could explain the parallel fixation of identical alleles in isolated caves [52, 70-73]. In addition, some alleles could have spread in several caves if they were connected to each other [74].

It is noteworthy that different ancestral loss-of-function or deleterious alleles would get fixed in different cave populations [70, 75] without the need of *de novo* mutations. Very often many deleterious mutations in the same gene coexist in a large population, each at very low frequency [76]. Thus, the finding of different mutations in different caves is not a definitive evidence that they are *de novo* mutations.

We do not exclude that cavefish populations of the *Astyanax mexicanus* species have existed for a very long time. But these cave populations may have experienced such a high extinction rate that very old populations cannot not be found. The application of our population genomic approach to other cave populations could help shed some light on this issue. The evolution of similar phenotypes in independent populations adapting to a new environment in a short period of time, that is in about ten thousand years, is actually not that unexpected and has already been observed in other fish species such as the stickleback [77], dwarf whitefishes [78] and African cichlids [79, 80]. Cavefish could thus be a new and striking illustration that several large phenotypic changes can accumulate in parallel and in a short period of time thanks to standing genetic variation [81]. The relative roles of selection and drift in allelic frequency changes is not yet understood, but if the recent origin of this cavefish population is confirmed, it would be a good model to analyze this issue using population genomics tools such as the quantification of selective sweep around candidate loci most likely involved in the adaptation to a cave environment.

Increased rate of fixation of deleterious mutations in the cavefish population

A higher number of polymorphic sites in SF compared with CF was observed at synonymous sites (ratio = 3.08), and to a lesser extent, at non-coding (ratio = 2.71) and non-synonymous (ratio = 2.34) sites (**Table 1**). These lower ratios may be the result of stronger selection against deleterious mutations at some non-coding and non-synonymous positions in the surface population than in the cave population because the surface population size is much larger than the cave population size, resulting in higher selection intensity. Indeed, the efficacy of selection, referred to as “selection intensity” depends on the product of selection coefficient (s) and effective population size (N_e). The evolutionary dynamics of weakly selected mutations (when s is very small) are thus highly sensitive to population size because such mutations can behave as neutral mutations in small populations but as selected mutations in large populations [82]. Most non-synonymous mutations are likely neutral or slightly deleterious. Whereas the distribution of selection coefficients of new mutations is not well established and is thought to vary among species, it is assumed that a large fraction of mutations are only slightly deleterious [82-87]. In humans for example, an excess proportion of segregating damaging alleles has been found in Europeans relative to Africans, most probably the consequence of the bottleneck that Europeans experienced at about the time of the migration out of Africa [88] but not necessarily because natural selection has been less effective [89]. We observed an excess of radical amino acid substitutions in cavefish, but it is not significant (chi-square test, p = 0.4) (**Figure 3**). The same trend was observed for STOP codons, but the numbers of polymorphisms and fixed STOP codons in both populations were so low that we could not evaluate the significance of the differences observed. This slight bias toward an accumulation of more deleterious mutations in cavefish genome may be the consequence of a small population size and high isolation, but not for a period of time long enough to have a clear and significant effect on the evolution of coding genes.

Darwin wrote: “I am only surprised that more wrecks of ancient life have not been preserved, owing to the less severe competition to which the inhabitants of these dark abodes will probably have been exposed.” [90]. Indeed, cave animals are often portrayed as degenerate organisms that have survived in low selection refuges. We think that it is more likely that these fish population experienced strong selection in order to adapt to a new and very challenging environment in a very short period of time and this is why it has not occurred very often. To test this hypothesis, the next step would be to search for evidence of selection at loci that could have been involved in this process.

Materials and Methods

Sampled populations

For fifteen years we have maintained laboratory stocks of *Astyanax mexicanus* cavefish and surface fish, founded with fish collected respectively in the Pachón cave (Sierra de El Abra, Mexico) and at the San Solomon Spring (Texas, USA), and obtained from W. R Jeffery in 2004. In 2012, we purchased thirty *Hyphessobrycon anisitsi* fish.

RNA samples and RNA-seq

In order to identify polymorphisms at the population level based on a Pool-seq approach [50], for each population, 50 to 200 embryos/larvae from several independent spawning events and at different developmental stages (6 hours post-fertilization to two weeks post-fertilization) were pooled and total RNA isolated. Total RNA was also isolated from the brain and the olfactory epithelium of two adult fish from each population. Five RNA samples were thus obtained for each population of *Astyanax mexicanus* (cavefish and surface fish) and five RNA samples for the other species, *Hyphessobrycon anisitsi* (**Additional file 2; Figure S1**).
Each RNA sample was sequenced on an Illumina HiSeq 2000 platform (2 x 100 bp paired-end). The pooled embryo samples had been previously sequenced using the Sanger and 454 methods [91] (**Additional file 2; Figure S1**).

Transcriptome assembly and annotation

The *Astyanax mexicanus* transcriptome was assembled with Newbler ver. 2.8 (Roche 454) sequence analysis software using 454 sequences (2.10^6 reads) of both the Pachón cave and surface fish pooled embryos (**Additional file 2; Figure S1**). We obtained 33,400 contigs (mean contig length = 824 bp). We also tried to generate a transcriptome assembly using the Illumina sequences, but whereas this resulted in more contigs (49,728) than the 454 sequences, many of them were concatenations of different transcripts and in some cases the same transcript was found in more than one contig. We therefore mapped the Illumina sequences onto the 454 contigs to identify and annotate SNPs. Putative coding sequences in each contig were identified using the zebrafish (Zv9) proteome available at EnsEMBL 73 as a reference [92]. A contig was considered protein coding if the e-value for the best hit was $< 10^{-5}$. We found 13,240 protein coding contigs (contig mean length = 530 bp). We identified contigs containing domains that matched different zebrafish proteins and which were most likely chimeric contigs. These contigs were removed (369, *i.e.* 3% of the protein coding contigs). In total, we analyzed 12,871 putative protein coding contigs.

SNP identification and annotation

Illumina sequences were aligned to contigs with BWA [93] using the default parameters for paired-end reads. *Hyphessobrycon anisitsi* sequences were aligned to *Astyanax* contigs using a lower maximum edit distance (n = 0.001).

SNPs calling was performed using GATK UnifiedGenotyper v2.4.9 [94]. Because we filtered SNPs after detection using different parameter thresholds described below, we used the allowPotentiallyMisencodedQuals and –rf BadCigar options. We detected 299,101 SNPs including 141,490 SNPs in annotated contigs.

When a complete coding sequence was identified, *i.e.* from the start codon to the stop codon and corresponding to a complete zebrafish protein, we could identify the non-coding flanking sequences (containing 18,743 SNPs), otherwise only the sequence matching the coding sequence of the zebrafish was annotated as coding and the flanking sequences were not annotated. The 55,950 SNPs in the coding sequences were annotated as synonymous or non-synonymous, according to which amino acid was coded for by the alternative codons resulting from the SNP. The ancestral allele and the derived allele were inferred according to the allele found in the outgroup *Hyphessobrycon anisitsi* (**Figure 1**). SNPs for which the ancestral allele and derived allele could not be identified, either because in *Hyphessobrycon anisitsi* no sequence could be identified or there was another allele present or the allele was polymorphic, were discarded.

SNP classification

The SNPs identified in *Astyanax mexicanus* SF and CF were classified into eight classes (**Figure 1**). The number of SNPs in the different classes depended on the thresholds used to consider a SNP as reliable and polymorphic in each population. The rationale for the set of thresholds selected is given below.

The populations being closely related (they belong to the same species) and the mutation rate for a SNP origin being very low (~10^{-8}), we would expect that the eighth class (divergent polymorphism) of SNPs would be a very rare outcome because it is the result of two independent mutations at the same site, either in the ancestral population or in the CF and SF populations. We found only one SNPs in this class **(Table 1)**. It suggests that Illumina sequencing did not generate a number of sequencing errors that would significantly inflate the number of SNPs identified.

Parameter thresholds for SNP selection

We examined the effect of the thresholds applied to parameters used to discard SNPs before their classification and population genomics analyses.

First we looked at the effect of sequencing depth. Whereas the mean sequencing depth was 820, the standard deviation was very large (9,730). When the minimal number of reads per population at a SNP site was set to 100 or higher, the relative frequencies of the eight SNP classes were very stable, indicating that 100 was a good compromise between the stability of the distribution of the SNPs into different classes and the number of SNPs discarded (**Additional file 2; Figure S2**).

We then considered the effect of the e-value of the blast between the *Astyanax* contig and the zebrafish sequence used for annotation, in order to discard poorly conserved sequences that were misidentified as protein coding. It appeared that the SNP classification was stable whichever the threshold was used, *i.e.* e-value $< 10^{-5}$ (**Additional file 2; Figure S2**).

We also examined the effect of the interval between SNPs, because we would expect clusters of spurious SNPs in poorly sequenced regions. We tested the effect of selecting SNPs in regions without any other SNPs. As expected, there was an excess of shared polymorphisms (class 7) with a small window size. When the threshold was set to > 50 bp on each side of the SNP, the distribution was stable (**Additional file 2; Figure S2**).

Finally, we considered that the lowest value of minor allele frequency (MAF) in the lab populations should be set around 5% because the effective population size in the lab is low. All the above thresholds, apart from that for MAF, are trade-offs between quality and quantity of the data. The lowest MAF value possible in the pooled embryo samples depends on the unknown number of parents of the embryos, and the MAF threshold of >5% could therefore be considered arbitrary. Nevertheless, using MAF thresholds of 1%, 5% and 10% we obtained similar SNP class frequencies (**Table 1** and **Additional file 2; Table S1a** and **Table S1b**).

The results were thus also robust according to this parameter, and the use of different sets of parameters led to similar distribution of SNP classes that led to the same conclusion.

Therefore, all analyses in this paper were performed using the following thresholds: MAF > 5%; depth > 100; e-value < 10^{-5}; SNP isolation > 50 bp.

Tests of the reliability of the observed polymorphisms

First of all we examined whether Illumina sequencing generated polymorphism artifacts due to sequencing errors. To evaluate the extent of this bias we looked at the mitochondrial gene polymorphisms. Since the transmission of this genome is clonal, we would expect to find fixed differences between populations and no shared polymorphisms. Some polymorphic sites within a population would be expected if several haplotypes coexist. We found 18 SNPs in the mitochondrial genome that are fixed differences between Texas SF and Pachón CF. Most of them were synonymous (15/18) and most of them were transitions C:G <-> T:A (17/18) as expected. The frequency of the sum of the minor alleles was about 0.1%. These results suggest that using the threshold MAF > 5%, the within population polymorphic sites we identified were not the results of sequencing errors, which has a much lower level (i.e. < 1%) (**Additional file 2; Table S6**).

Using SNPs with known frequencies, we tried to evaluate if estimation of allele frequencies were biased and the extent of the standard error. We looked at the frequencies of derived alleles in different organs (brain and olfactory epithelium) of two surface fish and two cavefish. As in one individual, polymorphic sites were heterozygous sites, the expected frequency of the derived allele is 0.5. When we looked at the distribution of the derived allele frequencies in these eight samples, we found a symmetric distribution centered on 0.50 with a low standard deviation (0.14), suggesting that estimations of allele frequencies are not biased even if they are not very accurate. Moreover we confirmed that sequencing did not generate a large number of artefactual polymorphism that would have been detected as an excess of derived alleles at low frequencies (**Additional file 2; Figure S9**).

Estimation of Pro106Leu *MAO* polymorphism in Pachón natural population and laboratory stock

Genomic DNA was extracted from fin-clips of 20 Pachón wild-caught individuals and 31 individuals of our laboratory stock of Pachón cavefish obtained in 2004-2006 from Jeffery laboratory at the University of Maryland, College Park, MD, USA, and since then bred in our local facility. PCR was performed to amplify MAO exon4. Each PCR product was sequenced to identify the genotype at the codon which encodes the amino-acid 106.

Simulations of the evolution of neutral polymorphisms in the populations

In order to estimate the age of the Pachón cave population, we compared the distribution of SNPs into seven classes (the divergent polymorphism class was empty and thus excluded) defined above with the distribution obtained in simulations of the evolutionary process (**Figure 1**). The full model is as following: an ancestral population with a given size and at mutation/drift equilibrium (which depends on the mutation rate and the population size) was split into two populations that could have different sizes. After a delay, one population settled in a cave. Following a bottleneck this population could have a new size. Migrations between the populations could also be simulated. The delay and the bottleneck could be set to zero. We also took into account that genetic drift could have occurred in the laboratory stocks. All mutations were neutral and each locus evolved independently. For a given set of parameters, each ten generations, we estimated the frequency of SNPs in each category and we estimated a score of goodness of fit with the observed frequencies. We ran the simulation and the test of goodness of fit with different sets of parameters in order to identify the sets of parameters, including the age of the Pachón cave population, that resulted in SNP frequency in each class that fitted well with observed frequencies (see for more details **Additional file 1**). The program was written in C and is available on Github (http://github.com/julienfumey/popsim).

Data storage and analyses

SNPs and their annotations are stored in a MySQL database and are available online at http://ngspipelines.toulouse.inra.fr:9022. Perl and R scripts for the data analyses and graphics are available upon request.

Additional Material

Additional file 1:

Additional file 2:

Additional file 3:

Acknowledgments

This work has benefited from the facilities and expertise of the high throughput sequencing platform of IMAGIF (Centre de Recherche de Gif - www.imagif.cnrs.fr). The work was supported by a collaborative ANR (Agence Nationale de la Recherche) grant BLINDTEST and IDEEV.

Authors’ contributions

DC and SR designed the study. JF wrote the program of simulation and analyzed the data. SR, HH collected the data. CN and JF generated the databases. DC drafted the manuscript. All authors contributed to the writing of the manuscript. All authors read and approved the final manuscript.

References

1. 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.
2. Gross JB: **The complex origin of *Astyanax* cavefish**. *BMC Evol Biol* 2012, **12**:105.
3. Jeffery WR: **Regressive evolution in *Astyanax* cavefish**. *Annu Rev Genet* 2009, **43**:25-47.
4. Jeffery WR: **Emerging model systems in evo-devo: cavefish and microevolution of development**. *Evol Dev* 2008, **10**(3):265-272.
5. Wilkens H, Strecker U: **Convergent evolution of the cavefish *Astyanax* (Characidae, Teleostei): genetic evidence from reduced eye-size and pigmentation**. *Biological Journal of the Linnean Society* 2003, **80**(4):545-554.
6. Hüppop K: **Oxygen-consumption of Astyanax-fasciatus (Characidae, Pisces) - A comparison of epigean and hypogean populations**. *Environmental Biology of Fishes* 1986, **17**(4):299-308.
7. Moran D, Softley R, Warrant EJ: **Eyeless Mexican Cavefish Save Energy by Eliminating the Circadian Rhythm in Metabolism**. *PLoS ONE* 2014, **9**(9):e107877.
8. Salin K, Voituron Y, Mourin J, Hervant F: **Cave colonization without fasting capacities: an example with the fish *Astyanax fasciatus mexicanus***. *Comparative biochemistry and physiology Part A, Molecular & integrative physiology* 2010, **156**(4):451-457.
9. Hüppop K, Wilkens H: **Bigger eggs in subterranean *Astyanax fasciatus* (Characidae, Pisces) - their significance and genetics**. *Zeitschrift Fur Zoologische Systematik Und Evolutionsforschung* 1991, **29**(4):280-288.
10. Teyke T: **Morphological differences in neuromasts of the blind cave fish Astyanax hubbsi and the sighted river fish Astyanax mexicanus**. *Brain Behavior and Evolution* 1990, **35**(1):23-30.
11. Yoshizawa M, Goricki S, Soares D, Jeffery WR: **Evolution of a behavioral shift mediated by superficial neuromasts helps cavefish find food in darkness**. *Curr Biol* 2010, **20**(18):1631-1636.
12. Yoshizawa M, Yamamoto Y, O'Quin KE, Jeffery WR: **Evolution of an adaptive behavior and its sensory receptors promotes eye regression in blind cavefish**. *BMC biology* 2012, **10**:108.
13. Duboué ER, Borowsky RL, Keene AC: **beta-adrenergic signaling regulates evolutionarily derived sleep loss in the Mexican cavefish**. *Brain, behavior and evolution* 2012, **80**(4):233-243.
14. Duboué ER, Keene AC, Borowsky RL: **Evolutionary convergence on sleep loss in cavefish populations**. *Curr Biol* 2011, **21**(8):671-676.
15. Elipot Y, Hinaux H, Callebert J, Retaux S: **Evolutionary shift from fighting to foraging in blind cavefish through changes in the serotonin network**. *Curr Biol* 2013, **23**(1):1-10.
16. Varatharasan N, Croll RP, Franz-Odendaal T: **Taste bud development and patterning in sighted and blind morphs of *Astyanax mexicanus***. *Dev Dyn* 2009, **238**(12):3056-3064.

17. Yamamoto Y, Byerly MS, Jackman WR, Jeffery WR: **Pleiotropic functions of embryonic sonic hedgehog expression link jaw and taste bud amplification with eye loss during cavefish evolution**. *Dev Biol* 2009, **330**(1):200-211.
18. Bibliowicz J, Alie A, Espinasa L, Yoshizawa M, Blin M, Hinaux H, Legendre L, Pere S, Retaux S: **Differences in chemosensory response between eyed and eyeless *Astyanax mexicanus* of the Rio Subterraneo cave**. *EvoDevo* 2013, **4**(1):25.
19. Espinasa L, Bibliowicz J, Jeffery W, Retaux S: **Enhanced prey capture skills in *Astyanax* cavefish larvae are independent from eye loss**. *EvoDevo* 2014, **5**(1):35.
20. Hüppop K: **Food finding ability in cave fish *(Astyanax fasciatus)***. *Int J Speleol* 1987, **18**:59-66.
21. Mayr E: **Cause and effect in biology**. *Science* 1961, **134**(3489):1501-1506.
22. Casane D, Rétaux S: **Evolutionary Genetics of the Cavefish Astyanax mexicanus**. In: *Advances in Genetics.* Edited by Nicholas SF, vol. Volume 95: Academic Press; 2016: 117-159.
23. Ma L, Jeffery WR, Essner JJ, Kowalko JE: **Genome Editing Using TALENs in Blind Mexican Cavefish, *Astyanax mexicanus***. *PLoS ONE* 2015, **10**(3):e0119370.
24. McGaugh SE, Gross JB, Aken B, Blin M, Borowsky R, Chalopin D, Hinaux H, Jeffery WR, Keene A, Ma L *et al*: **The cavefish genome reveals candidate genes for eye loss**. *Nat Commun* 2014, **5**:5307.
25. O'Quin KE, Yoshizawa M, Doshi P, Jeffery WR: **Quantitative genetic analysis of retinal degeneration in the blind cavefish *Astyanax mexicanus***. *PLoS One* 2013, **8**(2):e57281.
26. Hinaux H, Blin M, Fumey J, Legendre L, Heuze A, Casane D, Retaux S: **Lens Defects in *Astyanax mexicanus* Cavefish: Evolution of Crystallins and a Role for alphaA-Crystallin**. *Developmental Neurobiology* 2015, **75**(5):505-521.
27. Jeffery WR: **Pleiotropy and eye degeneration in cavefish**. *Heredity* 2010, **105**(5):495-496.
28. Wilkens H: **Genes, modules and the evolution of cave fish**. *Heredity* 2010, **105**(5):413-422.
29. Borowsky R: **Eye regression in blind *Astyanax* cavefish may facilitate the evolution of an adaptive behavior and its sensory receptors**. *BMC biology* 2013, **11**(1):81.
30. Gross JB, Powers AK, Davis EM, Kaplan SA: **A pleiotropic interaction between vision loss and hypermelanism in Astyanax mexicanus cave x surface hybrids**. *BMC Evolutionary Biology* 2016, **16**(1):1-16.
31. Retaux S, Casane D: **Evolution of eye development in the darkness of caves: adaptation, drift, or both?** *EvoDevo* 2013, **4**(1):26.
32. Avise JC, Selander RK: **Evolutionary genetics of cave-dwelling fishes of genus *Astyanax***. *Evolution* 1972, **26**(1):1-19.
33. Bradic M, Beerli P, Garcia-de Leon FJ, Esquivel-Bobadilla S, Borowsky RL: **Gene flow and population structure in the Mexican blind cavefish complex (*Astyanax mexicanus*)**. *BMC Evol Biol* 2012, **12**:9.
34. Hausdorf B, Wilkens H, Strecker U: **Population genetic patterns revealed by microsatellite data challenge the mitochondrial DNA based taxonomy of Astyanax in Mexico (Characidae, Teleostei)**. *Mol Phylogenet Evol* 2011, **60**(1):89-97.
35. Panaram K, Borowsky R: **Gene flow and genetic variability in cave and surface populations of the Mexican Tetra, Astyanax mexicanus (Telcostei : Characidae)**. *Copeia* 2005(2):409-416.
36. Strecker U, Bernatchez L, Wilkens H: **Genetic divergence between cave and surface populations of *Astyanax* in Mexico (Characidae, Teleostei)**. *Molecular ecology* 2003, **12**(3):699-710.
37. Strecker U, Faundez VH, Wilkens H: **Phylogeography of surface and cave *Astyanax* (Teleostei) from Central and North America based on cytochrome b sequence data**. *Mol Phylogenet Evol* 2004, **33**(2):469-481.
38. Strecker U, Hausdorf B, Wilkens H: **Parallel speciation in Astyanax cave fish (Teleostei) in Northern Mexico**. *Mol Phylogenet Evol* 2012, **62**(1):62-70.
39. Dowling TE, Martasian DP, Jeffery WR: **Evidence for multiple genetic forms with similar eyeless phenotypes in the blind cavefish, *Astyanax mexicanus***. *Mol Biol Evol* 2002, **19**(4):446-455.

40. Barr TC: **Cave ecology and the evolution of troglobites**. In: *Evolutionary Biology.* Edited by Press P, vol. 2. New York; 1968: 35-102.

41. Chakraborty R, Nei M: **Dynamics of gene differentiation between incompletely isolated populations of unequal sizes**. *Theoretical Population Biology* 1974, **5**(3):460-469.

42. Ornelas-García CP, Domínguez-Domínguez O, Doadrio I: **Evolutionary history of the fish genus Astyanax Baird & Girard (1854) (Actinopterygii, Characidae) in Mesoamerica reveals multiple morphological homoplasies**. *BMC Evolutionary Biology* 2008, **8**(1):1-17.

43. Yokoyama R, Yokoyama S: **Convergent evolution of the red- and green-like visual pigment genes in fish, Astyanax fasciatus, and human**. *Proc Natl Acad Sci U S A* 1990, **87**(23):9315-9318.

44. Yokoyama R, Yokoyama S: **Molecular characterization of a blue visual pigment gene in the fish Astyanax fasciatus**. *FEBS Lett* 1993, **334**(1):27-31.

45. Yokoyama S, Meany A, Wilkens H, Yokoyama R: **Initial mutational steps toward loss of opsin gene function in cavefish**. *Mol Biol Evol* 1995, **12**(4):527-532.

46. Li W-H, Nei M: **Persistence of common alleles in two related populations or species**. *Genetics* 1977, **86**(4):901-914.

47. Cavallari N, Frigato E, Vallone D, Froehlich N, Fernando Lopez-Olmeda J, Foa A, Berti R, Javier Sanchez-Vazquez F, Bertolucci C, Foulkes NS: **A Blind Circadian Clock in Cavefish Reveals that Opsins Mediate Peripheral Clock Photoreception**. *Plos Biology* 2011, **9**(9):e1001142.

48. Yang J, Chen X, Bai J, Fang D, Qiu Y, Jiang W, Yuan H, Bian C, Lu J, He S *et al*: **The Sinocyclocheilus cavefish genome provides insights into cave adaptation**. *BMC biology* 2016, **14**(1):1-13.

49. Niemiller ML, Fitzpatrick BM, Shah P, Schmitz L, Near TJ: **Evidence for repeated loss of selective constraint in rhodopsin of amblyopsid cavefishes (teleostei: amblyopsidae)**. *Evolution* 2013, **67**(3):732-748.

50. Schlotterer C, Tobler R, Kofler R, Nolte V: **Sequencing pools of individuals - mining genome-wide polymorphism data without big funding**. *Nat Rev Genet* 2014, **15**(11):749-763.

51. Grantham R: **Amino acid difference formula to help explain protein evolution**. *Science* 1974, **185**(4154):862-864.

52. Elipot Y, Hinaux H, Callebert J, Launay J-M, Blin M, Rétaux S: **A mutation in the enzyme monoamine oxidase explains part of the Astyanax cavefish behavioural syndrome**. *Nat Commun* 2014, **5**:3647.

53. 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.

54. Goldstein DB, Linares AR, Cavallisforza LL, Feldman MW: **An evaluation of genetic distances for use with microsatellite loci**. *Genetics* 1995, **139**(1):463-471.

55. Moran PAP: **Wandering distributions and the electrophoretic profile**. *Theoretical Population Biology* 1975, **8**(3):318-330.

56. Goldstein DB, Linares AR, Cavallisforza LL, Feldman MW: **Genetic absolute dating based on microsatellites and the origin of modern humans**. *Proceedings of the National Academy of Sciences of the United States of America* 1995, **92**(15):6723-6727.

57. Shriner D, Tekola-Ayele F, Adeyemo A, Rotimi CN: **Genome-wide genotype and sequence-based reconstruction of the 140,000 year history of modern human ancestry**. *Scientific Reports* 2014, **4**:6055.

58. Yue GH, David L, Orban L: **Mutation rate and pattern of microsatellites in common carp (Cyprinus carpio L.)**. *Genetica* 2007, **129**(3):329-331.

59. Beerli P, Felsenstein J: **Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach**. *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**(8):4563-4568.

60. Pinho C, Hey J: **Divergence with Gene Flow: Models and Data**. In: *Annual Review of Ecology, Evolution, and Systematics, Vol 41.* Edited by Futuyma DJ, Shafer HB, Simberloff D, vol. 41; 2010: 215-230.
61. Kimura M: **Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution**. *Nature* 1977, **267**(5608):275-276.
62. Chamary JV, Parmley JL, Hurst LD: **Hearing silence: non-neutral evolution at synonymous sites in mammals**. *Nat Rev Genet* 2006, **7**(2):98-108.
63. Kimura M: **Evolutionary rate at molecular level**. *Nature* 1968, **217**(5129):624-626.
64. Kimura M, Ohta T: **Average number of generations until fixation of a mutant gene in a finite population**. *Genetics* 1969, **61**(3):763-771.
65. Winemiller KO: **Patterns of variation in life-history among South-American fishes in seasonal environments**. *Oecologia* 1989, **81**(2):225-241.
66. Avise JC: **Phylogeography: The History and Formation of Species**. Harvard: Harvard University Press; 2000.
67. Hedgecock D: **Does variance in reproductive success limit effective population sizes of marine organisms?** In: *Genetics and evolution of aquatic organisms.* Edited by Beaumont AR. London: Chapman & Hall; 1994: 122-134.
68. Rohner N, Jarosz DF, Kowalko JE, Yoshizawa M, Jeffery WR, Borowsky RL, Lindquist S, Tabin CJ: **Cryptic Variation in Morphological Evolution: HSP90 as a Capacitor for Loss of Eyes in Cavefish**. *Science* 2013, **342**(6164):1372-1375.
69. Crow JF, Kimura M: **An introduction to population genetics theory**. New York: Harper & Row; 1970.
70. Gross JB, Borowsky R, Tabin CJ: **A novel role for Mc1r in the parallel evolution of depigmentation in independent populations of the cavefish *Astyanax mexicanus***. *PLoS Genet* 2009, **5**(1):e1000326.
71. Bradic M, Teotónio H, Borowsky RL: **The Population Genomics of Repeated Evolution in the Blind Cavefish *Astyanax mexicanus***. *Molecular Biology and Evolution* 2013, **30**(11):2383-2400.
72. Kowalko JE, Rohner N, Linden TA, Rompani SB, Warren WC, Borowsky R, Tabin CJ, Jeffery WR, Yoshizawa M: **Convergence in feeding posture occurs through different genetic loci in independently evolved cave populations of *Astyanax mexicanus***. *Proceedings of the National Academy of Sciences* 2013, **110**(42):16933-16938.
73. Aspiras AC, Rohner N, Martineau B, Borowsky RL, Tabin CJ: **Melanocortin 4 receptor mutations contribute to the adaptation of cavefish to nutrient-poor conditions**. *Proceedings of the National Academy of Sciences* 2015, **112**(31):9668-9673.
74. Espinasa L, Espinasa M: **Hydrogeology of Caves in the Sierra de El Abra Region**. In: *Biology and Evolution of the Mexican Cavefish.* Edited by Keene AC, Yoshizawa M, McGaugh SE. Amsterdam: Academic Press; 2016: 41-58.
75. Protas ME, Hersey C, Kochanek D, Zhou Y, Wilkens H, Jeffery WR, Zon LI, Borowsky R, Tabin CJ: **Genetic analysis of cavefish reveals molecular convergence in the evolution of albinism**. *Nat Genet* 2006, **38**(1):107-111.
76. Bobadilla JL, Macek M, Fine JP, Farrell PM: **Cystic fibrosis: A worldwide analysis of CFTR mutations - Correlation with incidence data and application to screening**. *Human Mutation* 2002, **19**(6):575-606.
77. Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA: **Population Genomics of Parallel Adaptation in Threespine Stickleback using Sequenced RAD Tags**. *Plos Genetics* 2010, **6**(2).
78. Renaut S, Nolte AW, Rogers SM, Derome N, Bernatchez L: **SNP signatures of selection on standing genetic variation and their association with adaptive phenotypes along gradients of ecological speciation in lake whitefish species pairs (Coregonus spp.)**. *Molecular ecology* 2011, **20**(3):545-559.

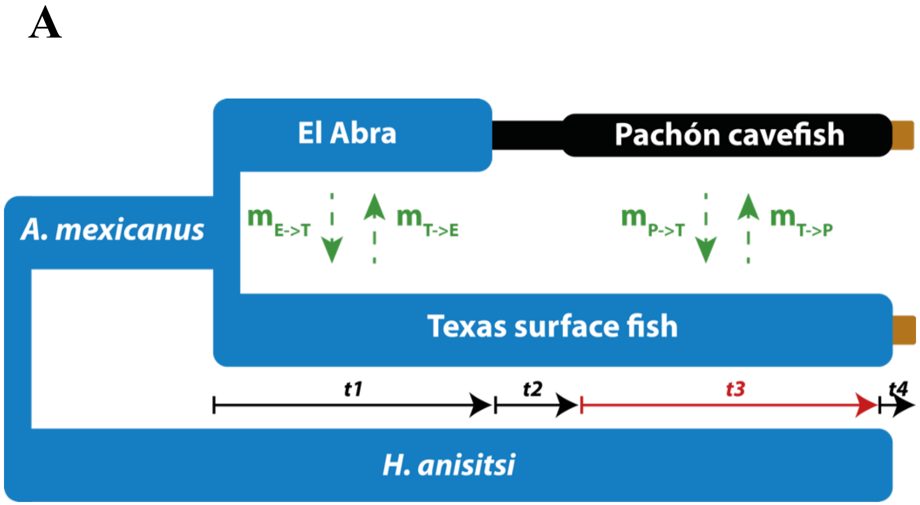
79. Johnson TC, Scholz CA, Talbot MR, Kelts K, Ricketts RD, Ngobi G, Beuning K, Ssemmanda I, McGill JW: **Late pleistocene desiccation of Lake Victoria and rapid evolution of cichlid fishes**. *Science* 1996, **273**(5278):1091-1093.
80. Brawand D, Wagner CE, Li YI, Malinsky M, Keller I, Fan SH, Simakov O, Ng AY, Lim ZW, Bezault E *et al*: **The genomic substrate for adaptive radiation in African cichlid fish**. *Nature* 2014, **513**(7518):375-381.
81. Paaby AB, Rockman MV: **Cryptic genetic variation: evolution's hidden substrate**. *Nat Rev Genet* 2014, **15**(4):247-258.
82. Akashi H, Osada N, Ohta T: **Weak selection and protein evolution**. *Genetics* 2012, **192**(1):15-31.
83. Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, Adams MD, Schmidt S, Sninsky JJ, Sunyaev SR *et al*: **Assessing the evolutionary impact of amino acid mutations in the human genome**. *PLoS Genet* 2008, **4**(5):e1000083.
84. Eyre-Walker A, Keightley PD, Smith NG, Gaffney D: **Quantifying the slightly deleterious mutation model of molecular evolution**. *Mol Biol Evol* 2002, **19**(12):2142-2149.
85. Eyre-Walker A, Woolfit M, Phelps T: **The distribution of fitness effects of new deleterious amino acid mutations in humans**. *Genetics* 2006, **173**(2):891-900.
86. Kousathanas A, Keightley PD: **A comparison of models to infer the distribution of fitness effects of new mutations**. *Genetics* 2013, **193**(4):1197-1208.
87. Nielsen R, Yang Z: **Estimating the Distribution of Selection Coefficients from Phylogenetic Data with Applications to Mitochondrial and Viral DNA**. *Molecular Biology and Evolution* 2003, **20**(8):1231-1239.
88. Lohmueller KE, Indap AR, Schmidt S, Boyko AR, Hernandez RD, Hubisz MJ, Sninsky JJ, White TJ, Sunyaev SR, Nielsen R *et al*: **Proportionally more deleterious genetic variation in European than in African populations**. *Nature* 2008, **451**(7181):994-997.
89. Do R, Balick D, Li H, Adzhubei I, Sunyaev S, Reich D: **No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans**. *Nature Genetics* 2015, **47**(2):126-131.
90. Darwin CR: **On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life**, 1st ed. edn. London: John Murray; 1859.
91. 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.
92. Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S *et al*: **Ensembl 2013**. *Nucleic Acids Research* 2013, **41**(D1):D48-D55.
93. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2009, **25**(14):1754-1760.
94. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M *et al*: **The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data**. *Genome Research* 2010, **20**(9):1297-1303.

Legends

Figure 1 Analysis of polymorphism in *Astyanax mexicanus* Texas surface *vs* Pachón cave population, using *Hyphessobrycon anisitsi* as outgroup. (A) Evolutionary model. (B) The eight SNP classes correspond to the polymorphism patterns that can be found within and between two populations. Class 1: Different fixed alleles in each population, derived allele in cavefish; Class 2: Different fixed alleles in each population, derived allele in surface fish; Class 3: Polymorphism in cavefish, ancestral fixed allele in surface fish; Class 4: Polymorphism in cavefish, derived fixed allele in surface fish; Class 5: Polymorphism in surface fish, ancestral fixed allele in cavefish; Class 6: Polymorphism in surface fish, derived fixed allele in cavefish; Class 7: Shared polymorphism; Class 8: Divergent polymorphism. x, y and z can be one of the four nucleotides A, T, G, C.

Figure 2 Goodness of fit to the data. The model parameters are: SF population size = 10,000; CF population size = 625; % migrants from surface to cave = 0.1; migration rate from surface to cave = 0.001 / year; SF generation time = 2 years; CF generation time = 5 years; lab population parameters: 10 fish, 10 generations. All the other parameters were set to zero. (A) Score of goodness of fit according to the age of the cave population (t3), the best fit is when the cavefish population is 25,500 years old. (B) Evolution of the SNP class frequencies during the simulation. Horizontal dotted lines are the observed SNP class frequencies. Observed and simulated frequencies at the age of the best fit are shown in the top right corner. (C) Evolution of the number of polymorphic sites in SF and CF during the simulation. (D) Evolution of the number of derived alleles that were fixed in SF and CF during the simulation. (E) Evolution of the SF/CF polymorphism ratio and the CF/SF derived allele ratio that reached fixation during the simulation. Horizontal dotted lines are the observed ratios. The vertical dotted line is the age of the cavefish population for which the best fit was observed.

Figure 3 Conservative and radical substitutions in CF and SF. (A) Numbers of substitutions. (B) relative frequencies.



B

y	*x*	*x/y*	*x/y*	*x*	*y*	*x/y*	*x/z*
x	*y*	*x*	*y*	*x/y*	*x/y*	*x/y*	*x/y*
x	*x*	*x*	*x*	*x*	*x*	*x*	*x*
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*

SNP class

Figure 1

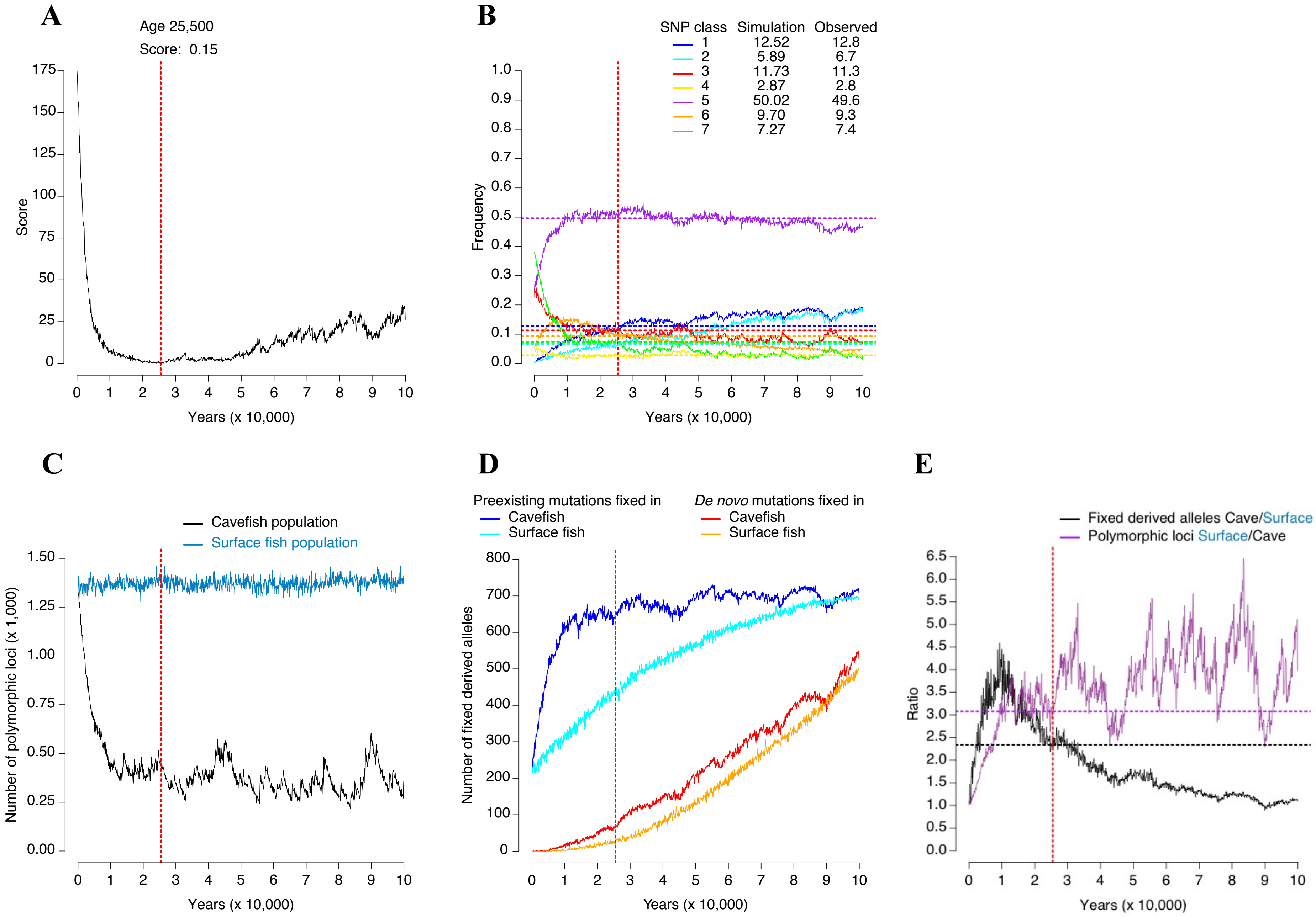

A
Age 25,500
Score: 0.15
Score
175
150
125
100
75
50
25
0
0 1 2 3 4 5 6 7 8 9 10
Years (x 10,000)
B
SNP class Simulation Observed
1 12.52 12.8
2 5.89 6.7
3 11.73 11.3
4 2.87 2.8
5 50.02 49.6
6 9.70 9.3
7 7.27 7.4
Frequency
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0 1 2 3 4 5 6 7 8 9 10
Years (x 10,000)
C
Cavefish population
Surface fish population
Number of polymorphic loci (x 1,000)
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0 1 2 3 4 5 6 7 8 9 10
Years (x 10,000)
D
Preexisting mutations fixed in
Cavefish
Surface fish
De novo mutations fixed in
Cavefish
Surface fish
Number of fixed derived alleles
800
700
600
500
400
300
200
100
0
0 1 2 3 4 5 6 7 8 9 10
Years (x 10,000)
E
Fixed derived alleles Cave/Surface
Polymorphic loci Surface/Cave
Ratio
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0 1 2 3 4 5 6 7 8 9 10
Years (x 10,000)

Figure 2

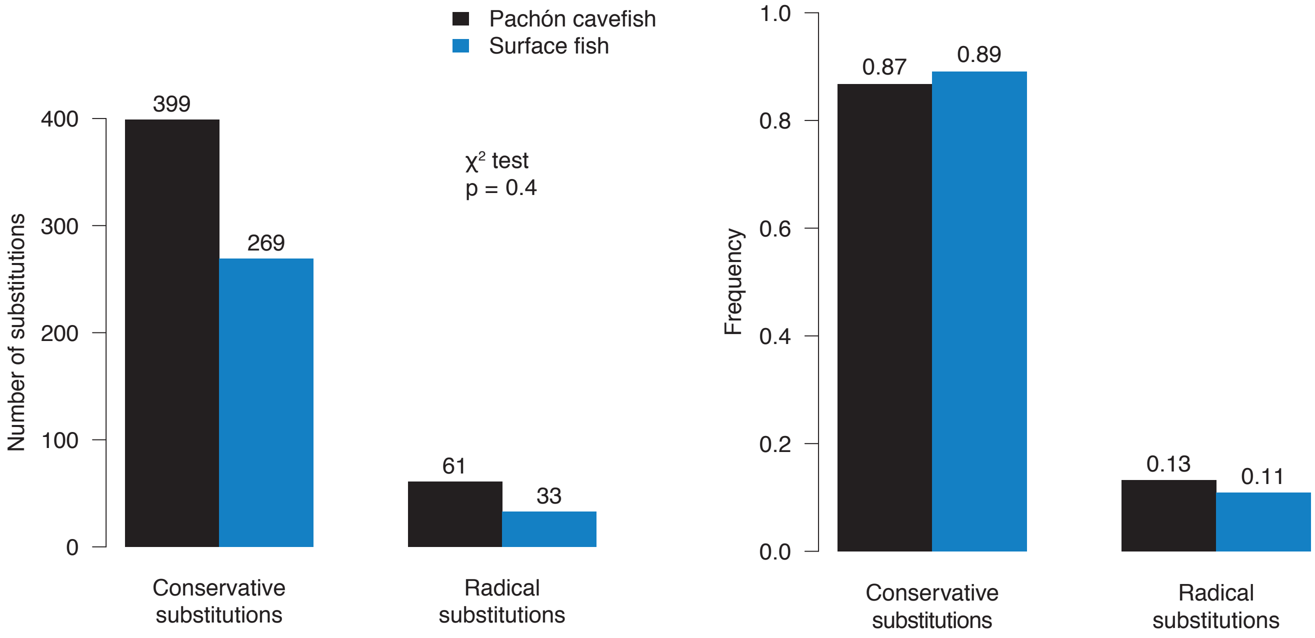

■ Pachón cavefish
■ Surface fish
400
300
200
100
0
Number of substitutions
399
269
61
33
χ^2 test
p = 0.4
Conservative substitutions
Radical substitutions
1.0
0.8
0.6
0.4
0.2
0.0
Frequency
0.87
0.89
0.13
0.11
Conservative substitutions
Radical substitutions

Figure 3

Table 1. Classification of polymorphisms in *Astyanax mexicanus* Texas surface *vs* Pachón cave populations

		Synonymous		Non-coding		Non-synonymous		Conservative		Radical		New Stop		Stop loss	
	Class	n	*%*	n	*%*	n	*%*	n	*%*	n	*%*	n	*%*	n	*%*
Ancestral fixed SF, derived fixed CF	(1)	540	*12.8*	157	*12.7*	301	*14.3*	254	*13.7*	47	*18.5*	4	*57.1*	0	*0.0*
Ancestral fixed CF, derived fixed SF	(2)	280	*6.7*	111	*9.0*	211	*10.0*	188	*10.1*	23	*9.1*	0	*0.0*	0	*0.0*
Polymorphism CF, ancestral fixed SF	(3)	476	*11.3*	146	*11.8*	302	*14.5*	269	*14.5*	33	*13.0*	2	*28.6*	0	*0.0*
Polymorphism CF, derived fixed SF	(4)	119	*2.8*	57	*4.6*	91	*4.4*	81	*4.4*	10	*3.9*	0	*0.0*	0	*0.0*
Polymorphism SF, ancestral fixed CF	(5)	2,086	*49.6*	601	*48.5*	923	*43.6*	809	*43.6*	114	*44.9*	1	*14.3*	0	*0.0*
Polymorphism SF, derived fixed CF	(6)	393	*9.3*	87	*7.0*	159	*7.8*	145	*7.8*	14	*5.5*	0	*0.0*	0	*0.0*
Shared polymorphism	(7)	309	*7.4*	80	*6.5*	123	*5.9*	110	*5.9*	13	*5.1*	0	*0.0*	0	*0.0*
Divergent	(8)	1	*0.0*	0	*0.0*	0	*0.0*	0	*0.0*	0	*0.0*	0	*0.0*	0	*0.0*
Total		4,204	*100.0*	1,239	*100.0*	2,110	*100.0*	1,856	*100.0*	254	*100.0*	7	*100.0*	0	*0.0*
Polymorphism SF	(5+6+7)	2,788		768		1,205		1,064		141		1		0	
Polymorphism CF	(3+4+7)	904		283		516		460		56		2		0	
Ratio SF/CF		3.08		2.71		2.34		2.31		2.52		0.5		n.a.	
Derived and fixed SF	(2+4)	399		168		302		269		33		0		0	
Derived and fixed CF	(1+6)	933		244		460		399		61		4		0	
Ratio CF/SF		2.34		1.45		1.52		1.48		1.85		n.a.		n.a.	

Thresholds: 100; MAF > 5%; Score Blast < 10^{-5}; interval > 50bp (see materials and methods for threshold definitions).
CF: Cavefish; SF: Surface fish; numbers in brackets are class identifiers described in Figure 1.