Tempo et mode de l'évolution des populations cavernicoles de l'espèce Astyanax mexicanus

Julien Fumey

NNT : 2016SACLS528

THÈSE DE DOCTORAT
DE
L'UNIVERSITÉ PARIS SACLAY
PRÉPARÉE
À
L'UNIVERSITÉ PARIS SUD

École doctorale n°577
SDSV Structure et Dynamique des Systèmes Vivants (SDSV)

Spécialité du doctorat
Sciences de la Vie et de la Santé

par
M. Julien Fumey

Tempo et mode de l'évolution des populations cavernicoles de l'espèce *Astyanax mexicanus*

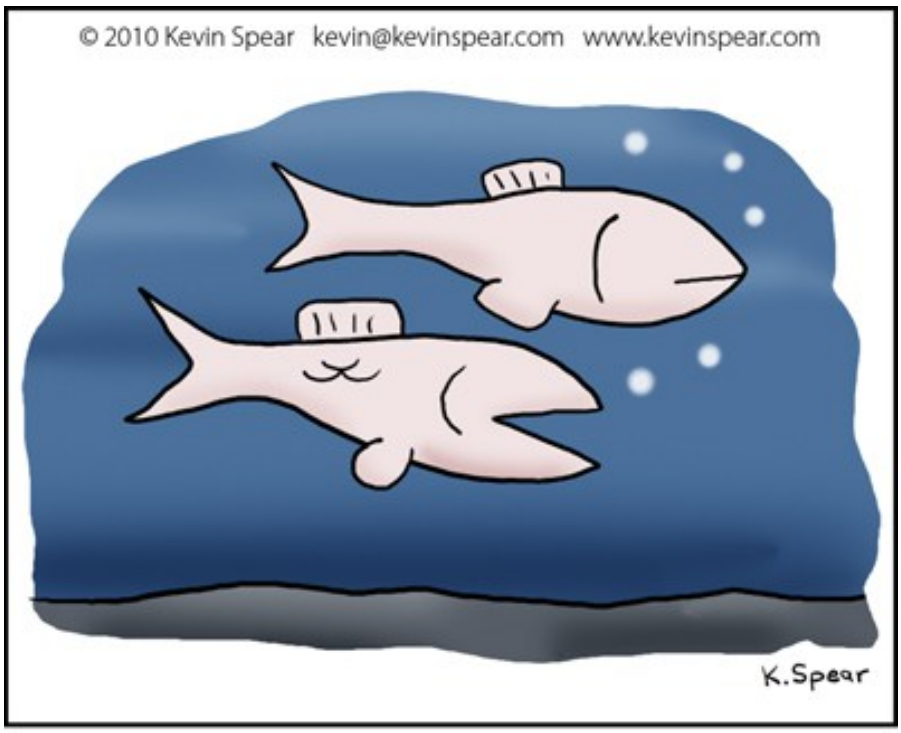
Thèse présentée et soutenue à Gif-sur-Yvette, le 12 décembre 2016

Composition du jury :

M. Pierre Capy	Professeur, Université Paris Sud	Président du jury
M. Guillaume Achaz	Maître de Conférences, Université Pierre et Marie Curie	Rapporteur
M. Christophe Douady	Professeur, Université Claude Bernard - Lyon 1	Rapporteur
M. Jean-François Agnèse	Directeur de Recherche, ISEM	Examinateur
M. Didier Casane	Professeur, Université Paris Diderot	Directeur de thèse

Julien Fumey

Tempo et mode de l'évolution des populations cavernicoles de l'espèce *Astyanax mexicanus*



"The one thing I regret about being a blind cave fish is I'll never be able to see the desert."

"Mon seul regret en tant que poisson cavernicole aveugle est que je n'aurai jamais la possibilité de voir le désert."

REMERCIEMENTS

Voici donc la partie la plus difficile à écrire mais aussi la plus importante d'une thèse : les remerciements. Ceux que je pourrais oublier m'excuseront.

Je tiens en premier lieu à remercier les membres du jury d'avoir accepté d'évaluer mon travail.

Cette thèse n'aurait pu voir le jour sans un grand nombre de personnes. Je tiens donc à remercier l'ensemble des membres de l'équipe RESGEN. Merci à Didier de m'avoir accueilli dans son équipe et de m'avoir initié à la génétique des populations. Le travail fût parfois intense mais a été ponctué de moments de détente bienvenus. Le terrain de tennis près de la Mérantaise se souvient encore de ~~quelques~~ nombreuses balles ratées. Je remercie Alice pour m'avoir donné le goût de la vulgarisation et avoir été à mon écoute dans certains moments difficiles. Merci pour la relecture attentive de cette thèse à la traque de la moindre petite faute d'orthographe (j'espère les avoir toutes supprimées cette fois !). Merci à Patrick pour les passionnantes discussions sur la politique, sur l'anti-darwinisme et pour son amour des trolls poilus. Merci aussi à Isa, Véro, Magalie et Jean-Luc.

Merci également à l'équipe du café du matin notamment David, Arnaud, Sylvie et Émilie, qui permettait de commencer la journée en pleine forme et de bonne humeur (mais également frigorifié en hiver). Vous allez devoir faire votre café vous-mêmes maintenant !

Cette thèse aurait été bien moins agréable sans la bonne humeur (la folie ?) des neurobiologistes de l'équipe DÉCA à l'INAF. Les deux missions d'échantillonnage et d'expérimentation auxquelles j'ai pu participer avec Sylvie, Hélène, Stéphane, Yoni, Alex, Maryline, Lucie, Laurent, Yannick, Victor et Carole ont permis de mieux comprendre les *Astyanax* et de me familiariser un peu avec la « bioinformatique de terrain » (même si ma cheville gauche n'a pas beaucoup apprécié la visite de la grotte Pachón...).

Un grand merci à Céline de la plateforme GénoToul de l'INRA de Toulouse pour son aide sur la génomique et la transcriptomique.

Merci aux anciens et actuels doctorants du labo notamment Gwenaëlle, Quentin, Antoine, Florian, Hanna, Bastien et Estelle. Les pauses café ou coca avec Arnaud au milieu de l'après-midi permettaient de relancer la machine lorsque je rédigeais ce qui suit. Et promis, je vais arrêter de changer tes fonds d'écran... ou pas.

Le travail des stagiaires, Nina, Maxime et Jean ont permis de faire avancer cette thèse. Merci à eux !

Merci à l'équipe de la plateforme de séquençage IMAGIF de m'accueillir pour la suite.

Je remercie les ananas et Madame Pineau pour ses délicieuses tartes aux abricots et ses british mac.

Un grand merci à Cath' qui a su attiser ma curiosité scientifique lors de mes premières années de fac.

Il me faut également remercier les membre de l'association des Jeunes Bioinformaticiens qui m'ont fait confiance ces deux dernières années en me donnant la responsabilité de présider l'association. Cela fut un plaisir et m'a permis de décompresser. Un coup de chapeau à Léopold (coin-coin) qui m'a suppléé avec brio lors de ces derniers mois. Et merci à tous les membres du CA qui ont pu faire avancer les projets de l'association : Sylvain, Alex, Axelle, Micaela, David, Romy, Marouen, Malvina, Nina, Nolwenn, Manue, Cédric, Leopold, Gwenaëlle, Bérénice, Lambert, Hugo.

Je n'oublie pas l'ensemble de la communauté de Bioinfo-fr. Il a été très plaisant de venir ~~troller~~ discuter bioinfo et d'aider ou de recevoir de l'aide lorsque cela était nécessaire. Et évidemment de kicker Yo du canal le jour où il a nommé tout le monde opérateur ;).

Il me faut remercier tous mes ami·e·s qui m'ont supporté durant toutes ces années. Une mention spéciale à Pauline qui le fait depuis déjà 20 ans...

Merci à la jeune femme qui m'a supporté pendant ces mois de rédaction et qui a rendu la vie plus facile et plus douce durant cette période.

Merci à mes parents qui sont un peu responsables, d'une certaine façon, de ce qui suit, bien qu'ils ne comprennent pas grand chose à tout ce charabia. Merci à ma sœur pour les différents voyages à travers l'Europe.

TABLE DES MATIÈRES

INTRODUCTION

0.1 Les organismes cavernicoles : de bons modèles d'études en biologie évolutive

Les animaux hypogés* peuvent être répartis en trois catégories [1] : les trogloxènes qui vivent de façon temporaire sous terre, les troglophiles qui vivent principalement sous terre et les troglobies qui sont des hypogés obligatoires. Le nombre d'espèces cavernicoles décrites ou estimées est de 7 000 pour les espèces aquatiques et d'au moins 21 000 pour les espèces terrestres [2]. Selon une autre estimation, le nombre total d'espèces troglobies pourrait atteindre 50 000 à 100 000 à travers le monde [3].

Hypogés *: Animaux vivant sous-terre, contrairement aux animaux* **épigés**

Parmi les animaux troglobies, nous allons nous intéresser en particulier aux animaux cavernicoles, qui vivent de façon pérenne dans des grottes où l'obscurité est totale et permanente. Cette obscurité implique généralement une réduction des ressources alimentaires disponibles par l'absence de producteurs primaires qui ne peuvent réaliser de photosynthèse en l'absence de lumière. De plus l'isolement de la surface, permanent ou saisonnier, de certaines grottes empêche l'arrivée de ressources alimentaires par des flux d'eau. Cela a pour autre conséquence, pour ces espèces, une diminution, voire, dans certains cas, une absence de prédation.

Une des difficultés majeures dans l'étude de l'évolution et de l'adaptation est de définir l'état ancestral (plésiomorphie) et l'état dérivé (apomorphie) d'un caractère. L'étude des animaux cavernicoles permet de résoudre cette difficulté. En effet, en comparant une espèce cavernicole avec une espèce épigée proche, l'état dérivé devrait pour de nombreux caractères être celui présent chez l'espèce cavernicole.

C'est pourquoi les animaux cavernicoles sont de bons modèles en biologie évolutive, en particulier pour l'étude de l'adaptation à l'environnement et pour celle des processus évolutifs impliqués dans cette adaptation.

Les animaux cavernicoles sont retrouvés dans de nombreux groupes, surtout des arthropodes mais aussi quelques vertébrés (Figure 1). Il est possible d'observer une convergence évolutive* de ces espèces cavernicoles. En effet, les animaux cavernicoles possèdent généralement des caractéristiques communes. Certaines de ces caractéristiques, telles que l'absence de pigmentation ou la perte des yeux, sont facilement identifiables (Figure 1).

Convergence évolutive *: Acquisition d'un phénotype* similaire de façon indépendante par des espèces ou des populations soumises à une même contrainte environnementale.
Phénotype *: Caractère observé, issu de l'interaction du génotype et de l'environnement.*

La perte de la pigmentation et des yeux pourrait être due, comme le pensait Darwin dans l'*Origine des espèces* [4], à leur non-usage dans l'obscurité complète.

> « As it is difficult to imagine that eyes, although useless, could be in any way injurious to animals living in the darkness, I attribute

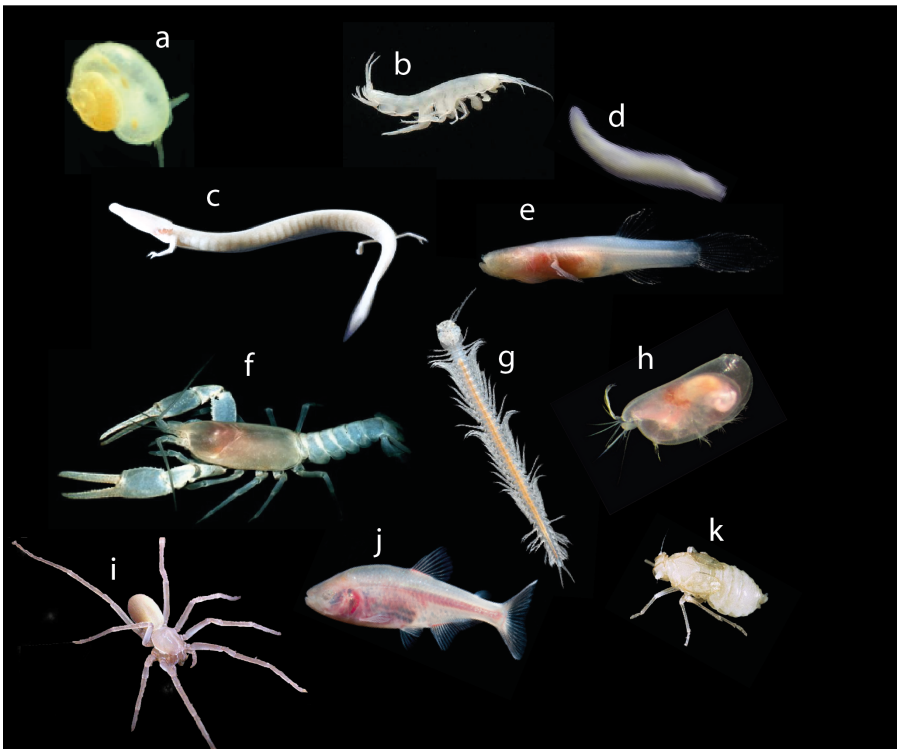


Figure 1. Exemples d'animaux cavernicoles. a. *Antrobia culveri* (Gastéropode) b. *Palaemonias ganteri* (Arthropode) c. *Proteus anguinus* (Amphibien) d. *Spalloplana sp.* (Planaire) e. *Amblyopsis rosae* (Téléostéen) f. *Cambarus tartarus* (Arthropode) g. *Xibalbanus tulumensis* (Arthropode) h. *Spelaeoecia capax* (Arthropode) i. *Amauropelma matakecil* (Arthropode) j. *Astyanax mexicanus* (Téléostéen) k. *Oliarus polyphemus* (Arthropode).

> their loss wholly to disuse [...] 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 ».

« Comme il est difficile de supposer que l'œil, bien qu'inutile, puisse être nuisible à des animaux vivant dans l'obscurité, j'attribue la perte de cet organe au non-usage [...] Je suis plutôt étonné que nous ne retrouvions pas dans les cavernes un plus grand nombre de vestiges de formes de vie passées, en raison du peu de concurrence à laquelle les habitants de ces sombres demeures ont été exposés ».

Contrairement à l'affirmation de Darwin, les animaux hypogés ne sont pas des « vestiges de formes de vie passées », mais au contraire des animaux dérivés et adaptés à leur environnement. En effet, outre la perte des yeux et de la pigmentation, ils possèdent souvent des caractères leur permettant de survivre dans les grottes comme, par exemple, une augmentation en taille et en nombre d'organes sensoriels non-visuels [5].

Une des interrogations récurrentes porte sur la connaissance du mode d'acquisition du phénotype cavernicole. En particulier, les mutations res-

ponsables du phénotype cavernicole préexistent-elles dans l'espèce épigée dont est issue l'espèce cavernicole ou au contraire sont-elles apparues après la colonisation de l'environnement cavernicole ?

Afin de répondre à la question du mode d'acquisition du phénotype cavernicole, de nombreuses études se sont employées à identifier les causes proximales [6], c'est-à-dire les mutations qui vont changer le développement, la physiologie ou le comportement. Les causes distales [6], c'est-à-dire les autres forces évolutives (sélection naturelle*, migration et dérive génétique*) sous-jacentes à ces changements, sont plus complexes à identifier et restent à ce jour très mal comprises.

Parmi les causes distales, les rôles relatifs de la dérive génétique et de la sélection sont particulièrement intéressants et certaines caractéristiques des populations permettraient de les analyser. Une de ces caractéristiques est particulièrement importante, le temps (le tempo), c'est-à-dire l'âge des populations cavernicoles et le temps nécessaire pour la mise en place du phénotype cavernicole. En effet, si les populations sont anciennes, il y a eu assez de temps pour que de nouvelles mutations spécifiques de ces populations apparaissent et se fixent*. Au contraire, si les populations sont récentes, il n'y a probablement pas eu suffisamment de temps pour l'apparition de nombreuses nouvelles mutations et encore moins pour leur fixation. Dans ce dernier cas, le phénotype cavernicole serait dû principalement à la fixation de mutations préexistantes à la colonisation des grottes c'est-à-dire présentes dans les populations de surface au moment de cette colonisation. Les populations cavernicoles étant généralement de taille réduite, la fixation de ces allèles responsables du phénotype cavernicole a lieu par dérive génétique d'allèles devenus neutres, c'est-à-dire du simple fait du hasard [7] ou par sélection d'allèles favorables dans cet environnement.

Aussi, afin de déterminer les mécanismes impliqués dans l'adaptation à l'environnement cavernicole, il est essentiel de déterminer l'âge de ces populations.

Sélection naturelle *: Avantage reproductif de certains individus par rapport à d'autres dans un environnement donné favorisant la diffusion des caractères portés par ces individus.*

Dérive génétique *: Variation aléatoire des fréquences des allèles due à l'échantillonnage de ces allèles lors de la reproduction.*

Fixation *: Un allèle* est dit fixé lorsque sa fréquence dans la population est de 1, c'est-à-dire qu'il n'existe pas d'autre allèle.*

Allèle *: Variant à un locus* donné.*

Locus *: Position sur un chromosome.*

0.1.1 *Méthodes de datation*

Les méthodes de datation se divisent en deux catégories : les méthodes de datation relative et les méthodes de datation absolue.

La datation relative permet d'ordonner des événements dans un ordre chronologique sans toutefois donner ni leur âge ni le temps séparant ces événements. En paléontologie, une des méthodes de datation relative est la biostratigraphie [8] qui se base sur l'étude des roches dans lesquelles des fossiles sont retrouvés. Ainsi, sauf dans le cas où des événements de plissements ont eu lieu, des fossiles trouvés dans une strate* inférieure seront plus vieux que ceux trouvés dans une strate supérieure. Dans le cas qui nous intéresse, aucun fossile n'est disponible.

Strate *: Couche géologique homogène*

La datation absolue permet de donner un âge, plus ou moins précis, à un événement. Différentes méthodes existent : méthodes radiométriques basées sur la diminution constante au cours du temps de la quantité d'un

Radioisotope *: Atome dont le noyau est instable par excès de neutrons ou de protons.*

radioisotope* (par exemple, la datation au carbone-14) ou la dendrochronologie [9] basée sur l'étude des cernes de croissance des arbres. En biologie évolutive, on utilise la datation moléculaire, basée sur l'étude des changements dans les séquences nucléiques ou protéiques.

La première méthode de datation moléculaire a été proposée par Zuckerkandl et Pauling en 1965 [10]. Ils ont postulé que, dans deux espèces différentes, la divergence, c'est-à-dire le nombre de différences observées, entre séquences homologues, était proportionnelle au temps écoulé depuis la séparation de ces deux espèces. C'est le concept d'horloge moléculaire.

Depuis lors, de nombreuses méthodes de datation qui reposent sur ce même principe ont été mises au point. À partir d'une distance génétique entre deux séquences homologues un temps de divergence peut être estimé si on a une estimation du taux de mutation [11]. Il faut donc d'abord calculer cette distance génétique. Ce calcul nécessite de spécifier un modèle de substitutions. Plusieurs modèles de substitutions existent, certains supposant l'équiprobabilité de l'ensemble des substitutions (par exemple le modèle JC69 [12]), d'autres supposant au contraire des probabilités de substitutions différentes en fonction du type de substitution (comme le modèle GTR [13]).

0.2 *Astyanax mexicanus* : un organisme modèle en éco-évo-dévo

Parmi les téléostéens cavernicoles, nous allons nous intéresser plus spécialement aux poissons dont au moins 90 espèces [14, 15] ont été décrites comme vivant dans l'obscurité totale. Ces espèces sont situées dans différents clades et le mode de vie cavernicole est donc apparu, chez ces poissons, plusieurs fois de façon indépendante.

Parmi ces poissons, nous nous intéresserons plus particulièrement au characidé *Astyanax mexicanus* qui est un modèle particulièrement intéressant pour les études d'éco-évo-dévo*. Chez la plupart des espèces cavernicoles, lorsque des populations épigée vivent à proximité, il s'agit généralement d'une espèce sœur. Chez *Astyanax mexicanus*, un poisson téléostéen d'eau douce, vivant au Mexique et au sud des États-Unis, les deux morphotypes, cavernicole et de surface, appartiennent à la même espèce (Figure 2). Cette présence au sein de la même espèce des deux morphotypes facilite beaucoup leur comparaison. Les deux morphotypes sont interfertiles, ce qui permet de réaliser des études de génétique.

Éco-évo-dévo *: Discipline à l'interface de la biologie évolutive, de la biologie du développement et de l'écologie.*

Différentes grottes sont connues, 30* à ce jour, pour abriter des populations de cette espèce, et situées dans la Sierra de El Abra dans la Sierra Madre Orientale (voir carte Figure 3).

Nombre de grottes *: Le nombre de grottes actuellement décrit est de 29, mais lors de notre dernière expédition au Mexique, nous avons identifié une nouvelle grotte, nommée Chiquitita.*

Il est frappant d'observer, dans une région limitée, autant de populations cavernicoles appartenant à la même espèce, alors qu'on ne connait qu'une population cavernicole de cette espèce en dehors de cette région. On peut s'interroger sur l'origine de ces différentes populations. Proviennent-elles toutes d'une unique colonisation et adaptation à l'environnement cavernicole suivie de migrations dans les différentes grottes *via* des passages

Figure 2. *Astyanax mexicanus* : morphotype de surface (à gauche), morphotype cavernicole (à droite). *(Photo : S. Rétaux)*

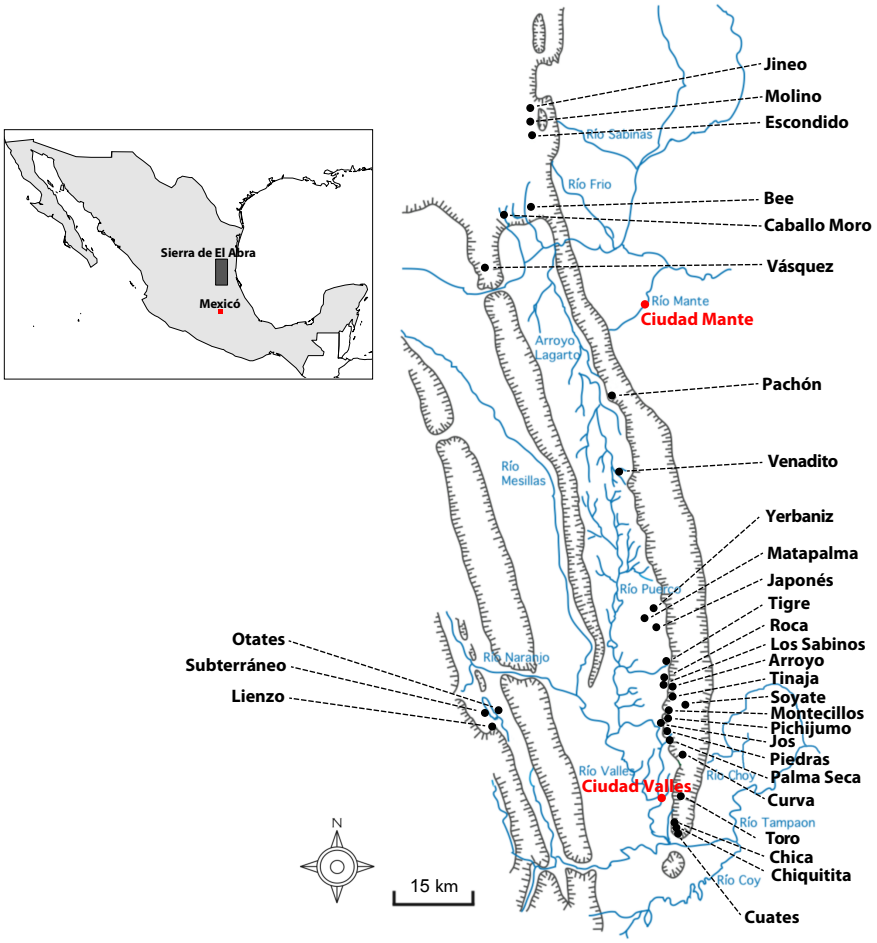


Figure 3. Carte de la Sierra de El Abra et localisation des 30 grottes identifiées pour abriter des populations d'*Astyanax mexicanus*. La situation de la Sierra de El Abra est visible sur la carte du Mexique dans l'encadré. Modifié d'après [16].

souterrains ? Ou au contraire, y-a-t'il eu différents événements de colonisation ?

L'origine des allèles responsables du phénotype cavernicole est également une question majeure chez *Astyanax mexicanus*. De nombreuses études des différences phénotypiques observables entre populations de surface et cavernicoles ont été réalisées et ont permis d'identifier les causes proximales de certaines de ces différences. En particulier, des mutations différentes ont été identifiées dans différentes populations et sont responsables d'un même phénotype, par exemple la dépigmentation. On peut se demander si ces allèles étaient présents dans les populations de surface lors de la colonisation des grottes ou s'ils sont apparus une fois les populations cavernicoles établies.

Bien qu'importante, la question de l'âge des populations cavernicoles a été très peu étudiée chez *Astyanax mexicanus*. La seule estimation, rarement citée, de l'âge des populations a été réalisée en 1974 par Chakraborty et Nei [17] comme application de leur méthode de datation à partir de données moléculaires. À partir de données de polymorphisme d'allozymes* publiées par Avise et Selander [18], ils ont estimé l'âge de la population de la grotte La Cueva de Pachón à 710 000 ans et celui de la grotte La Cueva de Los Sabinos à 525 000 ans. L'écart-type sur ces estimations (s) est néanmoins très grand : 460 000 ans pour la première et 330 000 ans pour la deuxième population. En considérant un intervalle de confiance à 95% ($\pm\ 1{,}96\ s$), ces populations peuvent donc être anciennes comme très récentes : entre 1,6 millions d'années et aujourd'hui pour Pachón et entre 1,2 millions d'années et aujourd'hui pour Los Sabinos. Ces estimations permettent seulement de conclure que ces deux populations ne sont pas âgées de plusieurs millions d'années.

***Allozyme** : Variants d'une enzyme codés par des allèles différents à un même locus.*

Pourtant, de nombreuses publications concernant *Astyanax mexicanus* mentionnent de tels temps de divergence entre population de surface et populations cavernicoles ou entre différentes populations cavernicoles. Les divergences peuvent être en fait et selon les publications, très récentes, très anciennes ou avec des âges intermédiaires : 10 000 ans [16, 19–25], des centaines de milliers d'années [26–30], entre un million d'années et 8 millions d'années [16, 19, 24, 25, 31–60].

L'objet de cette thèse est donc d'estimer avec plus de précision l'âge d'une population cavernicole. Afin de déterminer ce temps de divergence, nous utiliserons une nouvelle méthode que nous avons développée. Cette méthode est basée sur le polymorphisme identifié entre et à l'intérieur des populations, et en particulier sur les différences de fixation d'allèles dérivés.

Nous nous concentrerons sur la population de la grotte Pachón, qui est une des plus étudiées et qui est considérée comme une des plus anciennes et des plus isolées.

Dans le Chapitre 1, nous passerons en revue les différences entre morphotype de surface et morphotype cavernicole de l'espèce *Astyanax mexicanus* ainsi que les différentes hypothèses sur l'origine des populations de surface et des 30 populations cavernicoles.

Nous nous concentrerons ensuite (Chapitre 2) sur l'étude du polymorphisme observé entre une population de surface et la population de la

grotte Pachón et à l'intérieur de ces populations. Nous nous intéresserons dans un premier temps à l'étude du polymorphisme non-synonyme qui nous permettra d'évaluer le niveau de sélection pour l'ensemble des populations et pour chaque population prise séparément. Puis nous nous intéresserons au polymorphisme synonyme qui nous permettra d'émettre des hypothèses concernant l'âge de la population cavernicole.

Ces hypothèses seront testées au Chapitre 3. Les résultats concernant le temps de divergence entre la population de surface et la population cavernicole Pachón seront présentés et discutés.

Au Chapitre 4 nous utiliserons d'autres données de la littérature, des microsatellites, afin de réaliser une autre datation du temps de divergence entre population de surface et populations cavernicoles.

Enfin, nous discuterons (Chapitre 5) de ces résultats et de leur implication pour l'évolution des populations cavernicoles de l'espèce *Astyanax mexicanus*.

1 CARACTÉRISTIQUES MORPHOLOGIQUES ET ORIGINE DES POPULATIONS CAVERNICOLES DE L'ESPÈCE *ASTYANAX MEXICANUS*

1.1 Phénotype des populations cavernicoles d'*Astyanax mexicanus*

Les poissons des populations cavernicoles se distinguent de leurs homologues de surface par un grand nombre de différences morphologiques, physiologiques et comportementales.

1.1.1 *Dépigmentation*

La pigmentation est, chez les animaux, utilisée pour différentes fonctions comme le camouflage ou la protection aux rayonnements solaires. Les populations cavernicoles d'*Astyanax mexicanus* sont dépigmentées. En fonction des grottes, le niveau de dépigmentation est très variable : dans certaines grottes, comme Pachón, les poissons sont fortement dépigmentés alors que dans d'autres grottes, comme Río Subterráneo dans laquelle on peut trouver des poissons hybrides, des poissons ne le sont que très légèrement (Figure 4).

Lors d'un test de croisement de poissons provenant de différentes populations, on observe en F1 une restauration du phénotype de surface pour certains de ces croisements (Figure 5). Ce résultat montre qu'il y a complémentation fonctionnelle et donc que différents loci sont impliqués dans cette perte de pigmentation.

La mise en place de la pigmentation est un processus complexe. En effet, il existe trois types de cellules pigmentées : les mélanophores contenant de la mélanine (noir), les iridophores contenant de la guanine (argenté) et les xanthophores contenant de la ptéridine (orange). Chez les *Astyanax mexicanus* cavernicoles, les mélanophores sont en nombre réduit, voire, dans certaines populations, totalement absents, et ces mélanophores produisent moins, voire pas du tout, de mélanine. La réduction ou la suppression des mélanophores est aussi appelée phénotype *brown*.

Deux gènes ont été identifiés pour la diminution ou la perte de synthèse de mélanine chez *Astyanax mexicanus* : *oca2* et *mc1r*.

Le gène *oca2* (*oculocutaneous albinism 2*) est un transporteur du précurseur de la mélanine, la L-tyrosine, dans les mélanosomes (Figure 6). Quatre mutations ont été identifiées dans ce gène. Les deux premières, identifiées dans la population Pachón, sont des mutations ponctuelles* (SNP) entraî-

***Mutations ponctuelles** : (Single Nucléotide Polymorphism, SNP) Mutation d'un seul nucléotide.*

Figure 4. Photos de poissons cavernicoles de l'espèce *Astyanax mexicanus* provenant de différentes grottes. Modifié d'après [48].

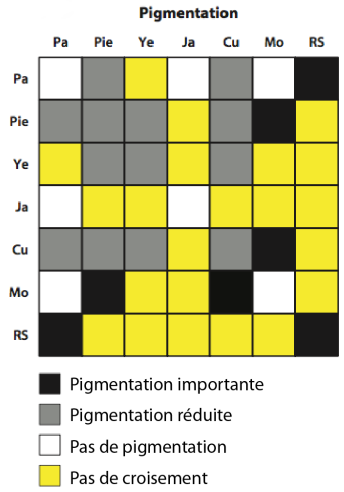


Figure 5. Échiquier de Punnet présentant les résultats du croisement de poissons de différentes grottes. En jaune les croisements qui n'ont pas marché ou qui n'ont pas été testés, en blanc les croisements produisant des poissons dépigmentés, en gris ceux produisant des poissons avec une pigmentation réduite et en noir ceux produisant des poissons ayant une pigmentation semblable aux poissons de surface. Pa : Pachón, Pie : Piedras, Ye : Yerbaniz, Ja : Japonès, Cu : Curva, Mo : Molino, RS : Río Subterráneo. Modifié d'après [48].

nant un changement d'acide aminé dans la protéine. Ces deux mutations ne semblent pas avoir d'effet sur la synthèse de mélanine [61]. Les deux autres mutations sont des délétions* qui ont pour conséquence un arrêt de la synthèse de mélanine [61]. Dans une autre population, Japonès, le gène *oca2* est également impliqué dans la perte de pigmentation, mais sa séquence codante est intacte [61].

Délétion *: Suppression d'une partie de séquence*

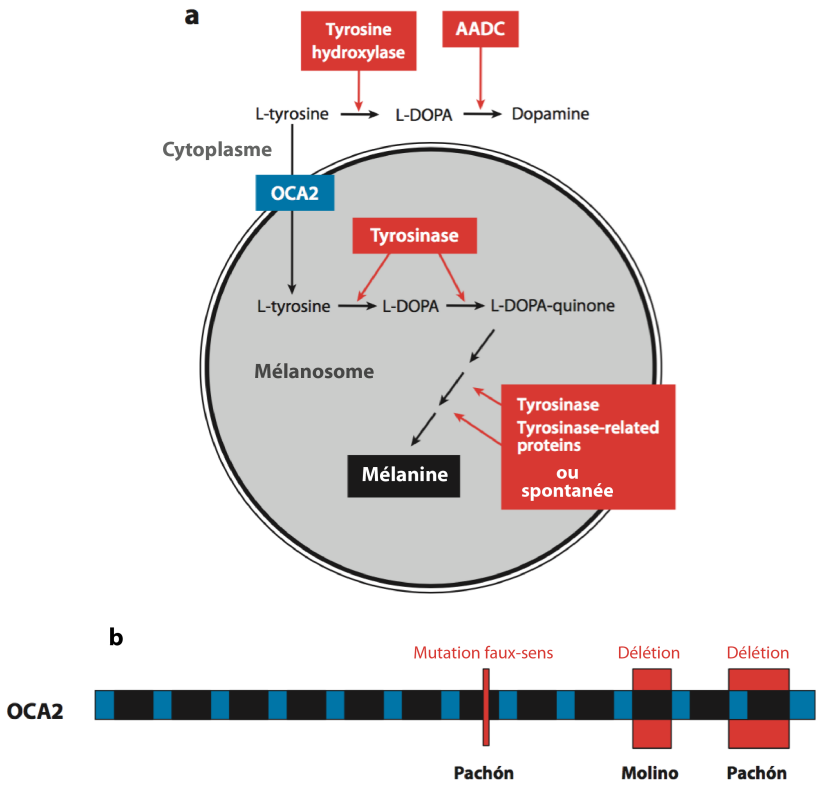


Figure 6. (a) Voie de synthèse de la mélanine. (b) Mutations observées dans la protéine codée par le gène *oca2*. Modifié d'après [48]

Le gène *mc1r* (MelanoCortin 1 Receptor, Récepteur de la mélanocortine de type 1) est un récepteur transmembranaire spécifique des mélanocytes.

Dans différentes populations, dont celle de la grotte Pachón, une délétion de deux paires de bases (2 pb) a été identifiée à l'extrémité 5' codante du transcrit, dans le domaine N-terminal de la protéine. Cette délétion a pour conséquence un codon stop prématuré et une protéine tronquée [62]. Dans d'autres grottes, une mutation ponctuelle change l'acide aminé codé d'arginine à cystéine. Chez l'Homme, une mutation ponctuelle à la même position entraînant le changement de l'arginine en tryptophane est impliqué dans le phénotype roux caractérisé notamment par une peau pâle, des cheveux roux et un risque accru de cancer de la peau [63].

1.1.2 *Perte des yeux*

Comme chez de nombreuses espèces cavernicoles, les populations cavernicoles d'*Astyanax mexicanus* ont perdu leurs yeux. Plus précisément, si les yeux se développent normalement dans les premières heures du développement chez les poissons cavernicoles (Figure 7), leur dégénérescence commence par une entrée en apoptose* du cristallin environ 24h après la fécondation [64]. La dégénérescence du cristallin entraîne par la suite une dégénérescence de l'ensemble de l'oeil. Le cristallin semble être le seul responsable de cette dégénérescence. En effet, une opération de transplantation d'un cristallin de poisson de surface dans la cupule optique (voir Figure 7) d'un embryon de poisson cavernicole a permis une restauration de l'oeil [53]. L'opération inverse, transplantation d'un cristallin de poisson cavernicole chez un poisson de surface, a entrainé la perte de l'oeil [53].

Apoptose *: Mort cellulaire programmée.*

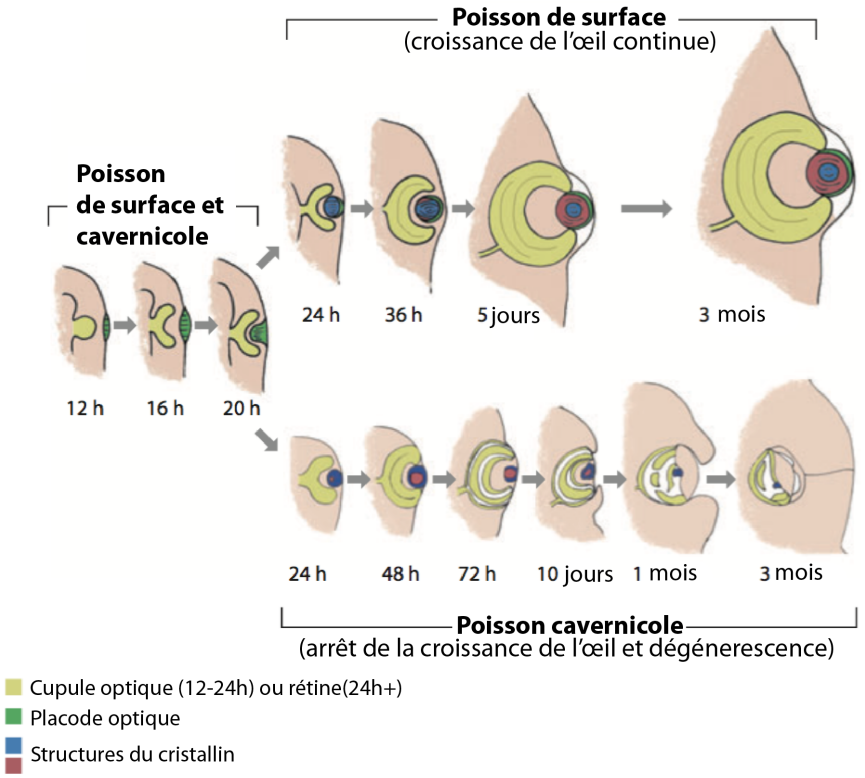


Figure 7. Schéma du développement de l'oeil chez *Astyanax mexicanus*. L'oeil se développe chez les deux morphotypes, mais à 48 heures après fécondation, le cristallin entre en apoptose chez les poissons cavernicoles. Modifié d'après [48].

Hétérotopie *: Modification du territoire d'expression d'un gène.*

Hétérochronie *: Modification du moment d'expression d'un gène.*

Si les yeux se développent dans les premières heures après fécondation chez les poissons cavernicoles, la taille de ces yeux est plus petite que chez les poissons de surface, en particulier le quadrant ventral de la rétine [65]. Une hétérotopie* de l'expression du gène *shh* entraîne une hétérochronie* de l'expression d'un gène (*fgf8*) qui entraîne, elle-même, une hétérotopie

de l'expression du gène *pax6*. Cette hétérotopie entraîne une diminution du territoire disponible pour la formation de l'oeil et serait à l'origine de la dégénérescence de l'œil.

1.1.3 *Recherche alimentaire et métabolisme*

Les poissons cavernicoles vivent dans un environnement où la nourriture est peu abondante et irrégulièrement disponible. De plus, dans l'obscurité il est difficile de la trouver. Dans cet environnement, les poissons cavernicoles ont des capacités à trouver de la nourriture supérieures à celles des poissons de surface [66, 67]. Les poissons cavernicoles ont également un métabolisme modifié.

1.1.3.1 *Stockage des graisses*

Les poissons cavernicoles stockent plus de graisses que les poissons de surface [68], en particulier les triglycérides [69]. Ce stockage de graisses plus important permet aux poissons cavernicoles de mieux résister à l'absence de nourriture dans l'environnement. Ainsi, après un mois de jeûne, il n'y a pas d'augmentation de l'appétit chez les poissons cavernicoles des populations Pachón et Tinaja, contrairement aux poissons de surface [69].

Comme chez de nombreuses espèces cavernicoles, les œufs de poissons cavernicoles sont moins nombreux mais sont plus gros et possèdent plus de vitellus (environ 50% de plus) que ceux des poissons de surface [29], permettant aux embryons de vivre plus longtemps sur leurs propres réserves et donc de moins dépendre des ressources alimentaires disponibles dans l'environnement.

1.1.3.2 *Économie d'énergie*

Chez les vertébrés, la rétine consomme plus d'énergie dans le noir qu'à la lumière [70]. Ainsi, une des raisons permettant d'expliquer la perte de l'œil serait non pas son inutilité mais une réduction de l'énergie consommée par les poissons cavernicoles. L'économie d'énergie apportée par l'absence des yeux est comprise entre 5% chez les adultes et 15% chez les juvéniles [71].

L'économie d'énergie la plus importante est apportée par la perte du rythme circadien dans le métabolisme [72] grâce à laquelle les poissons cavernicoles dépensent 27% d'énergie de moins que les poissons de surface le jour et 38% de moins dans l'obscurité.

La perte des yeux et du rythme circadien dans le métabolisme permet donc de substantielles économies d'énergie, en particulier chez les juvéniles. Ces deux caractères, perte des yeux et perte du rythme circadien, pourraient avoir été sélectionnés directement car les poissons porteurs de ces caractères pourraient avoir été avantagés dans un environnement aux ressources limitées.

1.1.4 *Chimiosensibilité*

La chimiosensibilité est un des plus anciens systèmes sensoriels et serait apparue il y a 500 millions d'années [73, 74]. Elle est impliquée dans deux fonctions primordiales à la survie individuelle et de l'espèce : l'alimentation et la reproduction [73].

Chez les poissons, la chimiosensibilité regroupe l'olfaction et la gustation.

1.1.4.1 *Olfaction*

Une étude a montré que les poissons cavernicoles ont de meilleures capacités olfactives que les poissons de surface [75]. Un test d'olfaction *in situ* a été réalisé en mars 2013 dans la grotte Río Subterráneo. Dans cette grotte, peu isolée de la surface, il est possible d'observer des poissons cavernicoles et des poissons de surface, ainsi que des poissons présentant des phénotypes intermédiaires.

Lors de cette expérience, une dizaine de poissons ont été placés dans une piscine gonflable pendant 24h pour habituation. À l'aide de seringues et de tubulures de perfusion, de l'eau et un extrait alimentaire sont injectés de part et d'autre de la piscine. Il est alors possible de voir de quel côté de la piscine se trouvent les poissons ainsi que que le temps passé à proximité de l'arrivée des tubulures (Figure 8).

Les poissons dont le phénotype est plus proche de celui de surface n'étaient pas attirés par l'odeur contrairement aux poissons avec un phénotype cavernicole (Figure 8) [76].

De nouvelles études ont été réalisées au laboratoire et *in situ* dans différentes grottes en utilisant des acides aminés au lieu d'extrait alimentaire (Annexe A). Ces tests suggèrent que les poissons de la grotte Pachón sont capables de détecter l'odeur de l'alanine, un acide aminé pouvant être utilisé comme *proxy* de l'odeur de nourriture, jusqu'à des concentrations de 10^{-10} M alors que les poissons de surface sont incapables de détecter ce même acide aminé à une concentration inférieure à 10^{-5} M [77]. Les poissons cavernicoles auraient donc un odorat sensible à des concentrations 100 000 fois plus faibles que les poissons de surface.

Cette augmentation des capacités olfactives pourrait être due, en partie, à une augmentation de la taille de l'épithélium olfactif chez les poissons cavernicoles : dans les embryons de poissons cavernicoles, la placode olfactive, tissu précurseur de l'épithélium olfactif, est environ 1,5 fois grande que chez les embryons de poissons de surface [77].

1.1.4.2 *Gustation*

La gustation dépend de bourgeons gustatifs. Chez les poissons, les bourgeons gustatifs sont situés sur les ouïes, les barbillons, les nageoires, dans la cavité bucale et dans le pharynx, mais pas sur la langue [73]. Les poissons cavernicoles possèdent plus de bourgeons gustatifs que les poissons de surface [78, 79].

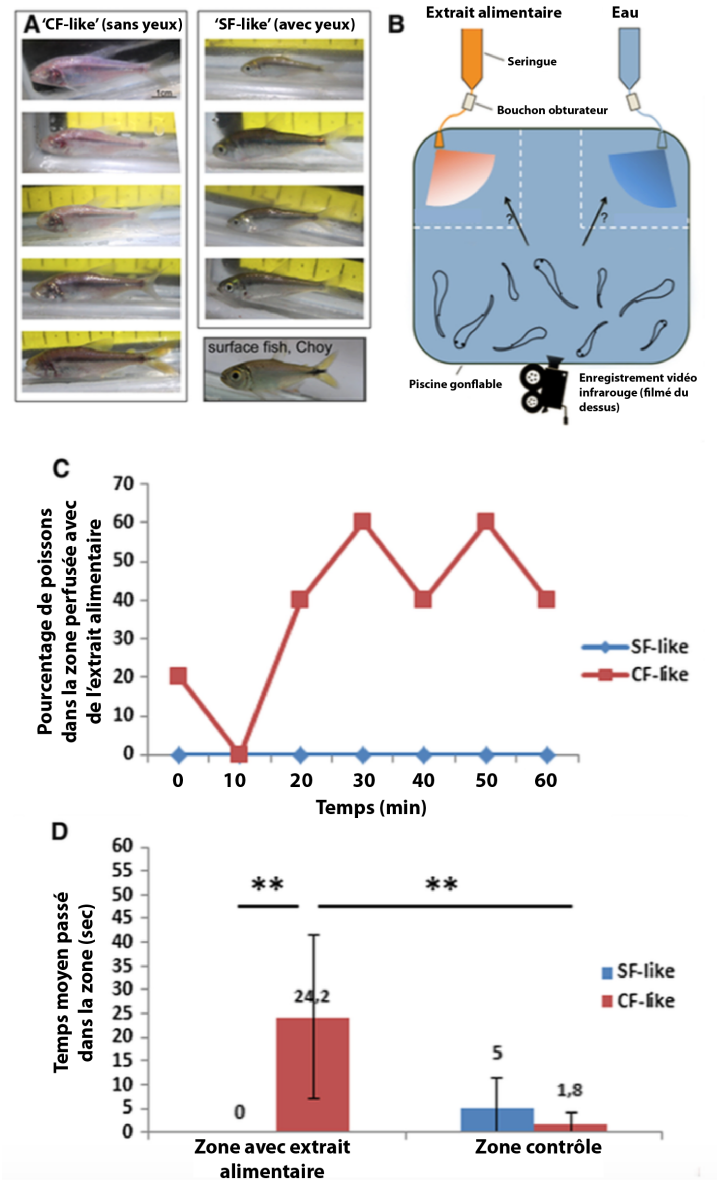


Figure 8. Test de comportement en réponse à un extrait alimentaire chez des poissons de la grotte Río Subterráneo. (A) Photos des poissons testés. Les poissons utilisés étaient de phénotype cavernicole (CF-like) ou de surface (SF-like). (B) Schéma du dispositif expérimental. (C) Pourcentage de poissons présents dans la zone perfusée en extrait alimentaire en fonction du temps pour les deux morphotypes. (D) Temps moyen passé dans la zone contrôle et dans la zone perfusée en extrait alimentaire pour les deux morphotypes. Modifié d'après [76].

1.1.5 *Augmentation du nombre de neuromastes et modification morphologique*

Les poissons cavernicoles ne pouvant se repérer grâce à la vision ont surdéveloppé une autre méthode d'orientation. Chez de nombreux vertébrés aquatiques le système mécano-sensoriel de la ligne latérale est utilisé pour

détecter les changements de pression de l'eau, ce qui permet d'identifier le sens du courant, les proies, etc. La ligne latérale est composée de neuromastes. Les neuromastes sont un ensemble de cellules ciliées entourées de cellules supports qui vont sécréter la cupule, sorte de capuchon gélatineux protégeant les cils (Figure 9). Cette cupule est déplacée par les mouvements d'eau, et ce déplacement entraîne une déformation des cils. Cette déformation des cils résulte dans un changement du potentiel électrique de la cellule qui est ensuite transmis aux fibres afférentes. Chaque cellule ciliée possède un axe de sensibilité maximale. Le déplacement des cils sur cet axe sera excitateur dans un sens et inhibiteur dans le sens inverse [80].

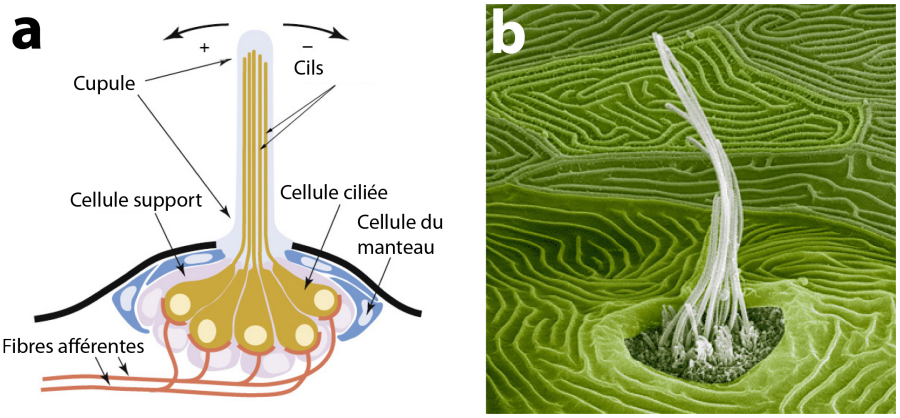


Figure 9. (a) Schéma d'un neuromaste. Modifié d'après [81] (b) photo en microscopie électronique à balayage d'un neuromaste de larve de poisson zèbre (*Danio rerio*). (Photo Jürgen Berger, couverture Neuron février 1998 20(2)).

Les neuromastes sont de deux types. Les neuromastes superficiels situés au niveau de l'épiderme qui sont sensibles à des fréquences faibles de vibration (< 50 Hz) et qui permettent principalement la détection du sens du courant. Les neuromastes canaux sont, eux, enfermés dans une gouttière épidermique (Figure 10). Ils sont sensibles à des fréquences plus élevées et sont impliqués dans la détection de proies.

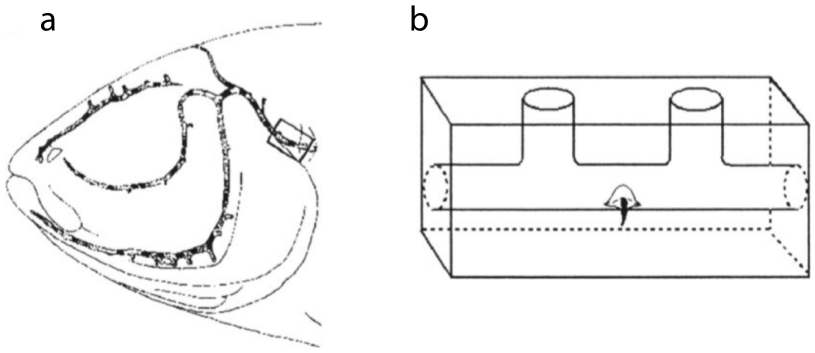


Figure 10. (a) Localisation des canaux de la ligne latérale au niveau de la tête chez le poisson cavernicole *Astyanax mexicanus*. (b) Schéma d'une section de canal avec un neuromaste. Modifié d'après [80].

Bien que le nombre de neuromastes superficiels soit similaire chez les deux morphotypes, les poissons cavernicoles en ont près de deux fois plus que les poissons de surface au niveau de la tête (Figure 11 A-B) et de la nageoire caudale. Les poissons de surface possèdent, eux, plus de deux fois plus de neuromastes superficiels au niveau du tronc [78, 82]. Les neuromastes canaux sont présents en même quantité et sont répartis de façon similaire dans les deux morphotypes (Figure 11 C-D) [78, 82].

Chez les poissons cavernicoles, la structure des neuromastes superficiels est modifiée. La cupule (la partie ciliée) est bien plus longue : plus de 100 µm en moyenne avec un maximum de plus de 300 µm chez les poissons cavernicoles contre une moyenne de 40 µm (et un maximum de 50 µm) chez les poissons de surface. La base de la cupule est également plus large chez les poissons cavernicoles. Cette augmentation de la longueur des cupules permet une plus grande sensibilité aux changements hydrodynamiques [83]. Il semble qu'il n'y ait pas de modification de structure des neuromastes canaux chez le morphotype cavernicole.

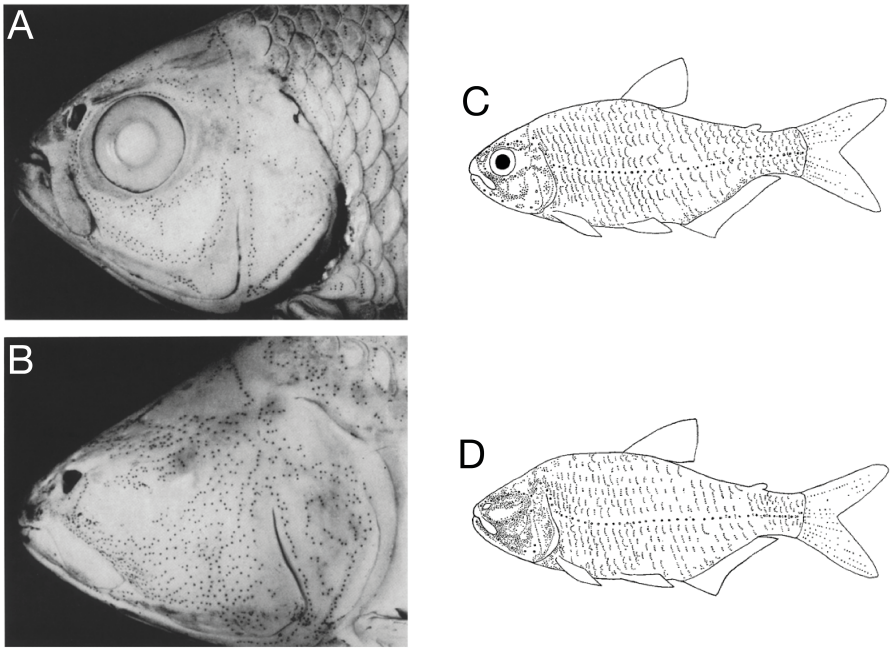


Figure 11. (A et B) Photos noir et blanc du côté gauche de la tête d'un poisson de surface (A) et d'un poisson cavernicole (B) après coloration des neuromastes au bleu de méthylène. (C et D) Schéma de répartition des neuromastes chez un poisson de surface (C) et un poisson cavernicole (D). Les points fins représentent les neuromastes superficiels alors que les gros points au niveau de la ligne latérale représentent les neuromastes des canaux épidermiques. Modifié d'après [78].

1.1.6 *Comportement d'attraction aux vibrations (VAB)*

Les poissons cavernicoles sont très sensibles aux modifications de pressions de l'eau. Cette sensibilité a permis l'apparition d'un nouveau comportement d'attraction aux vibrations (VAB : Vibration Attraction Behavior).

Pour caractériser ce comportement, des poissons cavernicoles et des poissons de surface ont été placés dans le noir dans un aquarium dans lequel se trouvait un bâtonnet vibrant (Figure 12) [84]. Il a ainsi pu être observé que les poissons cavernicoles s'approchaient plus souvent du bâtonnet vibrant, pendant plus de temps et après une période de latence plus courte que les poissons de surface (Figure 12). En faisant varier la fréquence des vibrations il a pu être déterminé que les poissons cavernicoles détectaient les vibrations comprises entre 10 Hz et 50 Hz, avec un pic à 35 Hz (Figure 13). Ce pic correspond à peu près à la fréquence émise par les mouvements d'insectes piégés dans l'eau (Figure 13) [85], ce qui ferait penser à une adaptation permettant de trouver de la nourriture plus facilement bien que le régime alimentaire des poissons cavernicoles ne soit pas bien connu pour le moment.

Locus de caractère quantitatif *(Quantitative Trait Loci, QTL) Portion du génome présentant un polymorphisme corrélé à la variation d'un caractère quantitatif.*

Une recherche de QTL* a permis de montrer que le nombre de neuromastes suborbitaux et le comportement d'attraction aux vibrations à 35Hz était génétiquement liés [44]. De façon intéressante, ces deux caractères sont également génétiquement liés à la taille des yeux, ce qui pourrait laisser penser à une sélection indirecte de la disparition des yeux des poissons cavernicoles. Les zones cartographiées sont très grandes et cette hypothèse est très discutée [40, 44].

Les poissons cavernicoles sont également capables de se repérer grâce à une augmentation de la fréquence du mouvement d'ouverture et de fermeture de leur bouche lorsqu'ils s'approchent de la paroi à des niveaux bien plus élevés qu'observé chez d'autres espèces de poissons [86]. Cette augmentation de fréquence permet de détecter les obstacles non mouvants par les neuromastes. Ce mécanisme est un ordre de grandeur plus efficace que la détection avec les changements de pression dus aux mouvements du poisson pour des vitesses faibles et à proximité des obstacles. De plus, le mouvement du poisson génère plutôt des variations du courant faible alors que l'ouverture/fermeture de la bouche génère plutôt des oscillations. Or, les neuromastes d'*Astyanax mexicanus* sont plus sensibles aux oscillations qu'aux faibles variations du courant [86].

1.1.7 *Prise alimentaire*

Les poissons cavernicoles sont, contrairement aux poissons de surface, capables de nager tout en se nourrissant grâce à une modification de leur position lors de la prise alimentaire. Ainsi les poissons de surface cherchent la nourriture en pleine eau alors que les poissons cavernicoles la cherche plutôt au fond. Au sol, les poissons de surface cherchent de la nourriture en formant un angle avec le sol d'environ 90° alors que les poissons cavernicoles

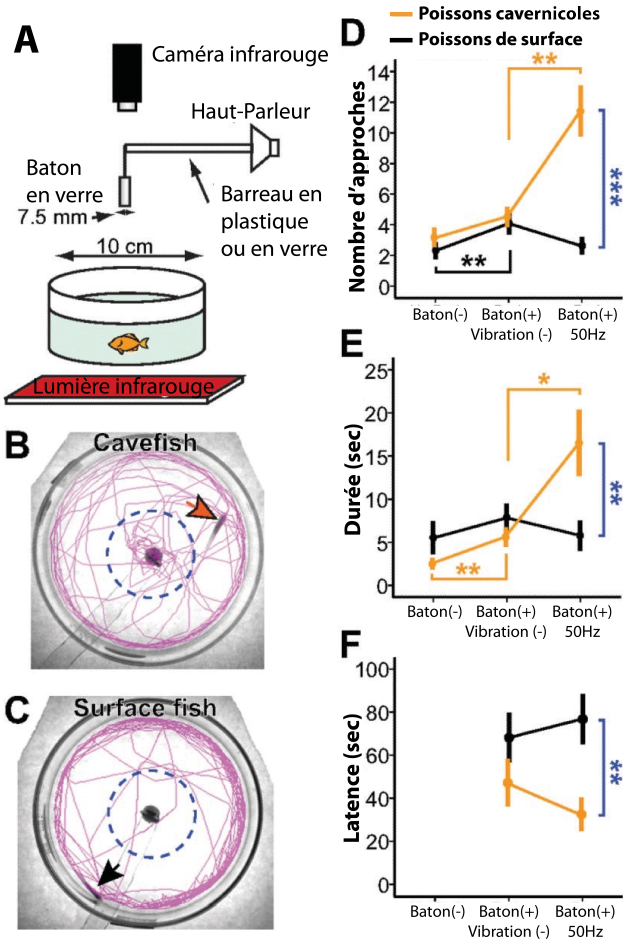


Figure 12. (A) Dispositif expérimental permettant d'enregistrer le VAB dans le noir : un aquarium est placé sur une source lumineuse infrarouge. Un bâtonnet en verre relié à un haut-parleur est placé dans l'aquarium. La nage d'un poisson est enregistrée grâce à la caméra CCD infrarouge. (B et C) Chemin parcouru par les poissons cavernicoles (B) et de surface (C) pendant 3 minutes. La ligne en pointillé représente une zone de 2 cm autour du bâtonnet en verre. La flèche indique la position du poisson au début de l'expérience. (D à F) Quantification du comportement d'approche chez les poissons cavernicoles (en noir) et les poissons de surface (en orange) : nombre d'approches en fonction de la fréquence de vibration (D), durée d'approche (E) et latence (F). Image d'après [84].

adoptent un angle d'environ 45° [87] (Figure 14). Comparés aux poissons de surface, les poissons cavernicoles ont une bouche plus large et en forme de pelle. Cette différence leur permet de « scanner » plus efficacement le fond des bassins afin de trouver de la nourriture [88].

1.1.8 *Comportement social et agressivité*

Les poissons de surface nagent en bancs et une hiérarchie entre individus se met en place au sein des populations. Ces comportements ne sont pas re-

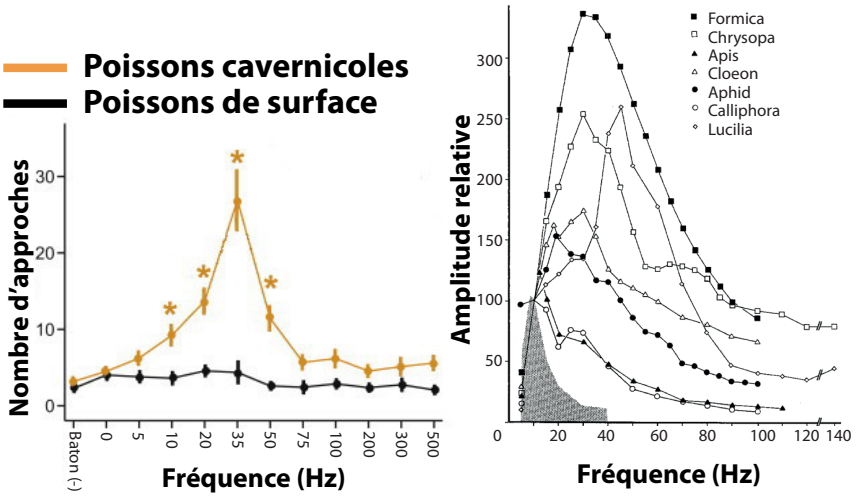


Figure 13. *(À gauche)* Nombre d'approches du bâtonnet vibrant en fonction de sa fréquence de vibration pour les poissons de surface (en noir) et les poissons cavernicoles (en orange). Modifié d'après [84]. *(À droite)* Spectres des ondes produites par des insectes piégés dans l'eau pour différents genres. Modifié d'après [85].

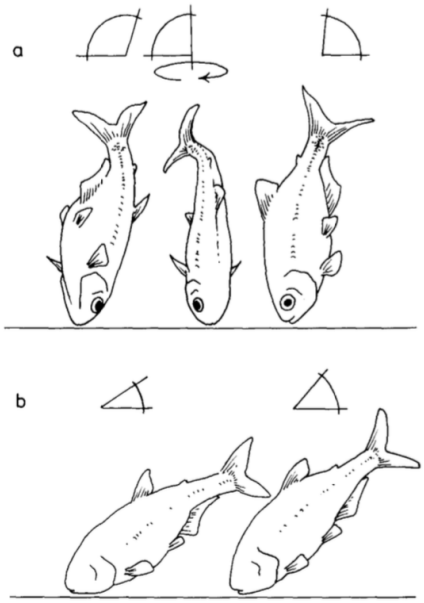


Figure 14. Posture de prise alimentaire. Les poissons de surface (a) adoptent un angle d'environ 90° par rapport au sol, alors que les poissons cavernicoles (b) adoptent un angle d'environ 45°. Image d'après [87].

trouvés chez les populations cavernicoles [60, 89]. Les poissons cavernicoles sont également moins agressifs que les poissons de surface [60]. L'agressivité des poissons cavernicoles serait liée à une recherche alimentaire alors que chez les poissons de surface elle serait plutôt liée à la mise en place d'une hiérarchie sociale [60].

1.2 Origines des populations cavernicoles d'*Astyanax mexicanus*

1.2.1 *Géologie de la Sierra de El Abra*

La Sierra de El Abra, où l'on trouve les populations cavernicoles d'*Astyanax mexicanus*, constitue la partie la plus orientale de la Sierra Madre Orientale. Elle s'étend sur 150 km du Nord-Ouest au Sud-Est et de 7 km à 15 km d'Est en Ouest à la limite entre l'état du Tamaulipas et de San Luis Potosí. Elle s'élève à l'Est abruptement entre 250 m et 300 m au dessus du niveau de la pleine côtière et à l'Ouest à 150 m au dessus de la vallée [90].

Les roches calcaires de la Sierra de El Abra sont issues d'un dépôt marin qui a eu lieu entre 100 et 98 millions d'années au milieu du Crétacé (Figure 15) [18]. Il y a 80 à 35 millions d'années a eu lieu l'orogenèse* laramienne, *i.e.* le passage de la plaque Farallon, aujourd'hui disparue, sous la plaque nord-américaine. Cet événement géologique majeur est responsable de la formation de nombreuses chaînes montagneuses d'Amérique du Nord, comme les Montagnes Rocheuses ou la Sierra Madre Orientale où se trouve la Sierra de El Abra. C'est cet événement qui a soulevé les roches calcaires déposées lors du Crétacé dans la Sierra de El Abra et les a exposées [16].

Orogenèse *: Formation de montagne.*

Les premières cavités de la région ont été formées par un processus de spéléogenèse* par corrosion sulfurique [91] : du sulfure d'hydrogène, provenant de pétrole, s'échappe et en remontant dans les couches supérieures est mélangé avec de l'eau. Ce mélange provoque la formation d'acide sulfurique qui va alors dissoudre le calcaire.

Spéléogenèse *: Formation de cavités.*

La géochimie de la région a ensuite changé et la spéléogenèse s'est alors faite de manière plus conventionnelle par dissolution du calcaire causée par de l'eau acidifée avec du CO_2 [91].

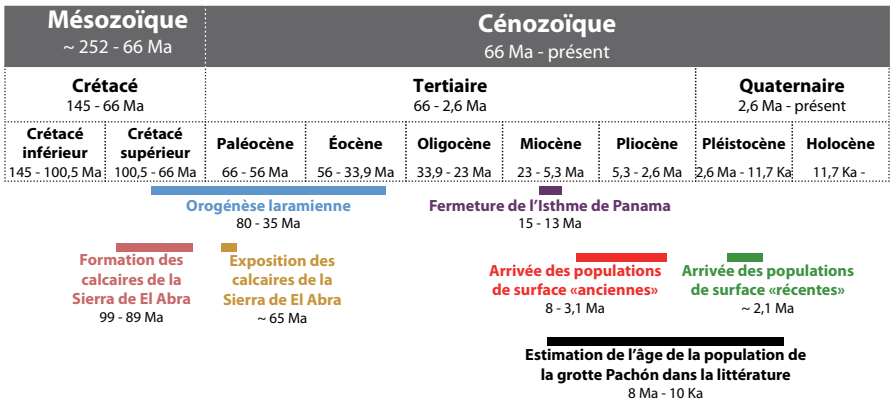


Figure 15. Évènements clés dans l'évolution des populations cavernicoles d'*Astyanax mexicanus*. Modifié d'après [16].

1.2.1.1 *Fermeture de l'Isthme de Panama*

L'Isthme de Panama est un étroit morceau de terre reliant l'Amérique du Nord à l'Amérique du Sud. Sa fermeture il y a 15 à 13 millions d'années (Fi-

gure 15) [92, 93] est un des événements géologiques les plus importants et a entraîné des changements climatiques, océaniques et biologiques majeurs sur l'ensemble du globe. Par exemple, les courants marins entre l'océan Atlantique et l'océan Pacifique ont été arrêtés et ont été détournés entraînant une forte intensification du Gulf Stream [94, 95].

Les populations marines de l'océan Pacifique et de l'océan Atlantique ont été isolées par cette nouvelle barrière géographique et ont ainsi évolué distinctement (spéciation par vicariance) [96]. Les espèces terrestres et dulçaquicoles, ont au contraire, pu migrer entre l'Amérique du Sud et l'Amérique du Nord. Ces migrations sont appelées le Grand Échange Interaméricain et se sont produites il y a environ 3 millions d'années.

L'ordre des Characiformes est retrouvé sur le continent africain et dans l'écozone* Néotropicale (Figure 16) [98]. Elle a une origine gondwanienne (avant la séparation de l'Afrique et de l'Amérique du Sud) [98]. La plus grande diversité d'espèces de cette famille est retrouvée en Amérique du Sud dans le bassin amazonien. Parmi les Characiformes, seul le genre *Astyanax* est retrouvé dans l'écozone Néarctique (Figure 16) [99]. Sa répartition géographique va de la Patagonie, au sud de l'Argentine, jusqu'au Sud des États-Unis, au Texas. Le genre *Astyanax* serait ainsi originaire d'Amérique du Sud et aurait pu coloniser l'Amérique du Nord suite à la fermeture de l'Isthme de Panama.

Écozone *: Une partie du globe représentant une unité écologique [97].*

Cordillère Néovolcanique *(*Trans Mexican Volcanic Belt, *TMVB) : ceinture volcanique traversant le Mexique d'Est en Ouest à la latitude de Mexico, au sud des Sierras Madre Occidentale et Orientale. Elle est considérée comme la séparation géologique entre Amérique du Nord et Amérique Centrale.*

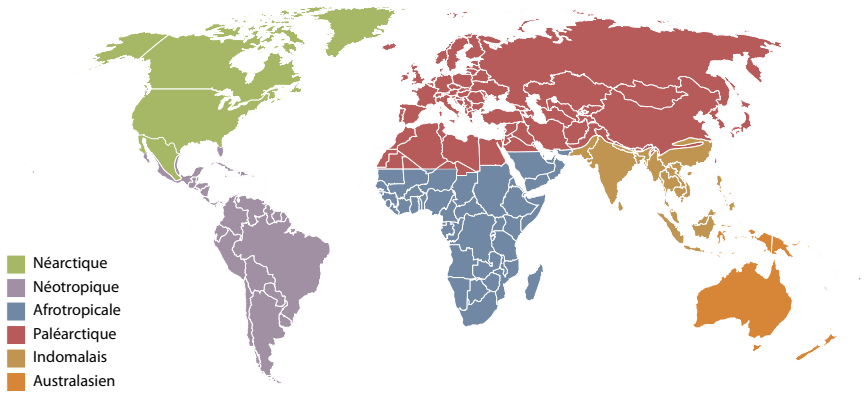


Figure 16. Planisphère présentant six des huit écozones. Les écozones Arctique et Antarctique ne sont pas représentées sur la carte. Modifié d'après https://commons.wikimedia.org/wiki/User:CarolSpears

1.2.2 *Populations d'*Astyanax *dans la Sierra de El Abra*

Après le passage de l'Isthme de Panama, le genre *Astyanax* a pu s'étendre rapidement dans toute l'Amérique Centrale il y a huit à trois millions d'années (Figure 15). Après avoir passé la Cordillère Néovolcanique* les *Astyanax* ont pu atteindre le nord du Mexique et le sud des États-Unis il y a six à trois millions d'années.

En 1936, une population cavernicole de l'espèce *Astyanax mexicanus* est découverte [100] dans la Sierra de El Abra. Cette population, provenant

de la grotte La Cueva* Chica (voir carte Figure 3), est, dans un premier temps, attribuée à une nouvelle espèce appartenant à un nouveau genre, *Anoptichtys jordani*, bien que les poissons de cette grotte ressemblaient fortement à ceux de l'espèce *Astyanax mexicanus* présente en surface dans les rivières proches.

La découverte d'autres populations cavernicoles a également fait l'objet de création de nouvelles espèces : *Anoptichtys antrobius* pour la population de la grotte Pachón [101] et *Anoptichtys hubbsi* pour Los Sabinos [102].

Depuis de nombreuses autres populations, 30 à ce jour (voir carte Figure 3), ont été découvertes dans la région de la Sierra de El Abra, et ont fait l'objet de nombreuses publications. Par la suite, les populations de surface et cavernicoles ont été classées dans l'espèce *Astyanax mexicanus* ou *Astyanax fasciatus*. Avise et Selander [18] notent qu'il n'y a aucune raison de classer les populations cavernicoles et les populations de surface dans des genres et des espèces différents. En effet, il est possible, au laboratoire, de croiser des poissons de surface et cavernicoles et d'obtenir une descendance fertile. Dans différentes grottes, comme Chica ou Río Subterráneo, des poissons ayant un phénotype intermédiaire issus du croisement d'un poisson de surface avec un poisson cavernicole sont observés.

__Cueva__ : Terme désignant une grotte dont l'entrée est horizontale et peut se faire en marchant, contrairement aux __Sotanos__ dont l'entrée est verticale.

1.2.3 *Nombre d'événements de colonisation de l'environnement cavernicole*

L'observation de différentes grottes avec des populations cavernicoles appartenant à la même espèce interroge sur l'origine de ces populations et a nourri de nombreuses discussions sur le nombre de colonisations de l'environnement cavernicole par cette espèce. Les différentes populations pourraient être issues d'un unique événement de colonisation de l'environnement cavernicole suivi de migrations. Il est également possible que chaque population soit indépendante et par conséquent que l'acquisition du phénotype cavernicole ait eu lieu à 30 reprises. Enfin, il y a pu y avoir différentes colonisations indépendantes suivies de migrations. L'acquisition du phénotype cavernicole aurait ainsi eu lieu de façon indépendante à quelques reprises.

Actuellement, la plupart des publications font mention de deux lignées cavernicoles distinctes. Cette séparation est basée sur l'observation, en 2002, de deux groupes d'haplotypes (haplogroupes) du gène mitochondrial ND2 [103], nommés A et B. Ces deux haplogroupes divergent d'environ 3,5% [103]. Les populations des grottes Chica, Pachón et Río Subterráneo (Figure 3) appartiennent à l'haplogroupe A alors que les populations des grottes Los Sabinos, Tinaja et Curva (Figure 3) appartiennent à l'haplogroupe B. Les populations de surface de la Sierre de El Abra appartiennent, quant à elles, à l'haplogroupe A [103]. Les auteurs suggèrent que les populations cavernicoles appartenant à l'haplogroupe B ont colonisé l'environnement cavernicole à partir de populations de surface de l'haplogroupe B. Ces populations de surface auraient disparu et auraient été remplacées par des populations de surface appartenant à l'haplogroupe A (Figure 18). Les

populations cavernicoles appartenant à cet haplogroupe seraient dérivées de cette « nouvelle » population de surface [103].

En 2003, une étude est réalisée sur un autre gène mitochondrial, le cytochrome b (*cytb*). Les mêmes groupes ont été retrouvés [104]. Dans cette étude, des séquences microsatellites ont également été étudiées. Les résultats obtenus avec ces séquences diffèrent de ceux obtenus avec les gènes mitochondriaux. La population de la grotte Pachón est, avec les gènes mitochondriaux, regroupée avec les populations des grottes Chica et Río Subterráneo (aussi appelée Micos), alors qu'en utilisant des marqueurs nucléaires (les séquences microsatellites), elle est regroupée avec les populations des grottes Los Sabinos et Tinaja (Figure 17).

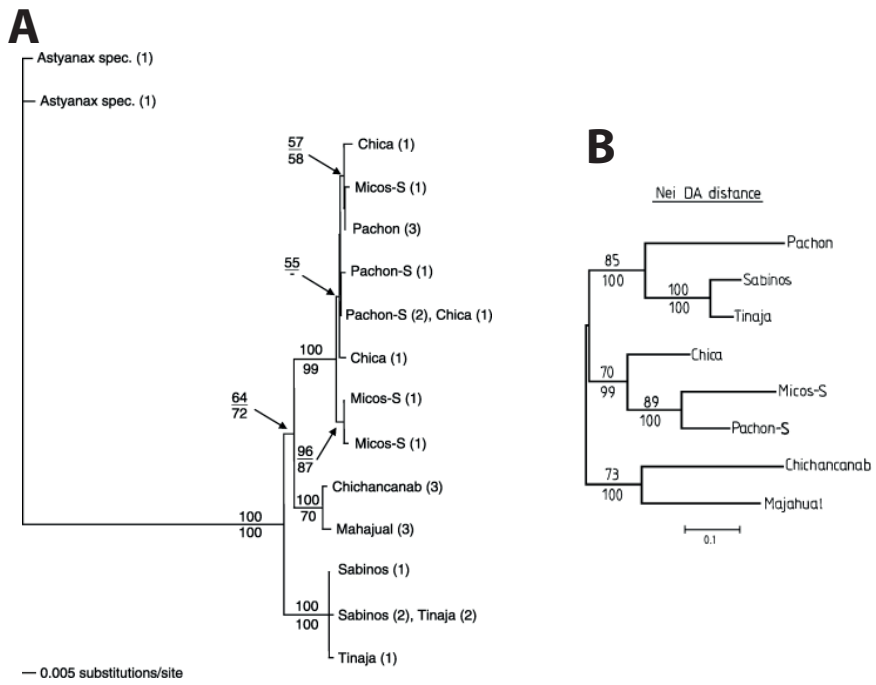


Figure 17. (A) Arbre basé sur les séquences de *cytb*. (B) Arbre basé sur les séquences microsatellites. Les populations dont le nom se termine par -S sont les populations de surface échantillonnées à proximité des grottes. Modifié d'après [104]

Il y a donc une contradiction entre les marqueurs nucléaires et les marqueurs mitochondriaux pour la population de la grotte Pachón. Pour expliquer cette contradiction, il a été proposé que la population Pachón soit une population ancienne ayant, initialement, un génome mitochondrial appartenant à l'haplogroupe B. Un événement d'introgression plus récent aurait permis le remplacement total ou partiel du génome mitochondrial par un génome de l'haplogroupe A (Figure 18) [104].

Dans une étude de ces mêmes auteurs en 2004 [105], des poissons cavernicoles et de surface ont été échantillonnés dans tout le Mexique (Figure 19) et leur cytochrome b séquencé. Sept haplogroupes (nommés de A à G, le groupe G correspondant au groupe B des études précédentes. . .) différents

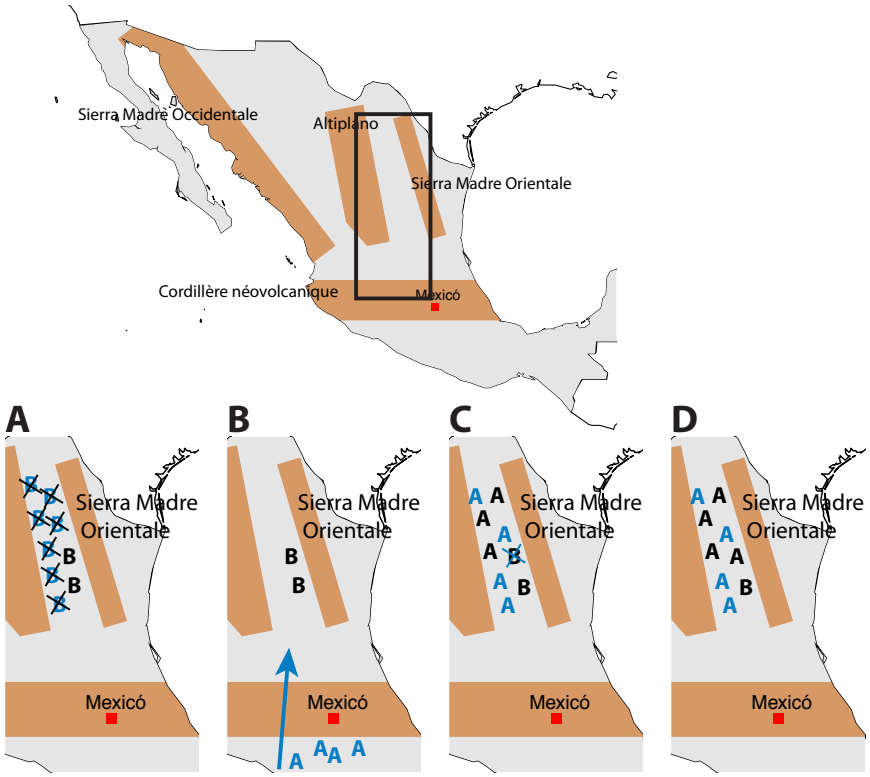


Figure 18. Schéma du scénario évolutif tel que décrit dans la littérature. (A) Suite au refroidissement de la région, les populations de surface de la Sierra de El Abra de l'haplotype B auraient disparu mais (B) les populations cavernicoles auraient survécu. (B-C) À la fin du dernier âge glaciaire, des populations de surface de l'haplotype A aurait recolonisé la région. (C-D) Puis, dans quelques grottes, l'ADN mitochondrial, de l'haplotype B, aurait été remplacé par de l'ADN mitochondrial de l'haplotype A.

ont été identifiés. Dans la Sierra de El Abra, seuls les haplogroupes A et G sont retrouvés dans les populations cavernicoles [105].

En 2008, Ornelas-Garcia *et al.* ont étudié la phylogénie du genre *Astyanax* en étudiant plusieurs gènes mitochondriaux (*cytb*, 16s et CO1) et le gène nucléaire RAG1 [52]. Les populations cavernicoles d'*Astyanax mexicanus* ont été identifiées dans deux groupes distincts : le groupe Ia, correspondant au groupe A précédemment étudié et le groupe II correspondant au groupe G (ou B) des études précédentes. Des populations de surface sont également identifiées dans ces deux groupes. En utilisant une horloge moléculaire, le temps de divergence entre ces deux groupes a été estimé à 6,7 millions d'années [52]. Les auteurs de ce papier ont considéré ces deux groupes comme faisant partie de deux espèces différentes, *Astyanax mexicanus* pour le premier et *Astyanax hubbsi* pour le deuxième. Pourtant des poissons appartenant à des populations possédant des haplotypes différents sont interfertiles (Figure 5). Par exemple, il est possible de croiser un poisson de

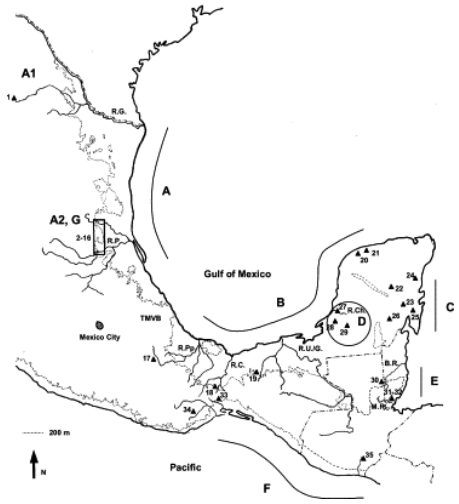


Figure 19. Carte des points d'échantillonnage [105] et des haplotypes identifiés. Les numéros permettent d'identifier les points d'échantillonnage et les lettres A à G les haplogroupes identifiés.

la population Pachón (groupe Ia) avec un individu de la population Piedras (groupe II).

1.2.4 *Divergence des haplotypes*

L'existence de plusieurs lignées mitochondriales serait la conséquence de plusieurs vagues de migration d'*Astyanax* d'Amérique du Sud vers l'Amérique du Nord (Figure 15) [16, 52]. En utilisant le concept d'horloge moléculaire, le temps de divergence entre les haplogroupes Ia (A) et II (B/G) a été estimé entre 1,8 millions d'années [105] et 6,7 millions d'années [52].

Ces temps de divergence ne donnent aucune information sur l'âge des populations cavernicoles puisque ces deux haplogroupes sont présents à la fois dans les populations de surface et dans les populations cavernicoles et les deux haplogroupes sont retrouvés en sympatrie dans au moins une population de surface.

Dans le chapitre suivant, nous allons nous intéresser à caractériser le polymorphisme existant entre une population de surface et la population de la grotte Pachón et à l'intérieur de ces deux populations. La grotte Pachón, qui est une des plus étudiées, est isolée des populations de surface. Le bassin principal de cette grotte se situe à 202 m, bien au dessus de la ligne d'eau de base, ce qui en fait la grotte abritant des *Astyanax mexicanus* la plus haute et la plus isolée [66].

2 ÉTUDE DU POLYMORPHISME GÉNÉTIQUE DES POPULATIONS DE SURFACE ET PACHÓN

Nous cherchons à estimer le temps de divergence entre une population cavernicole et une population de surface en utilisant le polymorphisme entre ces deux populations. Nous devons donc dans un premier temps caractériser ce polymorphisme. L'étude du polymorphisme entre notre population cavernicole d'intérêt, celle de la grotte Pachón, et une population de surface provenant du Texas est l'objet de ce chapitre.

2.1 Matériel et méthodes

2.1.1 *Poissons étudiés*

Les poissons de surface ont été pêchés au Balmorhea State Park, Texas (USA). Les poissons cavernicoles proviennent de la grotte Pachón, Antiguo Morelos, Tamaulipas (Mexique). Les poissons de la population de surface et de la grotte Pachón sont maintenus à l'animalerie centrale du campus CNRS de Gif-sur-Yvette avec un cycle jour/nuit de 12h/12h à une température d'environ 25 °C depuis 15 ans.

2.1.2 *Groupe externe*

Nous avions, à l'origine, choisi d'utiliser le poisson zèbre (*Danio rerio*) comme groupe externe. Ce cyprinidé représentait l'espèce la plus proche d'*Astyanax mexicanus* dont le génome fût séquencé. Or les cyprinidés et les characiformes (dont fait partie *Astyanax mexicanus*) divergent depuis au moins 150 millions d'années [106]. Il nous est rapidement apparu que l'ancienneté de cette divergence faisait du poisson zèbre un mauvais groupe externe puisqu'à de nombreuses positions plusieurs mutations successives avaient pu avoir lieu. Nous avons remarqué que lorsqu'il y avait un site polymorphe entre populations ou à l'intérieur d'une population d'*Astyanax mexicanus*, un troisième état était observé chez le poisson zèbre dans 20% des cas.

Nous avons donc choisi d'utiliser un autre groupe externe, le characiforme *Hyphessobrycon anisitsi*. Sur le gène *RAG2*, impliqué dans la recombinaison V(D)J et couramment utilisé en phylogénie (Figure 20) [107–109], la divergence entre ce poisson et *Astyanax mexicanus* est d'environ 4%. Pour ce gène, la divergence est d'environ 30% entre *Astyanax mexicanus* et le poisson zèbre.

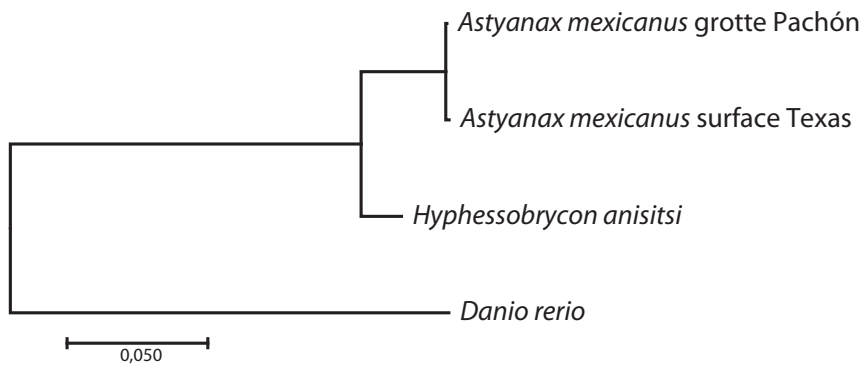


Figure 20. Arbre phylogénétique du gène *RAG2* inféré en utilisant la méthode du Neighbor-Joining [110]. Les distances évolutives ont été calculées en utilisant la correction de Kimura à 2 paramètres [111]. Les longueurs de branches ont pour unité le nombre de substitutions par site. Les positions contenant des indels ou des données manquantes ont été éliminées. Au total, il y a 631 positions dans le jeu de données. Les analyses ont été réalisées en utilisant le logiciel MEGA7 [112].

Avec ce nouveau groupe externe, seules 2% des positions pour lesquelles du polymorphisme est observé chez *Astyanax mexicanus* possèdent un troisième état.

Les poissons de l'espèce *Hyphessobrycon anisitsi* ont été achetés dans le commerce en 2012.

Figure 21. *Hyphessobrycon anisitsi*

2.1.3 *Transcriptomique*

Le transcriptome d'embryons poolés (entre 50 et 200 individus provenant de différentes pontes indépendantes) de la population Texas d'une part et de la population Pachón d'autre part ont été séquencés en utilisant

trois techniques différentes : Sanger [58] et Roche 454 dans un premier temps puis Illumina HiSeq2000 (2x100pb paired-end). En plus de ces deux populations, des embryons du groupe externe ont également été séquencés en utilisant la technologie Illumina (2x100pb paired-end).

Un premier assemblage a été réalisé en utilisant les données issues du séquençage Sanger et 454, en utilisant le logiciel Newbler v2.8 par la plateforme de Bioinformatique GenoToul de l'INRA de Toulouse. Cet assemblage a permis d'obtenir 33 400 contigs. Les lectures issues du séquençage Illumina (*Astyanax mexicanus* et *Hyphessobrycon anisitsi*) ont ensuite été alignées sur les contigs issus de l'assemblage des séquences 454 en utilisant BWA [113].

Nous avons ensuite réalisé une recherche de SNP* entre les deux populations d'*Astyanax mexicanus* mais aussi au sein de chaque population. Cette recherche a été réalisée grâce à GATK UnifiedGenotyper [114]. Comme les SNP identifiés seront nettoyés en utilisant différents filtres, nous avons, lors de cette recherche, utilisé les options -rf BadCigar et allowPotentiallyMisencodedQuals. Ces options permettent de ne pas tenir compte de la qualité des bases lues lors de la détection. Nous avons obtenu 224 847 SNP. Le nombre de lectures de chaque allèle dans chaque population pour les différents SNP identifiés a ensuite été retrouvé grâce à SamTools mpileup [115].

Single Nucleotide Polymorphism *: (SNP) Polymorphisme au niveau d'une seule paire de bases.*

Les SNP identifiés ont été enregistrés dans une base de données MySQL.

2.1.4 *Annotation des contigs*

L'annotation des contigs a été réalisée par comparaison avec les séquences protéiques du *Danio rerio*. En effet, si ce dernier est trop éloigné pour orienter de façon fiable des changements au niveau des nucléotides, il permet néanmoins d'identifier les séquences protéiques car c'est le génome le plus proche d'*Astyanax mexicanus* qui est bien annoté.

Les contigs assemblés ont été blastés (blastx, séquence nucléique contre banque de séquences protéiques) [116] contre l'ensemble des protéines du *Danio rerio* (Zv9) présentes dans EnsEMBL 73 [117]. Ce blast permet à la fois d'annoter le contig et de récupérer la séquence codante correspondante. Des régions flanquantes de la séquence codante étant également séquencées, les portions de contigs sont annotées comme codantes, non-codantes ou inconnues lorsque la séquence codante n'a pas été identifiée en intégralité (Figure 22).

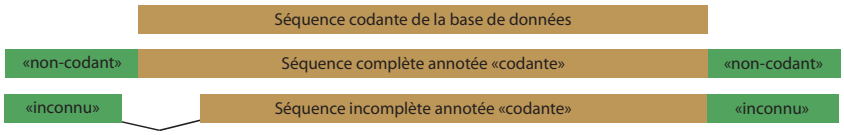


Figure 22. Annotation des régions codantes et non codantes des contigs identifiés. Si la séquence codante n'a pas été retrouvée entièrement, les séquences flanquantes sont annotées comme "inconnu".

e-value *: Statistique indiquant la probabilité d'obtenir un alignement entre la séquence requête et la séquence trouvée en utilisant une base de données de séquences aléatoires de même taille.*

L'identifiant de la séquence ayant la meilleure e-value* lors du blast permet ensuite, en utilisant BioMart [118], d'aller récupérer sur EnsEMBL les identifiants de la séquence dans différentes bases de données biologiques (EnsEMBL, Entrez, zfin).

2.1.5 *Filtrage des SNP*

Afin de tenir compte de possibles erreurs de séquençage nous avons utilisé plusieurs filtres afin de nettoyer notre jeu de données. Les filtres utilisés concernent la profondeur de séquençage, la fréquence de l'allèle minoritaire, la distance entre deux SNP consécutifs. Nous avons également utilisé un filtre sur la qualité de l'annotation en faisant varier la e-value des blasts. Pour chaque filtre, nous avons testé plusieurs valeurs seuils puis nous avons gardé un jeu de valeurs seuils permettant à la fois de nettoyer le jeu de données sans toutefois enlever trop de données.

2.1.5.1 *Profondeur de séquençage*

Le premier filtre utilisé concerne la profondeur de séquençage.

Dans chaque population, une profondeur minimale de séquençage est définie. Il doit donc y avoir un nombre minimum de lectures de l'ensemble des allèles présents au locus où est présent le SNP. Nous avons testé différents seuils pour ce filtre : 4, 10, 20, 30, 50, 100, 200 et 400 lectures. Au seuil de 100, chaque population doit avoir au moins 100 lectures à un locus donné pour que celui-ci soit conservé.

De plus, nous avons choisi de ne garder que les allèles qui avaient au moins deux lectures dans la population où ils sont retrouvés.

Enfin, chaque allèle doit être présent à une fréquence minimum (MAF, *Minimum Allele Frequency*) dans la population. Nous avons testé des fréquences de 1%, 2% 5% ou 10%.

2.1.5.2 *Isolement des SNP*

Dans une région mal séquencée, des successions de SNP peuvent être introduites par des erreurs de séquençage. Nous avons donc défini une zone en amont et en aval d'un SNP dans laquelle aucun autre SNP ne peut être présent. Nous avons testé plusieurs tailles d'isolement : 0, 20, 50, 100, 200 et 300 nucléotides de part et d'autre du SNP.

2.1.5.3 *blast*

Nous avons également utilisé différentes valeurs de e-value des blasts entre les contigs d'*Astyanax mexicanus* et les protéines du *Danio rerio* afin de ne garder que des séquences bien conservées et donc avec un homologue fiable et une bonne annotation. Pour ce filtre, les différentes valeurs seuils utilisées sont 10^{-5}, 10^{-10}, 10^{-20} et 10^{-50}.

2.1.6 *Classification des SNP*

Nous avons classé les SNP en huit catégories en fonction du polymorphisme intra et inter populations (Tableau 1). Nous considérons que l'allèle présent chez le groupe externe, *Hyphessobrycon anisitsi*, est l'allèle ancestral. L'allèle dérivé est défini par rapport à l'allèle présent chez *Hyphessobrycon anisitsi* : si l'allèle présent chez *Astyanax mexicanus* est différent de celui présent chez *Hyphessobrycon anisitsi*, alors il s'agit de l'allèle dérivé, sinon il s'agit de l'allèle ancestral. Seules les positions pour lesquelles *Hyphessobrycon anisitsi* est non-polymorphe sont utilisées. De plus l'allèle présent chez le groupe externe doit également être présent chez *Astyanax mexicanus* car sinon il est impossible de définir l'allèle ancestral et l'allèle dérivé chez *Astyanax mexicanus*.

Dans le Tableau 1, *x*, *y* et *z* représentent les allèles présents : *x* représente l'allèle ancestral alors que *y* et *z* représentent des allèles dérivés. On peut alors définir huit catégories de SNP. On peut observer du polymorphisme entre la population de surface et la population cavernicole mais pas au sein des populations (catégorie 1 et 2) : un allèle dérivé est fixé dans la population de surface (catégorie 1) ou dans la population cavernicole (catégorie 2) alors que l'allèle ancestral est fixé dans l'autre population. Une des deux populations peut présenter un polymorphisme alors que l'autre population a un allèle fixé : l'autre population a pu fixer l'allèle ancestral (catégories 3 et 5) ou un allèle dérivé (catégories 4 et 6). Enfin, les deux populations peuvent être simultanément polymorphes, que ce polymorphisme soit partagé (les mêmes allèles dans les deux populations) (catégorie 7) ou divergent (des allèles différents dans les deux populations) (catégorie 8). Ce dernier cas devrait être rare car il implique que deux mutations différentes ont eu lieu à une même position, ce qui est peu probable puisque le temps moyen entre deux mutations est égal à l'inverse du taux de mutation, de l'ordre de 10^8 générations, et donc bien plus grand que le temps moyen de fixation d'une mutation, 4N lorsque sa fréquence est la plus basse ($\frac{1}{2N}$) [7], de l'ordre de 10^4 à 10^5. Lors du séquençage des erreurs peuvent créer de nouveaux allèles et donc des locus avec plus de deux allèles. Aussi ce cas nous permet de réaliser un contrôle de la qualité du séquençage puisque si on observe de nombreux locus avec plus de deux allèles, il se peut qu'il y ait eu un problème lors du séquençage. Aucune autre catégorie n'est possible puisque l'allèle ancestral doit être retrouvé dans une des populations d'*Astyanax mexicanus* et qu'il doit y avoir un polymorphisme entre la population de surface et la population cavernicole.

Catégorie	1	2	3	4	5	6	7	8
Poisson de surface	x	y	x	y	x/y	x/y	x/y	x/y
Poisson cavernicole	y	x	x/y	x/y	x	y	x/y	x/z
Hyphessobrycon anisitsi	x	x	x	x	x	x	x	x

Tableau 1. Les huit catégories de SNP observables en fonction du polymorphisme. *x*, *y* et *z* représentent les allèles à un locus donné : *x* est l'allèle présent chez le groupe externe *Hyphessobrycon anisitsi* que l'on considère comme l'allèle ancestral, *y* et *z* sont des allèles dérivés. On peut observer du polymorphisme, c'est-à-dire deux allèles différents, entre les deux populations d'*Astyanax mexicanus* mais pas au sein de chacune d'entre elles : l'allèle ancestral est fixé dans une population, l'allèle dérivé dans l'autre (catégorie 1 et 2). Il peut y avoir du polymorphisme dans une des populations avec la fixation d'un allèle dans l'autre population : l'allèle fixé peut être l'allèle ancestral (catégorie 3 et 5) ou un allèle dérivé (catégorie 4 et 6). Les deux populations peuvent être simultanément polymorphes : le polymorphisme peut alors être partagé (catégorie 7) lorsque les deux populations ont les mêmes allèles ou divergent (catégorie 8) lorsque les allèles sont différents dans les deux populations. Les cases de couleurs représentent des regroupements de ces huit catégories : allèle dérivé fixé dans la population cavernicole (en bleu, catégorie 1 et 6), dans la population de surface (en rouge, catégorie 2 et 4), polymorphisme dans la population cavernicole (en violet, catégorie 3, 4 et 7), dans la population de surface (en orange, catégorie 5 à 7) et le polymorphisme partagé (ovale vert, catégorie 7).

2.2 Résultats

2.2.1 *Effet des filtres*

Pour chacun des filtres, nous avons testé différentes valeurs seuils. Nous avons ensuite regardé si les différentes valeurs testées modifiaient le classement des SNP dans les huit catégories précédemment présentées (Tableau 1).

2.2.1.1 *Profondeur de lecture dans chaque population*

Le premier filtre utilisé est la profondeur de lecture dans chaque population. Nous avons testé plusieurs valeurs pour ce filtre : 4, 10, 20, 30, 50, 100, 200 et 400 lectures.

La Figure 23 présente le nombre de SNP dans chacune des huit catégories définies Tableau 1 ainsi que la fréquence relative de ces différentes catégories en fonction de la profondeur de lecture minimale par population. Lorsque la profondeur minimale par population augmente, le nombre de SNP dans chacune des huit catégories diminue.

La fréquence relative des SNP dans les huit catégories est semblable quel que soit le seuil utilisé. L'augmentation du seuil pour ce filtre permet donc d'éliminer des positions avec peu de lectures sans toutefois biaiser les résultats obtenus, mais il ne semble pas nécessaire d'avoir une grande profondeur pour avoir des estimations fiables des fréquences relatives des huits catégories.

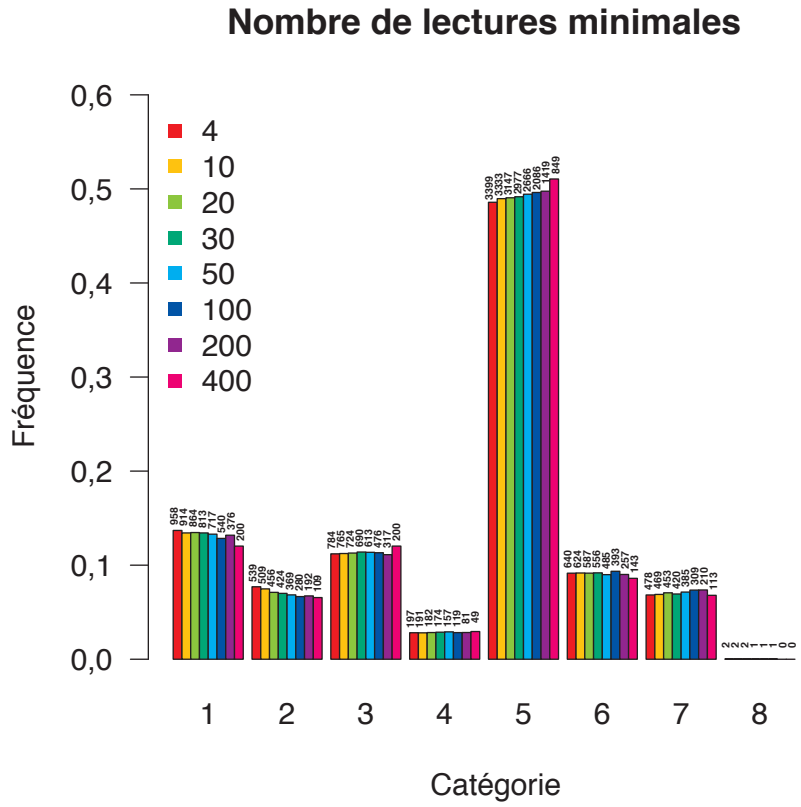


Figure 23. Fréquences relatives des différentes catégories et nombre de SNP dans ces huit catégories pour les différents seuils de profondeur de lecture dans chaque population. La hauteur des barres représente la fréquence relative des catégories. Le nombre de SNP est indiqué au dessus des barres. Les autres paramètres sont : fréquence de l'allèle minoritaire supérieure à 5%, isolement des SNP de 50 pb et une e-value inférieure à 10^{-5}.

2.2.1.2 *Fréquence minimale de l'allèle minoritaire*

Le deuxième filtre appliqué concerne la fréquence minimale de l'allèle minoritaire pour considérer que le site est polymorphe. Nous avons testé plusieurs valeurs seuils pour ce filtre : 1%, 2%, 5% et 10%.

La Figure 24 présente le nombre de SNP dans chacune des huit catégories définies Tableau 1 ainsi que la fréquence relative de ces différentes catégories en fonction de la fréquence minimale de l'allèle minoritaire. Le nombre

de SNP dans les catégories 1 et 2 (SNP fixés dans les deux populations, Tableau 1) augmente lorsque la fréquence minimale de l'allèle minoritaire augmente. Ceci est dû à des SNP polymorphes dans une ou dans les deux populations mais dont l'allèle minoritaire est à basse fréquence et en dessous du seuil. Ces SNP sont alors considérés comme étant fixés dans les deux populations. Par exemple, si à un locus polymorphe on observe deux allèles dont un est à la fréquence de 3% et l'autre à la fréquence de 97%, ce locus sera considéré comme polymorphe au seuil de 2% mais au seuil de 5% l'allèle majoritaire sera considéré comme fixé.

Dans les six autres catégories le nombre de SNP diminue lorsque la fréquence minimale des allèles est augmentée.

Les fréquences relatives des huit catégories varient faiblement lorsque l'on augmente le pourcentage de présence de l'allèle minoritaire. Cela permet donc d'éliminer des allèles à faible fréquence qui pourrait être dus à des erreurs de séquençage sans toutefois modifier fortement les fréquences des différentes catégories.

2.2.1.3 *Isolement des SNP*

Le troisième filtre utilisé est l'isolement des SNP. Pour ce filtre nous avons considéré des zones de différentes tailles : 0 pb, 20 pb, 50 pb, 100 pb, 200 pb ou 300 pb de part et d'autre du SNP.

La Figure 25 présente le nombre de SNP dans chacune des huit catégories définies au Tableau 1 ainsi que la fréquence relative de ces catégories en fonction de la taille de la zone d'exclusion.

Le nombre de SNP diminue dans toutes les catégories lorsque la taille de l'isolement augmente. Les fréquences relatives des différentes catégories de SNP varient faiblement lorsque l'isolement augmente. À partir d'un isolement de 20 pb, la fréquence de SNP partagé entre populations diminue et se stabilise. Ces SNP peuvent provenir de petites régions mal séquencées ou mal assemblées.

2.2.1.4 *e-value*

Le dernier filtre concerne la e-value des blasts réalisés pour l'annotation des contigs. Pour ce filtre, nous avons testé des seuils de 10^{-5}, 10^{-10}, 10^{-20} et 10^{-50}.

Le nombre de SNP dans les huit catégories définies Tableau 1 ainsi que la fréquence relative de ces catégories pour les différents seuils de ce filtre sont présentés Figure 26.

Lorsque la e-value est diminuée, le nombre de SNP reste assez stable. Cela montre que la plupart des contigs qui ont été annotés avaient une e-value bien inférieure à 10^{-5} et que cette e-value est assez faible pour obtenir des résultats fiables. Ainsi la plupart des séquences utilisées sont, au niveau protéique, très conservées entre *Astyanax mexicanus* et *Danio rerio*. La fréquence relative des différentes catégories est elle aussi stable.

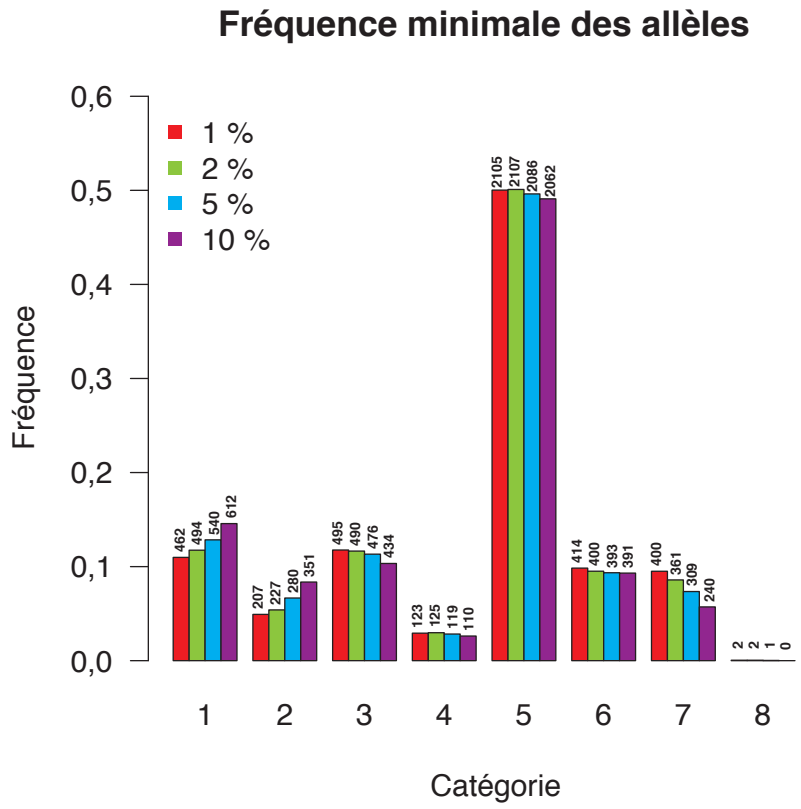


Figure 24. Fréquences relatives des huit catégories de SNP et nombre de SNP dans ces catégories pour les différentes valeurs de fréquence minimale de l'allèle minoritaire. La hauteur des barres représente la fréquence relative des catégories. Le nombre de SNP dans les catégories est indiqué au dessus des barres. Les autres paramètres sont : Nombre de lectures minimales par population de 100, isolement des SNP de 50 pb et une e-value inférieure à 10^{-5}.

Les résultats présentés par la suite ont été obtenus en utilisant les filtres avec les valeurs seuil suivantes :

- la profondeur de lecture à une position est d'au moins 100 par population ;
- l'allèle minoritaire doit être présent à une fréquence d'au moins 5% ;
- l'isolement des SNP est de 50 pb de part et d'autre du SNP ;
- une e-value inférieure à 10^{-5}.

Ces seuils permettent à la fois de s'assurer que les SNP utilisés dans l'analyse sont réellement des SNP tout en ayant un nombre suffisant grand de loci étudiés.

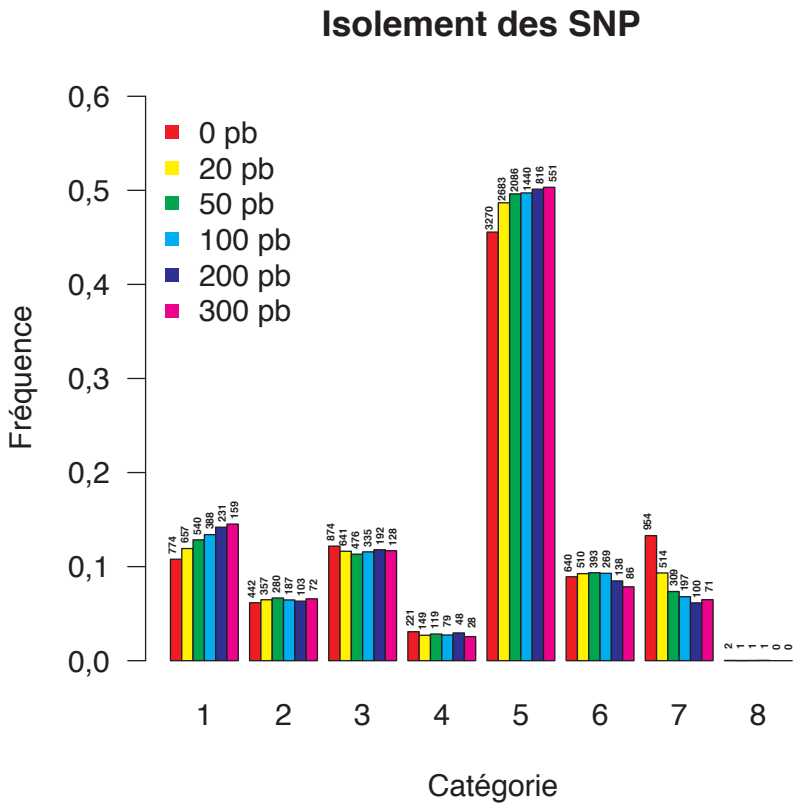


Figure 25. Fréquences relatives des huit catégories de SNP et nombre de SNP dans ces catégories pour les différentes tailles de zone d'exclusion. La hauteur des barres représente la fréquence relative des catégories. Le nombre de SNP dans les catégories est indiqué au dessus des barres. Les autres paramètres sont : Nombre de lectures minimales par population de 100, fréquence de l'allèle minoritaire supérieure à 5%, et une e-value inférieure à 10^{-5}.

2.2.2 *Nombre de SNP par catégorie*

Le nombre de SNP identifiés, avec les seuils définis précédemment ainsi que les fréquences relatives des différentes catégories de SNP sont présentés Tableau 2.

Les SNP observés peuvent provenir de la partie codante du contig ou d'une région flanquante non codante. Un SNP dans une séquence codante peut être synonyme si la mutation n'entraîne pas de changement d'acide aminé dans la séquence protéique ou non-synonyme dans le cas contraire.

Le nombre de SNP dans les huit catégories définies Tableau 1 ainsi que les fréquences relatives des différentes catégories dans les séquences non-codantes et pour les mutations synonymes et non-synonymes sont présentés Tableau 3. Les niveaux de polymorphisme et de fixation des allèles dérivés y sont également présentés par population.

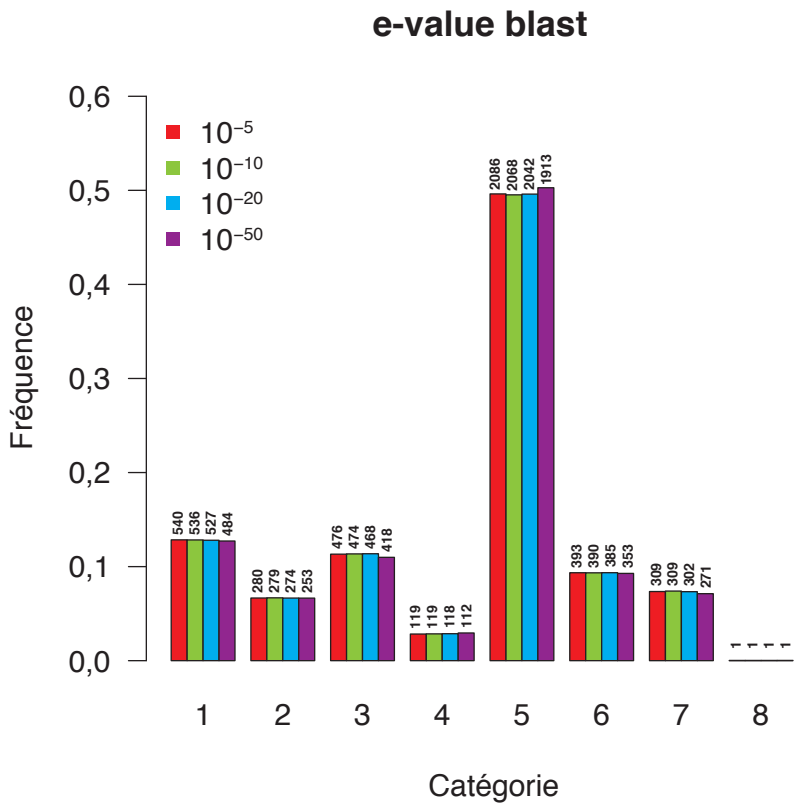


Figure 26. Fréquences relatives des huit catégories de SNP et nombre de SNP dans ces catégories pour les différents seuils de e-value. La hauteur des barres représente la fréquence relative des catégories. Le nombre de SNP dans les catégories est indiqué au dessus des barres. Les autres paramètres sont : Nombre de lectures minimales par population de 100, fréquence de l'allèle minoritaire supérieure à 5% et isolement des SNP de 50 pb.

Le nombre de SNP est très différent selon que l'on s'intéresse aux régions non-codantes (16% des SNP) ou codantes que la mutation soit synonyme (55% des SNP) ou non-synonyme (28% des SNP). Cela est dû au fait qu'il y a plus de codant que de non-codant dans le transcriptome.

Les fréquences relatives des différentes catégories sont significativement différentes entre mutations non-codantes, synonymes ou non-synonymes (test de χ^2, p-value $< 10^{-16}$). Si ces différences sont significatives, elles n'entraînent toutefois pas de changement de l'importance relative des différentes catégories.

Catégorie	Allèle SF	Allèle CF	Nombre de SNP	%
(1)	ancestral fixé	dérivé fixé	998	13 %
(2)	dérivé fixé	ancestral fixé	602	8 %
(3)	ancestral fixé	polymorphe	924	12 %
(4)	dérivé fixé	polymorphe	267	4 %
(5)	polymorphe	ancestral fixé	3 610	48 %
(6)	polymorphe	dérivé fixé	638	8 %
(7)	Polymorphisme partagé		512	7 %
(8)	Polymorphisme divergent		1	0 %
	Total		**7 552**	100 %

Tableau 2. Répartition des SNP dans les huit catégories définies Tableau 1.

Catégorie	Allèle SF	Allèle CF	SNP Non codants		SNP Synonymes		SNP non-synonymes		Total
(1)	ancestral fixé	dérivé fixé	157	*(12,7%)*	540	*(12,8%)*	301	*(14,3%)*	998
(2)	dérivé fixé	ancestral fixé	111	*(9%)*	280	*(6,7%)*	211	*(10%)*	602
(3)	ancestral fixé	polymorphe	146	*(11,8%)*	476	*(11,3%)*	302	*(14,3%)*	924
(4)	dérivé fixé	polymorphe	57	*(4,6%)*	119	*(2,8%)*	91	*(4,3%)*	267
(5)	polymorphe	ancestral fixé	601	*(48,5%)*	2 086	*(49,6%)*	923	*(43,7%)*	3 610
(6)	polymorphe	dérivé fixé	87	*(7%)*	393	*(9,3%)*	159	*(7,5%)*	638
(7)	Polymorphisme partagé		80	*(6,5%)*	309	*(7,4%)*	123	*(5,8%)*	512
(8)	Polymorphisme divergent		0	*(0%)*	1	*(0%)*	0	*(0%)*	1
	Total		**1 239**	*(100%)*	**4 204**	*(100%)*	**2 110**	*(100%)*	**7 552**
(5+6+7)	**Polymorphisme SF**		768		2 788		1 205		4 760
(3+4+7)	**Polymorphisme CF**		283		904		516		1 703
	Ratio polymorphisme SF/CF		2,71		3,08		2,34		2,80
(2+4)	**Dérivé et fixé SF**		168		399		302		869
(1+6)	**Dérivé et fixé CF**		244		933		460		1 636
	Ratio dérivé et fixé CF/SF		1,45		2,34		1,52		1,88

Tableau 3. Nombre de SNP dans chacune des huit catégories définies Tableau 1 et fréquences relatives de ces catégories pour les SNP non-codants, synonymes et non-synonymes. Niveaux de polymorphisme et de fixation des SNP par population et ratios entre populations.

2.2.3 *Comparaison entre le nombre de substitutions observées et le nombre attendu dans un modèle neutre*

Pour les SNP où au moins une des deux populations est polymorphe (catégorie 3 à 8, Tableau 1) et ceux où un allèle dérivé est fixé (catégories 1,2,4 et 6, Tableau 1), nous pouvons observer que respectivement 68% et 64% d'entre eux sont synonymes (Tableau 4).

	Sites polymorphes		Sites dérivés fixés	
Synonymes	3 384	*68%*	1 332	*64%*
Non-synonymes	1 598	*32%*	762	*36%*

Tableau 4. Nombre et pourcentage de mutations synonymes et non-synonymes pour les SNP où au moins une des deux populations est polymorphe (catégories 3 à 8, Tableau 1) et pour les SNP où un allèle dérivé est fixé (catégories 1,2, 4 et 6, Tableau 1).

On peut alors se demander si ces valeurs sont celles attendues dans un modèle d'évolution neutre.

K80 *: Modèle d'évolution moléculaire dans lequel on distingue les transitions* des transversions*.*
Transitions *: Mutations A/G et T/C.*
Transversions *: Mutations A/C, A/T, G/T et G/C.*

Pour calculer le nombre attendu, nous avons dans un premier temps compté, en utilisant le modèle d'évolution moléculaire K80* [111] et le code génétique (Tableau 5), le nombre de positions synonymes et non-synonymes. À chaque position d'un codon, trois types de mutations peuvent avoir lieu. Ces mutations peuvent être synonymes ou non-synonymes, en fonction de la dégénérescence du code génétique. Lorsque toutes les mutations possibles sont synonymes, la position est quatre fois dégénérée. Au contraire si toutes les mutations possibles sont non-synonymes, la position est zéro fois dégénérée. D'autres positions peuvent avoir une mutation synonyme et deux mutations non-synonymes. On parle alors de positions deux fois dégénérées. Une position quatre fois dégénérée donnera donc une mutation synonyme alors qu'une position zéro fois dégénérée donnera une mutation non-synonyme. Pour les positions deux fois dégénérées, il faut tenir compte du taux de transitions et de transversions pour calculer le nombre de mutations synonymes et non-synonymes. Les transversions sont moins fréquentes que les transitions. Le nombre de mutations synonymes et non-synonymes données par une position sera respectivement de $\frac{\alpha}{\alpha+2\beta}$ et de $\frac{2\beta}{\alpha+2\beta}$, où α est la fréquence des transitions et 2β la fréquence des transversions [111].

Dans les séquences d'*Astyanax mexicanus*, nous avons calculé, en utilisant l'ensemble des SNP synonymes, qu'il y avait trois transitions pour une transversion, la proportion de positions deux fois dégénérées donnant des mutations synonymes devient donc $\frac{3}{5}$ et le nombre de positions donnant des mutations non-synonymes $\frac{2}{5}$.

En appliquant ce calcul aux 2 456 691 codons présents dans les régions codantes du transcriptome d'*Astyanax mexicanus*, on trouve que 74% (5 432 051) des mutations possibles sont non-synonymes et 26% (1 938 022) sont synonymes, ce qui est conforme aux valeurs attendues.

Les valeurs observées sont donc significativement différentes de ces valeurs théoriques aussi bien pour les SNP polymorphes (test de χ^2, $\chi =$ 4 446,4, p-val = 0) que pour les SNP dérivés fixés ($\chi =$ 1 501,6, p-val = 0). Il y a donc beaucoup moins de mutations non-synonymes qu'attendu par le simple fait du hasard. Ce résultat suggère une sélection négative sur les SNP non-synonymes.

	T		C		A		G	
T	TTT	F	TCT	S	TAT	Y	TGT	C
	TTC	F	TCC	S	TAC	Y	TGC	C
	TTA	L	TCA	S	TAA	*	TGA	*
	TTG	L	TCG	S	TAG	*	TGG	W
C	CTT	L	CCT	P	CAT	H	CGT	R
	CTC	L	CCC	P	CAC	H	CGC	R
	CTA	L	CCA	P	CAA	Q	CGA	R
	CTG	L	CCG	P	CAG	Q	CGG	R
A	ATT	I	ACT	T	AAT	N	AGT	S
	ATC	I	ACC	T	AAC	N	AGC	S
	ATA	I	ACA	T	AAA	K	AGA	R
	ATG	M	ACG	T	AAG	K	AGG	R
G	GTT	V	GCT	A	GAT	D	GGT	G
	GTC	V	GCC	A	GAC	D	GGC	G
	GTA	V	GCA	A	GAA	E	GGA	G
	GTG	V	GCG	A	GAG	E	GGG	G

Tableau 5. Code génétique. Les couleurs montrent la dégénérescence du code : les positions en rouge sont deux fois dégénérées, celle en bleu sont quatre fois dégénérées.

2.2.4 *Conséquence des changements d'acides aminés*

Les SNP non-synonymes peuvent être classés en fonction du changement d'acide aminé. Pour réaliser ce classement, nous avons utilisé la matrice de Grantham [119] qui donne une distance entre chaque acide aminé en fonction de trois de leurs propriétés physico-chimiques : la composition, la polarité et le volume moléculaire.

La composition est définie comme le ratio du poids moléculaire des éléments non-carbonés dans le groupe terminal (ou le cycle) par rapport au poids des carbones dans la chaine latérale. Par exemple, pour la lysine, dont la formule développée est donnée Figure 27, le groupe terminal de la

chaîne latérale est NH_2 et la chaîne est composée de 4 groupements CH_2. La composition de la lysine est donc de $\frac{N+2\times H}{4*(C+2\times H)} = \frac{16}{48} \approx 0,33$.

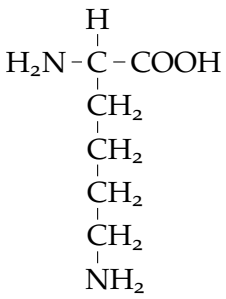


Figure 27. Formule développée de la Lysine

La matrice de Grantham est présentée Tableau 6. Pour établir qu'une substitution est conservative ou radicale, nous avons utilisé un seuil correspondant à la moitié de la plus grande valeur de cette matrice par analogie avec la méthode proposée par *Li et al.* [120]. Le changement le plus radical est la substitution Cys/Trp avec un score de 215. Ainsi un changement d'acide aminé avec un score supérieur à 107,5 sera considéré comme radical, alors qu'un changement avec un score inférieur sera considéré comme conservatif.

	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	195	126	107	113	60	86	94	106	96	84	111	27	91	112	99	58	64	148	112
C		154	170	205	159	174	198	202	198	196	139	169	154	180	112	149	192	215	194
D			45	177	94	81	168	101	172	160	23	108	61	96	65	85	152	181	160
E				140	98	40	134	56	138	126	42	93	29	54	80	65	121	152	122
F					153	100	21	102	22	28	158	114	116	97	155	103	50	40	22
G						98	135	127	138	127	80	42	87	125	56	59	109	184	147
H							94	32	99	87	68	77	24	29	89	47	84	115	83
I								102	5	10	149	95	109	97	142	89	29	61	33
K									107	95	94	103	53	26	121	78	97	110	85
L										15	153	98	113	102	145	92	32	61	36
M											142	87	101	91	135	81	21	67	36
N												91	46	86	46	65	133	174	143
P													76	103	74	38	68	147	110
Q														43	68	42	96	130	99
R															110	71	96	101	77
S																58	124	177	144
T																	69	128	92
V																		88	55
W																			37

Tableau 6. Matrice de Grantham [119]. Cette matrice présente la distance entre acide aminé déterminée en fonction de trois de leurs propriétés physico-chimiques : composition, polarité et volume moléculaire. La distance entre deux acides aminés i et j est égale à : $D_{i,j} = \sqrt{\alpha(c_i - c_j)^2 + \beta(p_i - p_j)^2 + \gamma(v_i - v_j)^2}$, où c, p et v sont respectivement la composition, la polarité et le volume moléculaire et α, β et γ sont respectivement les inverses des poids moyens de la composition, de la polarité et du volume moléculaire. *en vert : substitutions conservatives (distance inférieure à 107,5), en rouge : substitutions radicales (distance supérieure à 107,5).*

Nous avons compté le nombre de changements d'acides aminés conservatifs et le nombre de changements radicaux pour les huit catégories de SNP (Tableau 7).

Catégorie	Changements conservatifs		Changements radicaux	
1	254	*(13,7%)*	47	*(18,5%)*
2	188	*(10,1%)*	23	*(9,1%)*
3	269	*(14,5%)*	33	*(13%)*
4	81	*(4,4%)*	10	*(3,9%)*
5	809	*(43,6%)*	114	*(44,9%)*
6	145	*(7,8%)*	14	*(5,5%)*
7	110	*(5,9%)*	13	*(5,1%)*
8	0	*(0%)*	0	*(0%)*
Total	**1 856**	*(100%)*	**254**	*(100%)*

Tableau 7. Nombre de SNP non-synonymes par catégorie et fréquences relatives de ces catégories en fonction du type de changement d'acides aminés.

En utilisant uniquement les polymorphismes dérivés fixés (cas 1, 2, 4 et 6), on observe 668 substitutions conservatives et 94 substitutions radicales (Tableau 8).

Nous nous sommes demandés si ces nombres de changements conservatifs et radicaux observés correspondaient à ceux attendus si la pression de sélection est la même quel que soit le type de changement d'acide aminé.

En tenant compte du taux de transition et de transversion, du nombre de codons de chaque type présents dans notre transcriptome et de la matrice de Grantham (Tableau 6), on peut calculer les fréquences attendues et le nombre de mutations non-synonymes conservatives et radicales. Avec un seuil de 107,5 pour différencier changements conservatifs et radicaux, 78% des mutations possibles sont conservatives et 22% radicales.

Sur les 762 substitutions observées, il devrait donc y avoir 594 substitutions conservatives et 168 radicales. Or nous observons un excès de substitutions conservatives (668 au lieu de 594) et un déficit de substitutions radicales (94 au lieu de 168) (Tableau 8 et Figure 28). Cette différence est significative (test de χ^2, $\chi = 24{,}347$, p-val $= 8{,}04.10^{-7}$) et va dans le sens d'une plus grande sélection négative des changements radicaux.

Nous avons ensuite séparés ces substitutions par population. On observe alors qu'en nombre, il y a plus de substitutions dans la population cavernicole que dans la population de surface. Si on sépare les substitutions conservatives des substitutions radicales, on observe que ces dernières sont, en fréquence, plus nombreuses dans la population cavernicole que dans la population de surface (Figure 29). Ce résultat pourrait être considéré comme un argument en faveur d'un relâchement des pressions de sélection

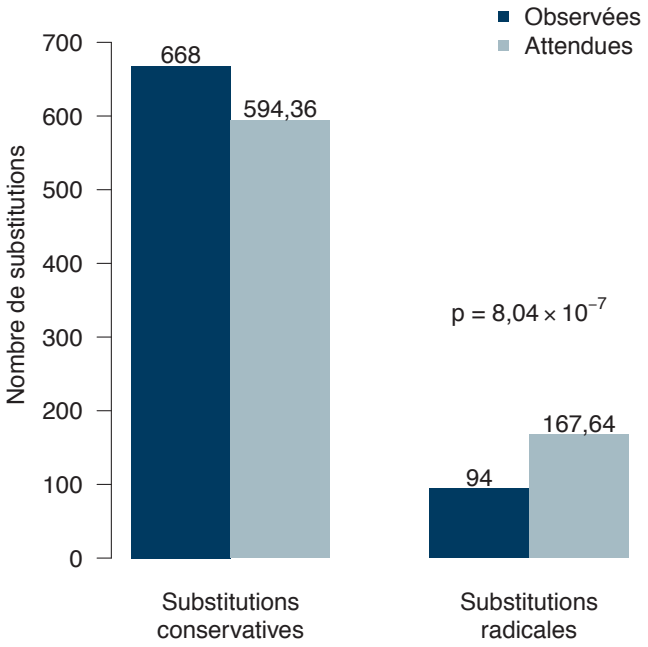


Figure 28. Nombre de substitutions conservatives et radicales observées et attendues.

dans les populations cavernicoles mais n'est pas statistiquement significatif (p-value = 0,4).

Population	Conservative		Radicale		Total
CF	399	*87%*	61	*13%*	460
SF	269	*89%*	33	*11%*	302
Total	668	*88%*	94	*12%*	762

Tableau 8. Répartition des SNP dérivés fixés par population et par type de changement d'acide aminé.

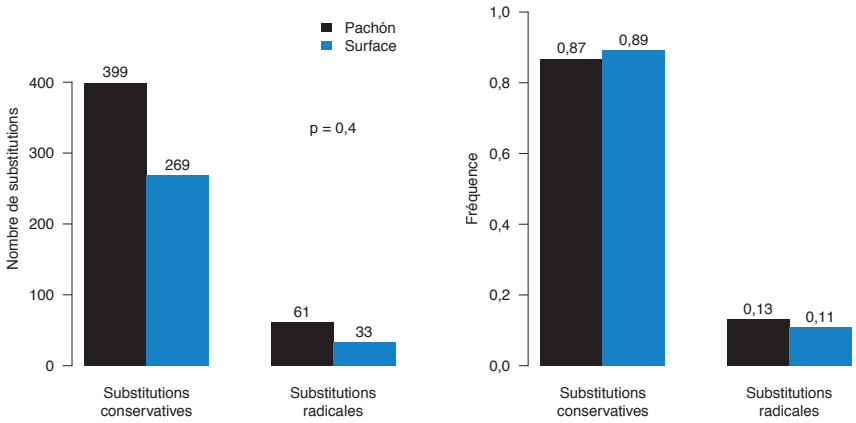


Figure 29. Nombre (à gauche) et fréquence (à droite) des substitutions synonymes et non-synonymes dans la population de surface (en bleu) et cavernicole (en noir).

2.3 Comparaison des niveaux de polymorphismes et de fixation entre les populations

Nous allons par la suite nous intéresser uniquement aux changements synonymes qui peuvent être considérés comme neutres d'un point de vue sélectif et qui peuvent être utilisés pour estimer des paramètres démographiques.

2.3.1 *Rôle de la taille des populations sur les mutations et leur fixation*

Nous considérons ici une population idéale Wright-Fisher dans laquelle la taille efficace est égale à la taille réelle. Le nombre de mutations apparaissant à chaque génération dans une population diploïde est proportionnel au nombre d'allèles et au taux de mutation. Le nombre de nouvelles mutations est donc $2 \times N_e \times \mu$, où N_e représente la taille efficace de la population et μ le taux de mutation.

Si une mutation est neutre, sa probabilité de fixation est égale à sa fréquence. Lorsqu'une nouvelle mutation apparait, un seul individu en est porteur. Sa fréquence est donc $\frac{1}{2 \times N_e}$.

Le temps moyen de fixation d'un nouvel allèle ne dépend que de la taille de la population dans laquelle il apparait ($\overline{t_{fixation}} \approx 4N_e$). Aussi dans les grandes populations, les locus polymorphes pour un nouvel allèle le restent plus longtemps que dans les petites populations.

Lorsqu'une mutation est fixée, on parle de substitution. La probabilité d'observer une substitution est égale à la probabilité d'apparition d'une nouvelle mutation multipliée par la probabilité de fixation. Au final, cette probabilité est donc égale au taux de mutation : $2 \times N_e \times \mu \times \frac{1}{2 \times N_e} = \mu$.

Pour un taux de mutation donné, le nombre de mutations qui apparaissent dans une population de grande taille est plus grand que dans une population de taille plus petite. Mais les mutations vont se fixer avec une

plus grande probabilité dans les petites populations que dans les grandes. Ainsi, pour un taux de mutation donné le nombre de substitutions observées est le même dans une petite population et dans une grande population après une longue période de temps.

Ainsi, quelle que soit la taille de la population, les substitutions neutres vont s'accumuler au même rythme, en terme de mutations par unité de temps.

2.3.1.1 *Simulation*

Nous avons réalisé une simulation de génétique des populations afin de mieux expliquer le devenir des mutations synonymes dans deux populations de tailles différentes après leur divergence.

Une seule population de grande taille ($N_e = 10\ 000$) est présente au début de la simulation. Une deuxième population, dont la taille efficace est deux fois plus petite que dans la première population ($N_e = 5\ 000$) est ensuite créée par échantillonnage de cette première population. Ces deux populations évoluent ensuite indépendamment pendant 100 000 générations (Figure 30). Les forces évolutives présentes sont la mutation et la dérive génétique. Il n'y a ni sélection ni flux de gènes entre les deux populations. On peut alors regarder le niveau de polymorphisme et de fixation d'allèles dérivés dans les deux populations au cours du temps.

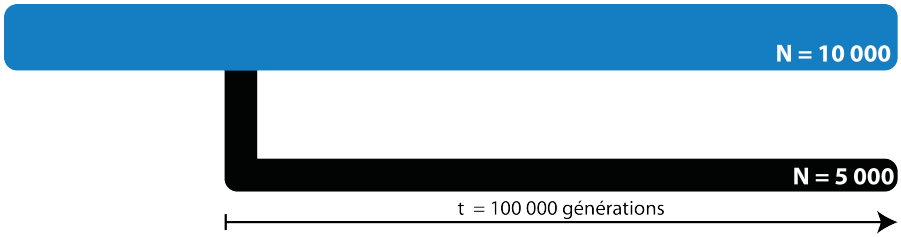


Figure 30. Modèle simple de modélisation de génétique des populations. Après un certain nombre de générations, une population de petite taille (en noir) est créée par échantillonnage dans la population de grande taille (en bleu).

2.3.2 *Positions polymorphes*

Dans la simulation, le nombre de sites polymorphes décroit rapidement dans la plus petite des deux populations avant de se stabiliser à deux fois moins de sites que dans la grande population (Figure 31), ce qui est attendu car à l'équilibre mutation/dérive, le polymorphisme ne dépend que de N_e.

Dans l'analyse du polymorphisme synonyme d'*Astyanax mexicanus* présentée précédemment (Tableau 3), nous avons observé environ trois fois plus de sites qui sont polymorphes dans la population de surface que dans la population cavernicole (2 788 vs 904). Une différence de niveau de polymorphisme est, comme nous l'avons vu avec la simulation, explicable par une différence de taille des populations. Si les populations sont à l'équilibre

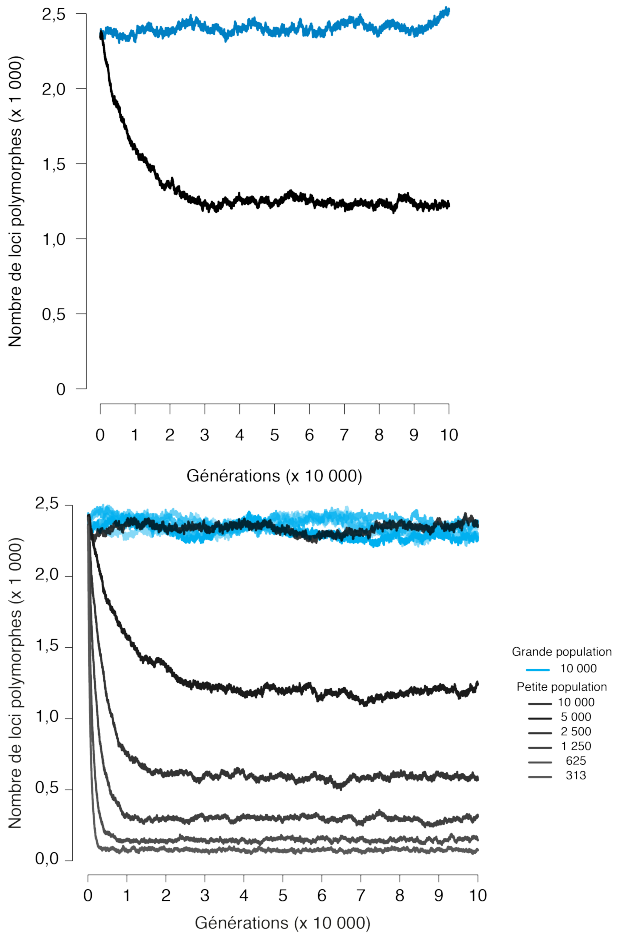


Figure 31. Nombre de sites polymorphes au cours du temps dans deux populations de taille différente lors de la simulation présentée Section 2.3.1.1. (En haut) En bleu, la plus grande des populations ($N_e = 10\ 000$) et en noir la plus petite ($N_e = 5\ 000$). (En bas) Six simulations différentes : En bleu, la plus grande des populations et cinq réplicats ($N_e = 10\ 000$) et en noir la plus petite des deux (N_e variable).

mutation/dérive cela suggère que la taille efficace de la population cavernicole est environ 3 fois plus petite que celle de la population de surface.

2.3.3 *Polymorphisme partagé*

Dans la simulation (Section 2.3.1.1), le nombre de sites où les deux populations sont polymorphes pour les mêmes allèles (polymorphisme partagé, catégorie 7 Tableau 3) est très élevé en début de simulation, car quasiment tout le polymorphisme est partagé : la plus petite des deux populations étant échantillonnée dans la plus grande, la plupart des sites polymorphes de la grande population se retrouvent à l'état polymorphe dans la petite. Puis générations après générations, le nombre de polymorphismes partagés décroit rapidement, par dérive génétique, avant de disparaître complètement après quelques dizaines de milliers de générations (Figure 32).

Dans l'analyse du transcriptome d'*Astyanax mexicanus*, nous avons observé un niveau assez important de polymorphisme partagé entre les populations ($\approx$ 7%). Ce résultat pourrait donc être expliqué par une divergence récente de la population cavernicole et de la population de surface.

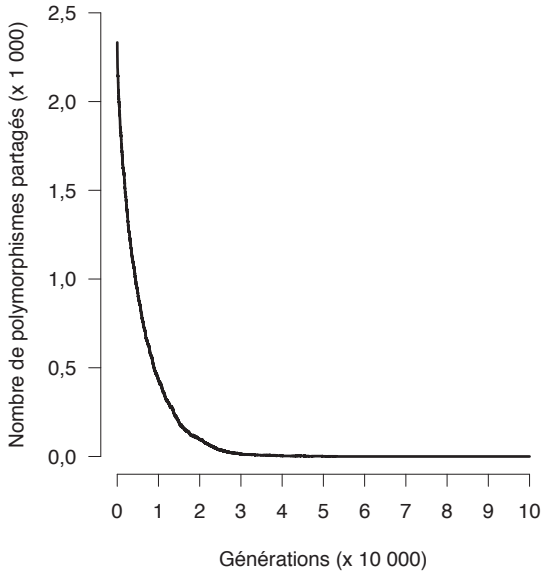


Figure 32. Nombre de sites polymorphes partagés entre populations au cours du temps lors de la simulation présentée Section 2.3.1.1.

Une autre hypothèse permettant d'expliquer ce fort taux de polymorphisme partagé entre les populations est des migrations entre les deux populations qui réhomogénéiseraient partiellement le polymorphisme. La population cavernicole de la grotte Pachón est une population relativement isolée des rivières de surface voisines [66], empêchant probablement les migrations fréquentes, bien qu'il ait été suggéré que des migrations récentes aient pu avoir lieu de la grotte vers la surface [121].

2.3.4 *Fixation d'allèles dérivés*

Dans la simulation, le nombre d'allèles dérivés fixés tend vers la même valeur dans les deux populations après une longue période de temps (1 000 000 de générations) (Figure 33, Figure 34 A), comme attendu.

Or, dans les données de transcriptomique, on observe qu'il y a environ deux fois plus d'allèles dérivés qui se sont fixés dans la population cavernicole que dans la population de surface (Tableau 3). Ce résultat est surprenant car la fixation des allèles dérivés se fait à la même vitesse dans deux populations de taille différente, si le taux de mutation dans chaque population est identique. Aussi après une longue période de temps, les populations cavernicoles d'*Astyanax mexicanus* étant considérées comme anciennes (environ 1 million d'années), le même nombre d'allèles fixés devrait être retrouvé dans les deux populations, ce qui n'est pas le cas.

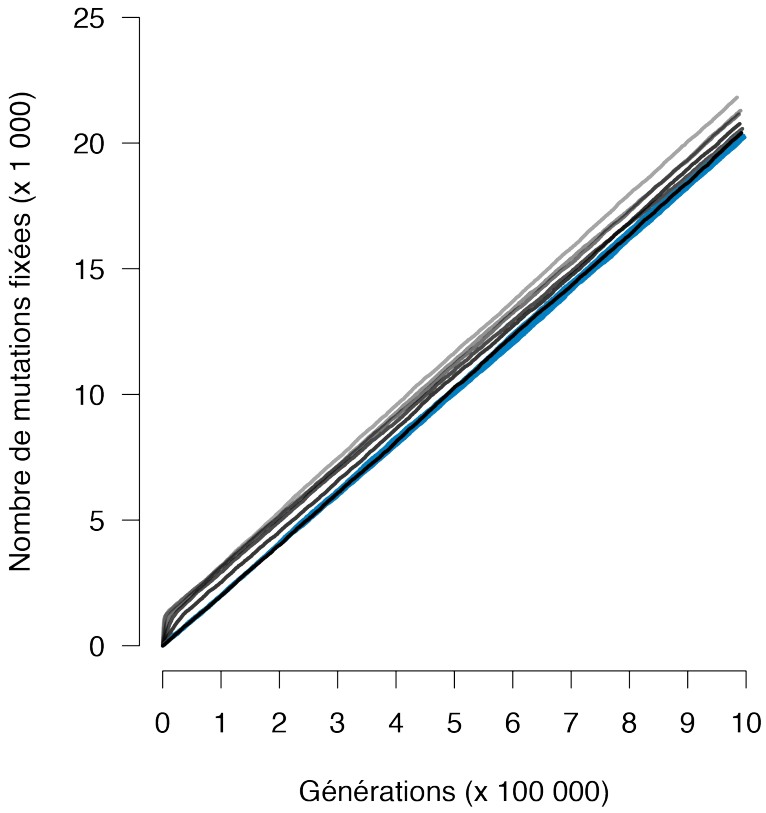


Figure 33. Nombre de sites ayant un allèle dérivé fixé dans une des deux populations lors de la simulation présentée Section 2.3.1.1 en utilisant différents nombres efficaces d'individus pour la plus petite des populations (en noir) et une taille efficace de $N_e = 10\,000$ pour la grande population (en bleu).

Si on s'intéresse maintenant aux 100 000 premières générations de la simulation on peut observer que, pour les mutations préexistantes à la séparation des populations (Figure 34 B), le même nombre d'allèles dérivés se fixe mais que la fixation est plus rapide dans la plus petite des deux populations. Les mutations apparues après la séparation des deux populations se fixent à la même vitesse dans les deux populations (Figure 34 C). Le temps moyen de fixation d'un nouvel allèle neutre dépend uniquement de la taille de la population, mais la fixation des allèles dérivés commence plus tôt dans la plus petite des deux populations. En additionnant les mutations préexistantes et les mutations apparues après la séparation des populations, on observe que les allèles se fixent plus rapidement dans la petite population pendant quelques milliers de générations puis par la suite à la même vitesse dans les deux populations (Figure 34 D).

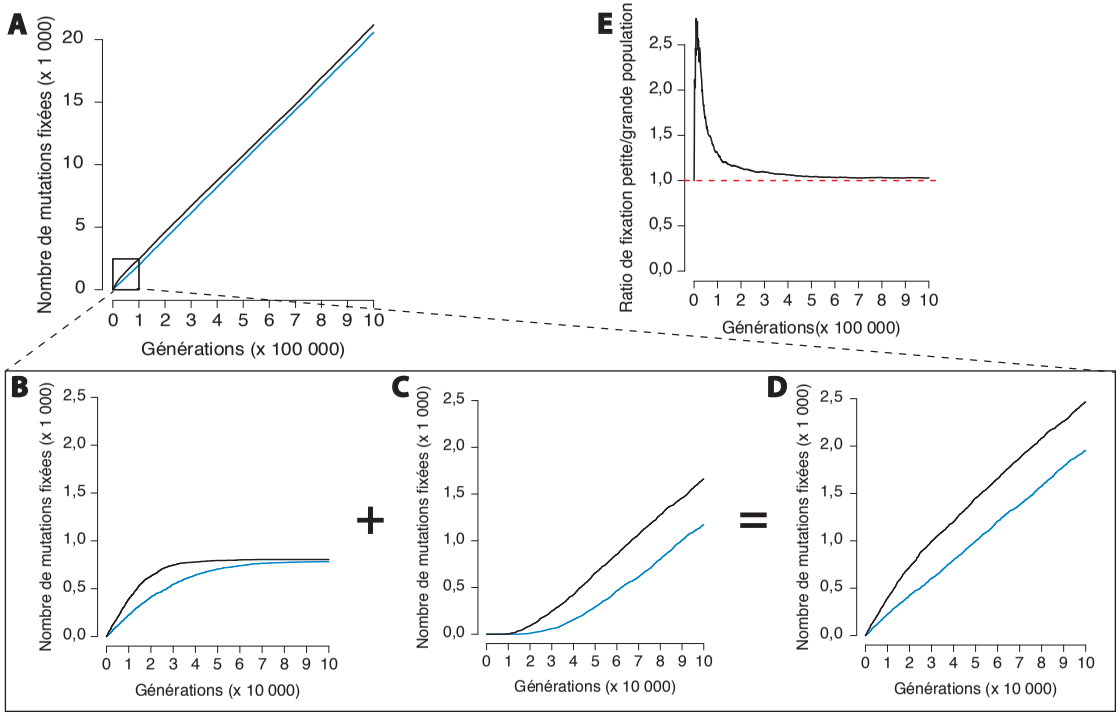


Figure 34. Fixation des allèles dérivés au cours du temps lors de la simulation présentée Section 2.3.1.1. (A) Nombre d'allèles dérivés fixés dans la grande population (en noir) et dans la petite population (en bleu) (1 000 000 générations). Nombre d'allèles dérivés fixés (B) apparus avant la séparation des populations, (C) après la séparation des populations, (D) total au cours des 100 000 premières générations. (E) Ratio de fixation petite/grande population au cours du temps.

On peut regarder le ratio de fixation entre la population de petite taille et celle de grande taille (Figure 34 E). Après un grand nombre de générations, le ratio de fixation tend, comme attendu, vers 1. Par contre, quelque temps après la séparation des populations, ce ratio est proche de 3. Cela signifie que dans la petite population il y a trois fois plus d'allèles dérivés qui sont fixés que dans la grande. Ainsi après séparation de deux populations de tailles différentes, on peut observer, de façon transitoire, dans la plus petite des deux populations une fixation plus importante d'allèles dérivés que dans la plus grande.

Aussi, l'observation de deux fois plus d'allèles dérivés fixés dans la population cavernicole de Pachón que dans la population de surface pourrait s'expliquer très simplement par une séparation récente de ces deux populations. Le test de cette hypothèse sera l'objet du Chapitre 3.

3 ÂGE DE LA POPULATION PACHÓN

3.1 Introduction

Dans le chapitre précédent, nous avons vu que la population de surface présente environ trois fois plus de polymorphismes que la population cavernicole de Pachón. Une hypothèse permettant d'expliquer ce résultat est la différence de taille entre ces deux populations. Une part importante du polymorphisme observé (environ 7%) est partagé par les deux populations. Ce résultat peut s'expliquer soit par des migrations entre populations soit par une divergence récente des deux populations. Enfin, nous avons observé que deux fois plus d'allèles dérivés étaient fixés dans la population cavernicole que dans la population de surface. L'hypothèse permettant d'expliquer simplement ce résultat est une récente divergence des populations.

L'estimation de l'âge de cette divergence est l'objet de ce chapitre. Pour la tester nous avons établis un modèle d'évolution des populations de surface et cavernicoles et nous avons ensuite simulé l'évolution du polymorphisme des populations dans le cadre de ce modèle.

3.2 Modélisation

3.2.1 *Différence entre modélisation prospective et rétrospective*

Les processus étudiés par la génétique des populations peuvent être modélisés selon deux types d'approches différentes : les approches prospectives et les approches rétrospectives.

Les approches rétrospectives, aussi appelées *backward* ou coalescence, existent depuis les années 1980 [122]. Dans ce type d'approches, la modélisation part d'un état final puis remonte le temps jusqu'à arriver à un ancêtre commun unique, appelé plus récent ancêtre commun (MRCA, *Most Recent Common Ancestor*). Le gros avantage de cette méthode est de ne simuler que les allèles observés dans la population finale et pas ceux qui ont disparu au cours de l'évolution. Cela permet des temps de calcul courts. Les approches prospectives (simulations *forward*), contrairement aux méthodes rétrospectives, nécessitent beaucoup de ressources informatiques. En effet, avec ce type d'approche, un état initial est généré. Cet état initial va ensuite évoluer, au cours du temps, sous l'effet des forces évolutives (mutations, dérive, sélection et migration). La simulation s'arrête lorsqu'un critère est rempli, par exemple après un nombre de générations fixé à l'avance. L'ensemble de l'information est donc conservé à chaque gé-

nération contrairement aux méthodes rétrospectives dans lesquelles seule l'information permettant d'obtenir l'état final est disponible.

Pour ces deux types d'approches, prospectives et rétrospectives, de nombreux programmes sont disponibles, chacun ayant ses spécificités. Une revue des différents programmes est disponible dans [123].

Nous avons choisi, ici, d'utiliser une méthode prospective puisque nous voulions voir l'évolution du rapport des nombres de substitutions dans les populations au cours du temps avant que ne s'établisse un équilibre mutation/dérive/migration. Nous avons également choisi d'implémenter nous-mêmes notre modèle démographique au lieu d'utiliser des logiciels existants. En effet l'utilisation de ces logiciels ne permettait pas d'estimer aisément des statistiques utiles à notre analyse.

3.2.2 *Modèle démographique implémenté*

Nous avons cherché à reproduire les données issues du séquençage en simulant des données de polymorphisme synonyme. La seule différence est la catégorie 8 (polymorphisme divergent, Tableau 1) qui n'est pas simulée puisque dans notre modèle toute nouvelle mutation apparait à un nouveau locus et il n'est donc pas possible qu'il y ait de polymorphisme divergent. Le modèle démographique utilisé est présenté Figure 35.

L'état initial est créé grâce à une population de surface ancestrale. Cette population ancestrale va ensuite donner deux populations de surface, la population du Texas que l'on étudie et une population de la Sierra de El Abra, plus proche, géographiquement, de la population cavernicole étudiée. C'est à partir de cette deuxième population que la population cavernicole est créée. Il peut y avoir un goulot d'étranglement au moment de la formation de la population cavernicole : la taille de la population est réduite à sa création avant de réaugmenter après quelques générations. Les trois populations vont alors évoluer pendant un temps plus ou moins grand. Enfin, quelques générations de dérive génétique sont réalisées afin de prendre en compte le temps passé au laboratoire des poissons échantillonnés dans la nature.

Des migrations entre les deux populations de surface et entre populations de surface et la population cavernicole sont possibles dans les deux sens.

Un grand nombre de paramètres peuvent être modifiés comme la taille des différentes populations, le taux de mutation, la probabilité de migration et le pourcentage de migrants et le temps de génération.

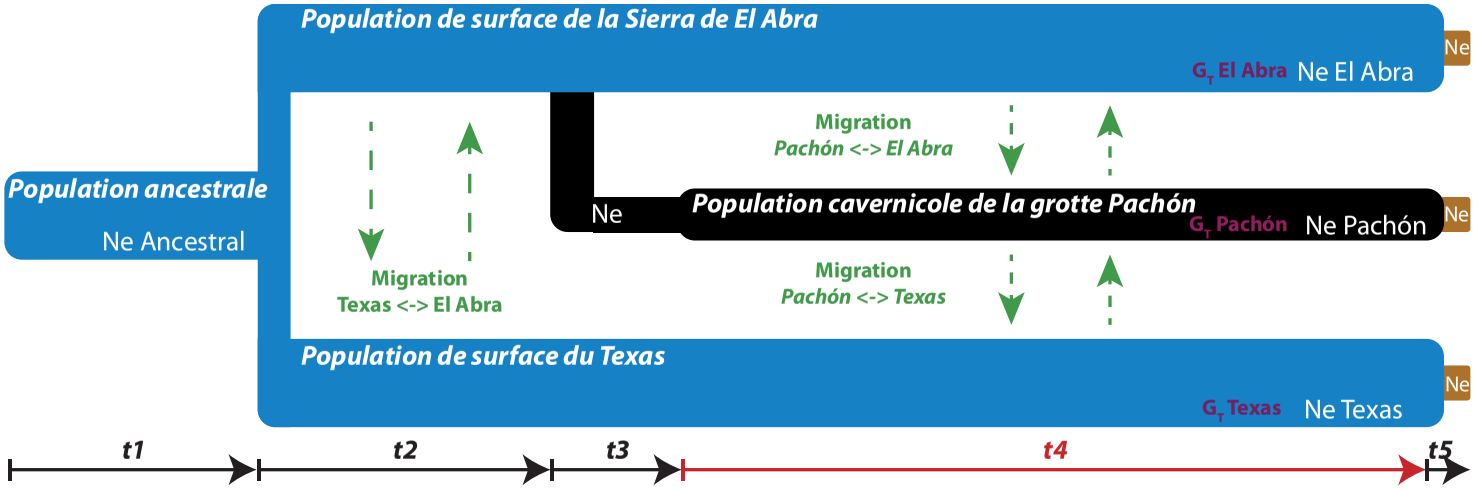


Figure 35. Modèle démographique implémenté. La modélisation commence par l'initialisation de la population ancestrale. Cette population donne deux populations de surface : celle du Texas, que nous étudions, et celle de la Sierra de El Abra qui est géographiquement la plus proche de la population cavernicole. Ces deux populations vont évoluer pendant un temps t_2. Après un temps t_2, une population cavernicole est échantillonnée dans la population de la Sierra de El Abra. Pendant une durée t_3 cette population cavernicole peut subir un goulot d'étranglement. On laisse ensuite évoluer les populations pendant une durée t_4 qui est le temps passé dans les grottes, c'est-à-dire le paramètre que nous intéresse. Nous avons tenu compte du temps passé au laboratoire des populations échantillonnées en simulant quelques générations de dérive génétique pendant un temps t_5. Des migrations entre les différentes populations peuvent avoir lieu, entre les deux populations de surface et entre les populations de surface et la population cavernicole. N_e *: taille efficace de la population,* G_T *: temps de génération*

3.2.2.1 *Mutations préexistantes*

La première étape des simulations est de générer l'état initial, c'est-à-dire un ensemble de sites polymorphes dans une population initiale permettant de simuler le polymorphisme existant chez l'ancêtre des populations de surface et cavernicoles. Nous faisons l'hypothèse que cette population est à l'équilibre mutation/dérive.

Ce polymorphisme peut-être généré de deux façons différentes. La première consiste à générer des allèles avec des fréquences aléatoires comprises entre 0 et 1 (0 et 1 non inclus, car un allèle à la fréquence 0 ou 1 est respectivement perdu ou fixé et non polymorphe). On laisse ensuite évoluer cet ensemble d'allèles sur un grand nombre de générations (étape de *burning*) en utilisant uniquement de la dérive génétique et des mutations. La deuxième méthode consiste à tirer des allèles aléatoirement dans une distribution théorique du nombre de sites possédant n allèles dérivés. Le nombre d'allèles dérivés à un site est compris entre 1 et $2N-1$, car si il n'y a pas d'allèle dérivé (0) alors l'allèle ancestral est fixé et si il n'y a que des allèles dérivés ($2N$) alors l'allèle dérivé est fixé.

La fréquence des loci possédant n copies de l'allèle est alors donnée par la formule [124] :

$$f_n = \frac{1}{n \times a_n} \tag{1}$$

avec

$$a_n = \sum_{i=1}^{2N-1} \frac{1}{i} \tag{2}$$

où N représente la taille de la population.

Par exemple, pour une population de taille 10 000, $a_n = \sum_{i=1}^{19\,999} \frac{1}{i} \approx 10,5$. On peut alors calculer la fréquence des sites ayant :

- 1 allèle dérivé parmi 20 000 : $\frac{1}{1 \times a_n} \approx 0,095$
- 2 allèles dérivés parmi 20 000 : $\frac{1}{2 \times a_n} \approx 0,047$
- 3 allèles dérivés parmi 20 000 : $\frac{1}{3 \times a_n} \approx 0,032$
- 100 allèles dérivés parmi 20 000 : $\frac{1}{100 \times a_n} \approx 9,5.10^{-4}$
- 19 999 allèles dérivés parmi 20 000 : $\frac{1}{19\,999 \times a_n} \approx 4,7.10^{-6}$.

Une fois ces fréquences calculées, on peut tirer aléatoirement des loci dans cette distribution.

Le nombre de loci polymorphes à générer est défini par la formule :

$$4 \times N \times \mu \times a_n \tag{3}$$

Le taux de mutation par génération (μ) utilisé dans les simulations a été calculé en fonction du nombre de positions n_{pos} produisant des mutations synonymes dans le génome d'*Astyanax mexicanus* et du taux de mutation par position et par génération (u) :

$$\mu = u \times n_{positions} \tag{4}$$

Dans nos séquences de transcriptome d'*Astyanax mexicanus*, nous avons identifié environ 2 millions de positions synonymes (voir Section 2.2.3). Le taux de mutation par position et par génération est, chez les eucaryotes, d'environ 10^{-8} dans les séquences nucléaires [125].

Le taux de mutation par génération est donc de $\mu = 2.10^{6} \times 10^{-8} = 2.10^{-2}$.

Pour un taux de mutation de 2.10^{-2} et une population de taille efficace 10 000, la simulation est donc initialisée avec 8 000 loci tirés au hasard dans la distribution calculée précédemment.

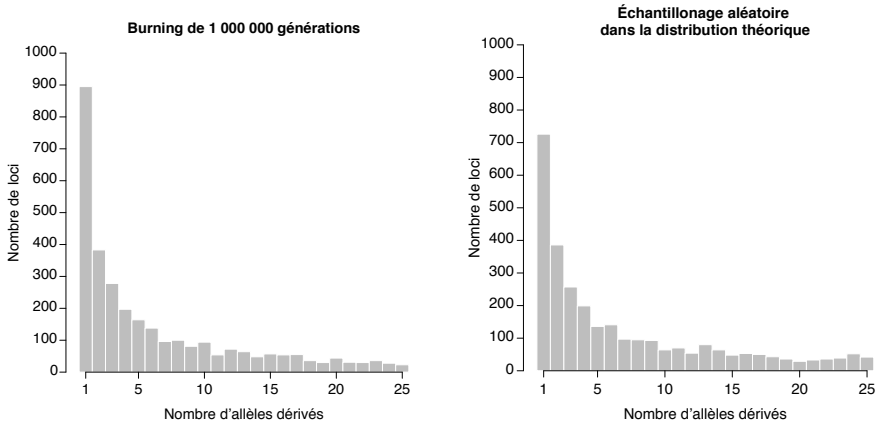


Figure 36. Comparaison de la distribution initiale des simulations en réalisant une étape de burning de 1 000 000 de générations (à gauche) et en utilisant l'échantillonnage aléatoire dans une distribution théorique calculée (à droite).

Les deux méthodes permettent d'obtenir un état initial similaire (Figure 36), mais l'échantillonnage dans une distribution théorique étant bien plus rapide, c'est cette méthode qui a été utilisée par la suite.

3.2.2.2 *Nouvelles mutations*

Tout au long de la simulation de nouveaux loci polymorphes vont apparaître par mutation. Le nombre de nouveaux loci polymorphes à chaque génération est calculé avec la formule suivante : $2 \times N_e \times \mu$ où N_e représente la taille efficace de la population et μ le taux de mutation. Ainsi, dans une population dont la taille efficace est de 10 000 avec un taux de mutation de 2.10^{-2}, 400 nouveaux loci polymorphes apparaitront à chaque génération. Le nouvel allèle à chacun de ces loci est à la fréquence $\frac{1}{2 \times 10\ 000} = 5.10^{-5}$.

Dans une autre population, par exemple de taille efficace 313, 13 nouveaux loci polymorphes apparaîtront à chaque génération et le nouvel allèle de chacun de ces loci sera à la fréquence $1{,}6.10^{-3}$.

3.2.2.3 *Migrations*

Les migrations entre deux populations dépendent de deux paramètres : la probabilité de migration entre ces deux populations et le nombre de poissons migrants.

Dans la simulation, à chaque génération, les migrations ont lieu de façon aléatoire et indépendante entre populations en tenant compte de la probabilité de migration. Si une migration a lieu d'une population vers une autre, la fréquence de l'ensemble des allèles est recalculée dans la population « receveuse ».

Considérons une migration depuis la population 2 vers la population 1. La fréquence des allèles dans la population 1 après migration est donnée par la formule :

$$f'_1 = \frac{(2 \times m_{2\to1} \times N_{e2} \times f_2) + (2 \times N_{e1} \times f_1)}{(2 \times m_{2\to1} \times N_{e2}) + (2 \times N_{e1})} \tag{5}$$

où f'_1 est la fréquence de l'allèle dans la population 1 après migration, $m_{2\to1} \times N_{e2}$ le nombre d'individus migrant de la population 2 vers la population 1, f_2 la fréquence de l'allèle dans la population 2, N_{e2} la taille efficace dans la population 2, f_1 la fréquence de l'allèle dans la population 1 avant migration et N_{e1} la taille efficace dans la population 1.

Prenons pour exemple un locus fixé dans une population de taille 10 000 et à la fréquence 0,2 dans une population de taille 1 000. Si 1% des poissons de la première population migrent vers la deuxième population, la fréquence dans cette population devient :

$$f'_1 = \frac{(2 \times 0{,}01 \times 10\ 000 \times 1) + (2 \times 1\ 000 \times 0{,}2)}{(2 \times 0{,}01 \times 10\ 000) + (2 \times 1\ 000)}$$
$$f'_1 = 0{,}27$$

3.2.2.4 *Temps de génération*

Le temps de génération est l'âge moyen des parents à la naissance de leurs descendants. Une différence de temps de génération entre les populations de surface et les populations cavernicoles va influencer la vitesse d'évolution des populations si celles-ci sont isolées. En effet, si dans une population il y a deux fois moins de générations que dans une autre et si le taux de mutation dépend du temps de génération, il y aura également deux fois moins de mutations qui apparaîtront pendant une période de temps donnée. Il a été suggéré que les animaux cavernicoles ont un temps de génération plus long que les animaux vivant en surface [126]. Par exemple, les olms (*Anguinus proteus*) peuvent vivre en moyenne près de 70 ans (avec un maximum prédit de 102 ans) et ont un temps de génération d'environ 40 ans [127].

En comparaison une espèce de protéidé épigée proche, *Necturus maculosus*, a une longévité d'environ 30 ans.

Aucune publication ne fait état du temps de génération chez les *Astyanax mexicanus* de surface mais il a été estimé à un an chez d'autres espèces du genre *Astyanax* [128].

Pour la population cavernicole de Pachón, le temps de génération a été estimé à 5 ans par Şadoğlu (non publiée mais reportée comme communication personnelle par Chakraborty et Nei [17]). Cette estimation ne repose sur aucunes données d'observations connues, mais il est possible que ses poissons puissent vivre une dizaine d'années, ce qui correspond à ce temps de génération s'ils commencent à se reproduire à l'âge de 1 an.

Puisque nous cherchons à tester l'hypothèse d'une origine récente de la population Pachón et que nous préférons donc surestimer que sous-estimer l'âge de cette population, nous avons utilisé un temps de génération pour la population de surface de deux ans, le double de la valeur estimée chez les autres espèces d'*Astyanax*. Lorsque le temps de génération est augmenté, la population évolue moins vite puisqu'il y a moins de générations produites pendant une même durée de temps. Le temps de divergence, en années, sera alors augmenté.

Nous avons également testé des temps de génération identiques (2 ans) dans les deux populations et plus grand dans la population de surface (5 ans) que dans la population cavernicole (2 ans). Bien que très peu probable, nous avons ainsi considéré un taux de mutation plus grand chez les poissons cavernicoles, ce qui pourrait, à première vue, être une explication simple à un plus grand nombre d'allèles dérivés fixés dans cette population.

3.2.2.5 *Dérive génétique au laboratoire*

Les poissons séquencés, auxquels nous comparerons les résultats des simulations, sont des animaux d'élevage. Nous avons considéré le nombre de générations dans l'élevage entre 5 et 10 pour chaque population et la taille efficace des populations à 10.

La taille de la population est ainsi fortement réduite par rapport aux populations sauvages et depuis un temps relativement important pour une population aussi petite. Ainsi une dérive génétique importante a pu avoir lieu dans l'élevage. Il faut donc tenir compte de cette dérive. Aussi, avant d'écrire les statistiques sur le nombre de SNP dans chacune des 8 catégories définies au Chapitre 2, nous avons ajouté une étape de dérive génétique pendant 0, 5 ou 10 générations avec 10 individus par population.

3.2.2.6 *Score d'ajustement de la simulation aux données*

Toutes les 10 générations, les pourcentages de SNP simulés dans chacune des 8 catégories sont comparés aux résultats de l'étude transcriptomique

présentée dans le chapitre précédent. Pour cela, une distance de χ^2 est calculée avec la formule suivante :

$$d_{simulation-observation} = \sum_{i=1}^{8} \frac{(Observation_i - Simulation_i)^2}{Observation_i} \quad (6)$$

où i représente une des 8 catégories de SNP.

Nous cherchons à obtenir des données simulées les plus proches possibles des données observées. Aussi, plus le score est proche de zéro, meilleure est la ressemblance entre données simulées et données réelles.

Pour chaque simulation nous avons calculé ce score toutes les 10 générations puis nous avons cherché le score minimum et après combien de générations dans la grotte ce score est obtenu.

3.2.3 *Modèles simplifiés utilisés*

Au vu du grand nombre de paramètres possibles dans le modèle démographique implémenté (Figure 35), nous avons choisi d'utiliser deux modèles simplifiés en jouant sur les paramètres des simulations (Figure 37). Nous parlerons de modèle complet et de modèles simplifiés A et B.

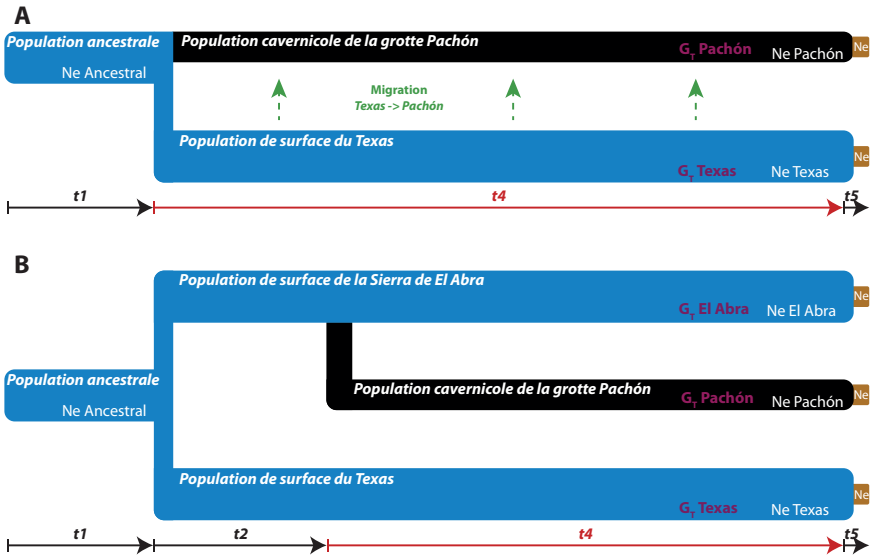


Figure 37. Modèles démographiques simplifiés. (A) Dans ce premier modèle simplifié, la population de surface de la Sierra de El Abra n'est pas simulée. La population cavernicole est échantillonnée directement dans la population ancestrale, et il n'y a pas de goulot d'étranglement à sa création. Les seules migrations possibles sont depuis la population de surface du Texas et la population de la grotte Pachón. (B) Dans ce second modèle simplifié, les mutations sont ignorées. Les deux populations de surface sont simulées et la population cavernicole est échantillonnée dans la population de la Sierra de El Abra après un nombre de générations, sans possibilité de goulot d'étranglement.

3.3 Résultats

Le modèle démographique complet a été implémenté dans un programme écrit en C [129]. Le programme est disponible en ligne sur Github (https://github.com/julienfumey/popsim).

Nous avons utilisé les modèles simplifiés A et B et pour chacun d'entre eux nous avons fait varier différents paramètres.

PARAMÈTRES COMMUNS AUX DEUX MODÈLES SIMPLIFIÉS

- Taille des populations de surface (N_e Texas, N_e El Abra) : 5 000, 10 000
- Taille de la population cavernicole (N_e Pachón) : 313, 625, 1 250, 2 500, 5 000, 10 000
- Taux de mutation : 2.10^{-3} par génome et par an
- Temps de génération (G_T) : 2 ans, 5 ans
- Nombre de générations au laboratoire : 0, 5, 10

PARAMÈTRES DU MODÈLE SIMPLIFIÉ A

- Probabilité de migration par année depuis la population de surface vers la population cavernicole : 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}
- Pourcentage de poissons de surface migrants "vers la population cavernicole : 10%, 1%, 0,1%, 0,01%

PARAMÈTRES DU MODÈLE SIMPLIFIÉ B

- Nombre d'années en surface avant apparition de la population cavernicole : 0, 10 000, 20 000, 40 000, 80 000

Lorsque le nombre de générations en surface avant apparition de la population cavernicole était différent de 0, aucune migration n'avait lieu entre population de surface et population cavernicole.

Les résultats obtenus avec le modèle simplifié A sont présentés Tableau 9 à 18. Ceux avec le modèle simplifié B sont présentés Tableau 19 à 23. Pour les scores, les cellules ont été colorées avec trois couleurs : vert pour des scores inférieurs à trois, orange pour des scores compris entre trois et quinze, et rouge pour des score supérieurs à quinze.

Lorsque la taille efficace des populations cavernicoles est supérieure à 5 000, le score minimal des simulations est systématiquement supérieur à 15. Avec ces tailles de population cavernicoles, il est impossible d'obtenir un bon ajustement entre données simulées et données observées. Ceci montre que ces simulations ne permettent pas de reproduire les données observées.

Par contre, pour des populations cavernicoles plus petites, il est possible d'observer de bons ajustements (score minimal de la simulation < 3, cellules

N_e CF	%M→	N_e SF 5 000							
		10%		1%		0,1%		0,01%	
	m↓	Score	Age	Score	Age	Score	Age	Score	Age
313	10^{-2}	204,75	7 200	132,69	64 600	21,29	98 600	10,64	5 200
	10^{-3}	11,82	87 500	3,08	83 700	2,96	27 300	10,91	5 000
	10^{-4}	5,06	48 800	3,24	36 100	9,03	10 900	9,73	6 600
	10^{-5}	2,20	42 000	6,57	9 400	9,04	6 300	12,22	5 200
625	10^{-2}	178,55	22 000	166,94	26 500	34,04	95 000	7,20	13 700
	10^{-3}	15,15	51 800	13,67	29 500	2,60	16 000	6,05	11 400
	10^{-4}	2,81	30 700	3,31	27 400	5,50	10 700	6,24	10 300
	10^{-5}	5,41	12 400	5,72	13 000	4,63	11 300	5,10	13 000
1 250	10^{-2}	229,99	62 700	159,12	47 600	43,57	29 500	8,24	20 700
	10^{-3}	58,13	93 200	27,67	16 700	9,86	18 500	7,30	21 000
	10^{-4}	9,73	22 200	10,88	20 900	7,03	20 700	9,91	20 300
	10^{-5}	8,22	17 300	12,28	18 600	11,11	20 600	7,05	17 600
2 500	10^{-2}	230,52	76 800	148,06	39 300	62,86	15 900	34,58	30 800
	10^{-3}	101,22	18 300	56,03	24 000	34,88	31 300	31,45	24 700
	10^{-4}	40,71	32 000	37,28	32 700	29,09	28 900	31,63	28 500
	10^{-5}	29,44	24 300	31,62	33 100	33,25	23 200	36,61	35 100
5 000	10^{-2}	218,15	31 300	138,45	26 000	88,33	18 000	68,80	15 000
	10^{-3}	108,85	36 000	85,48	16 300	68,07	27 700	63,65	21 300
	10^{-4}	78,70	13 100	69,51	17 200	73,70	51 600	68,12	40 500
	10^{-5}	74,86	27 200	68,11	28 600	74,66	53 300	72,58	23 600
10 000	10^{-2}	202,42	64 800	129,18	9 900	101,86	13 900	105,01	13 700
	10^{-3}	130,67	15 800	104,67	11 200	105,19	15 900	98,76	13 700
	10^{-4}	113,76	14 200	96,57	11 400	105,55	14 300	112,28	19 200
	10^{-5}	97,15	14 000	94,99	14 100	95,76	16 700	101,65	13 500

Tableau 9. Score minimum entre la population de surface et la population cavernicole pour différentes tailles de population cavernicole et différents taux de migration et temps de divergence lorsque ce score est atteint. Modèle démographique simplifié A. N_e en surface : 5 000. Simulations réalisées avec 5 générations de laboratoire, un temps de génération de 5 ans dans la population cavernicole et de 2 ans dans la population de surface.
%M : pourcentage de poissons migrants de la population de surface vers la population cavernicole, m : probabilité qu'un événement de migration ait lieu par an.
Vert : score inférieur à 3, orange : score compris entre 3 et 15, rouge : score supérieur ou égal à 15.

vertes) pour certains taux de migrations, lorsque le temps de génération est égal dans les deux populations ou lorsque le temps de génération est plus grand dans la population cavernicole que dans la population de surface. Ainsi, et en première approximation, nous voyons de bons ajustements de la

N_e CF	%M→	N_e SF 10 000							
		10%		1%		0,1%		0,01%	
	m↓	Score	Age	Score	Age	Score	Age	Score	Age
313	10^{-2}	201,73	70 300	176,26	60 300	65,39	92 100	21,30	33 800
	10^{-3}	19,10	93 700	21,12	80 900	1,50	70 100	21,26	5 000
	10^{-4}	8,80	99 700	2,20	56 300	6,12	38 300	21,24	32 700
	10^{-5}	7,87	47 700	20,92	35 200	10,30	21 900	22,24	27 600
625	10^{-2}	240,12	74 500	213,30	82 800	86,47	40 000	13,59	10 900
	10^{-3}	51,82	16 400	26,75	86 900	1,41	51 500	14,11	10 400
	10^{-4}	10,25	98 500	3,51	33 700	7,62	57 000	14,22	11 600
	10^{-5}	2,54	60 400	6,05	72 700	10,94	28 900	13,79	10 500
1 250	10^{-2}	254,24	32 500	215,22	91 800	75,19	52 300	4,70	25 800
	10^{-3}	77,39	27 000	31,53	61 900	3,35	56 100	5,84	22 600
	10^{-4}	4,77	50 700	3,38	31 400	3,08	30 500	5,34	25 800
	10^{-5}	4,96	49 800	6,08	21 600	5,85	22 700	6,70	24 300
2 500	10^{-2}	253,67	4 800	210,92	30 000	78,83	23 400	9,70	40 400
	10^{-3}	130,92	14 900	59,74	20 800	12,66	62 300	9,03	37 200
	10^{-4}	16,92	43 100	12,09	55 700	10,04	41 500	9,01	41 300
	10^{-5}	9,55	40 300	10,13	35 600	9,96	41 100	10,17	42 100
5 000	10^{-2}	265,28	57 900	187,78	38 700	89,65	28 800	32,98	61 400
	10^{-3}	137,38	28 300	70,75	24 300	37,57	58 800	33,49	52 100
	10^{-4}	43,81	64 400	36,48	70 400	33,28	60 500	34,99	65 800
	10^{-5}	34,23	61 000	33,30	69 800	37,21	60 200	32,82	50 400
10 000	10^{-2}	250,46	24 900	167,59	79 500	98,31	30 700	76,24	59 100
	10^{-3}	134,20	52 400	94,81	33 300	70,00	53 500	74,34	71 400
	10^{-4}	92,26	22 600	81,31	56 000	74,44	51 300	75,57	84 200
	10^{-5}	81,40	22 800	73,24	66 500	71,39	47 500	75,49	43 400

Tableau 10. Score minimum entre la population de surface et la population cavernicole pour différentes tailles de population cavernicole et différents taux de migration et temps de divergence lorsque ce score est atteint.
Modèle démographique simplifié A. N_e en surface : 10 000. Simulations réalisées avec 5 générations de laboratoire, un temps de génération de 5 ans dans la population cavernicole et de 2 ans dans la population de surface.
%M : pourcentage de poissons migrants de la population de surface vers la population cavernicole, m : probabilité qu'un événement de migration ait lieu par an.
Vert : score inférieur à 3, orange : score compris entre 3 et 15, rouge : score supérieur ou égal à 15.

simulation aux données observées que lorsque les jeux de paramètres utilisés doivent correspondre à la réalité démographique de ces populations.

Lorsque la taille de la population de surface est divisée par deux, les tailles de populations cavernicoles pour lesquels des bons scores sont observés sont

N_e CF	%M→	N_e SF 5 000							
		10%		1%		0,1%		0,01%	
	m↓	Score	Age	Score	Age	Score	Age	Score	Age
313	10^{-2}	67,73	39 100	57,75	74 700	11,96	22 300	22,72	2 500
	10^{-3}	20,67	74 500	13,28	36 900	8,87	16 500	23,91	15 500
	10^{-4}	7,27	97 200	4,73	20 500	14,17	13 200	22,98	14 800
	10^{-5}	6,52	68 300	23,81	19 800	23,71	17 700	24,16	20 700
625	10^{-2}	154,35	10 400	106,04	5 800	6,68	99 800	18,49	18 900
	10^{-3}	13,45	87 600	5,44	86 500	7,30	20 300	16,56	18 400
	10^{-4}	5,29	57 500	8,59	30 900	16,58	23 800	17,55	19 700
	10^{-5}	17,39	18 000	15,45	9 600	17,47	16 700	16,51	18 200
1 250	10^{-2}	212,58	41 000	107,22	40 400	16,68	45 400	14,28	18 700
	10^{-3}	22,64	48 100	9,59	77 000	12,00	19 900	14,40	14 000
	10^{-4}	10,57	30 100	9,61	26 400	13,54	21 500	13,95	19 300
	10^{-5}	8,96	23 800	9,29	22 900	14,30	17 900	11,81	14 400
2 500	10^{-2}	208,12	68 400	132,12	85 400	45,15	48 100	25,18	39 500
	10^{-3}	66,73	5 700	35,46	40 300	24,92	34 700	24,25	25 500
	10^{-4}	28,67	42 700	24,09	32 900	29,12	31 200	26,01	20 900
	10^{-5}	26,67	20 500	26,15	35 500	29,99	28 300	25,56	38 300
5 000	10^{-2}	226,55	78 200	143,58	20 600	90,12	34 300	64,60	63 500
	10^{-3}	106,54	22 300	83,19	42 200	68,81	91 100	68,49	80 600
	10^{-4}	90,32	58 400	68,43	78 100	68,58	71 000	72,85	64 400
	10^{-5}	70,47	66 900	70,16	81 400	63,51	85 500	65,60	64 500
10 000	10^{-2}	215,14	74 300	167,63	13 700	141,95	8 700	139,63	9 900
	10^{-3}	161,16	18 100	142,80	8 900	132,12	13 700	137,69	14 600
	10^{-4}	150,72	18 800	136,40	16 700	137,39	11 000	139,51	13 700
	10^{-5}	138,71	10 600	142,40	9 100	152,55	10 600	142,30	12 800

Tableau 11. Score minimum entre la population de surface et la population cavernicole pour différentes tailles de population cavernicole et différents taux de migration et temps de divergence lorsque ce score est atteint.
Modèle démographique simplifié A. N_e en surface : 5 000. Simulations réalisées avec 5 générations de laboratoire, un temps de génération de 2 ans dans la population cavernicole et de 5 ans dans la population de surface.
%M : pourcentage de poissons migrants de la population de surface vers la population cavernicole, m : probabilité qu'un événement de migration ait lieu par an.
Vert : score inférieur à 3, orange : score compris entre 3 et 15, rouge : score supérieur ou égal à 15.

également divisées par deux. Cela semble montrer que le point important n'est pas la taille des populations mais le ratio entre la taille de la population de surface et celle de la population cavernicole. Dans ce cas, l'âge de la population cavernicole est également divisé par deux.

N_e CF	%M→	N_e SF 10 000 10%		1%		0,1%		0,01%	
	m↓	Score	Age	Score	Age	Score	Age	Score	Age
313	10^{-2}	133,10	12 500	100,16	94 000	32,51	41 700	28,71	40 800
	10^{-3}	30,25	41 300	22,55	77 300	4,29	79 200	28,79	42 700
	10^{-4}	23,17	90 200	4,49	89 200	12,51	70 700	27,53	37 600
	10^{-5}	16,05	42 400	29,95	35 300	29,33	36 800	29,26	41 800
625	10^{-2}	190,31	20 000	144,75	34 600	26,43	50 700	24,65	41 700
	10^{-3}	27,58	89 600	18,40	42 700	4,39	72 500	24,07	36 300
	10^{-4}	17,26	37 200	4,19	80 800	12,34	56 400	24,12	38 000
	10^{-5}	22,46	38 500	27,18	34 600	17,39	39 100	23,88	27 700
1 250	10^{-2}	235,43	91 500	146,41	29 200	28,36	74 800	17,75	38 800
	10^{-3}	19,27	17 100	18,04	40 200	5,88	43 800	17,36	41 400
	10^{-4}	11,18	92 800	7,86	64 700	13,46	57 200	17,89	36 700
	10^{-5}	7,77	99 900	18,04	37 000	18,11	29 100	18,46	39 000
2 500	10^{-2}	257,59	63 800	186,97	92 600	45,80	21 600	14,10	45 500
	10^{-3}	54,80	30 400	20,01	83 500	10,45	43 500	13,83	37 600
	10^{-4}	11,93	82 500	10,82	53 300	13,45	51 400	14,93	35 900
	10^{-5}	15,41	44 400	14,36	43 400	14,04	37 000	16,14	44 800
5 000	10^{-2}	269,95	11 600	178,29	95 300	64,24	31 300	28,27	90 400
	10^{-3}	110,46	19 900	61,63	34 500	29,68	99 600	28,41	61 000
	10^{-4}	35,07	84 300	26,65	78 400	28,20	78 200	27,35	66 100
	10^{-5}	27,26	78 500	28,54	45 600	28,52	67 000	29,54	77 200
10 000	10^{-2}	258,59	76 100	182,13	74 800	111,72	54 900	76,70	98 800
	10^{-3}	141,13	42 800	110,06	30 300	84,65	94 100	73,85	99 100
	10^{-4}	88,65	28 500	79,71	98 400	80,14	99 600	78,20	95 000
	10^{-5}	83,39	98 600	79,91	80 300	80,74	99 600	78,07	97 600

Tableau 12. Score minimum entre la population de surface et la population cavernicole pour différentes tailles de population cavernicole et différents taux de migration et temps de divergence lorsque ce score est atteint.
Modèle démographique simplifié A. N_e en surface : 10 000. Simulations réalisées avec 5 générations de laboratoire, un temps de génération de 2 ans dans la population cavernicole et de 5 ans dans la population de surface.
%M : pourcentage de poissons migrants de la population de surface vers la population cavernicole, m : probabilité qu'un événement de migration ait lieu par an.
Vert : score inférieur à 3, orange : score compris entre 3 et 15, rouge : score supérieur ou égal à 15.

N_e CF	%M→	N_e SF 5 000 10%		1%		0,1%		0,01%	
	m↓	Score	Age	Score	Age	Score	Age	Score	Age
313	10^{-2}	100,36	6 600	70,47	17 300	8,30	76 700	4,86	5 600
	10^{-3}	4,57	79 300	0,55	77 400	1,70	13 100	4,13	4 600
	10^{-4}	0,52	68 500	0,48	12 800	3,70	4 600	4,25	5 300
	10^{-5}	2,29	70 100	1,94	16 500	3,36	4 800	4,46	4 200
625	10^{-2}	112,85	45 200	80,63	52 600	17,80	19 300	1,39	9 300
	10^{-3}	25,16	87 100	7,56	94 100	1,49	13 700	1,48	8 800
	10^{-4}	1,50	9 200	1,69	8 200	2,20	8 100	2,07	9 500
	10^{-5}	0,79	12 800	1,24	9 700	0,86	9 300	1,20	8 400
1 250	10^{-2}	128,03	99 900	92,96	44 300	27,82	28 400	7,15	13 600
	10^{-3}	51,49	38 800	25,04	17 000	5,97	13 500	5,19	14 800
	10^{-4}	7,09	13 700	7,18	17 600	6,42	13 400	10,92	15 200
	10^{-5}	6,61	14 500	7,48	13 800	7,43	15 800	5,19	15 700
2 500	10^{-2}	126,28	8 200	98,61	97 900	51,52	10 300	29,37	27 800
	10^{-3}	72,20	23 600	38,43	28 600	30,06	23 400	29,80	16 000
	10^{-4}	27,93	16 900	28,22	24 200	23,72	19 600	27,16	20 700
	10^{-5}	23,98	18 600	23,45	19 400	28,00	29 800	32,01	24 200
5 000	10^{-2}	128,74	89 900	96,73	10 100	64,33	19 100	55,83	23 800
	10^{-3}	83,35	13 400	69,75	17 400	61,35	28 800	55,91	19 300
	10^{-4}	57,55	13 700	55,09	18 500	53,50	16 100	57,10	21 800
	10^{-5}	55,30	18 300	53,40	23 300	59,40	25 600	56,15	20 300
10 000	10^{-2}	123,71	74 900	90,25	12 500	82,43	12 500	75,75	15 800
	10^{-3}	99,74	7 700	83,42	8 100	76,60	9 000	78,63	26 000
	10^{-4}	84,44	14 000	77,41	15 000	76,76	16 600	76,06	12 700
	10^{-5}	77,20	12 600	70,87	11 100	73,12	10 500	75,53	16 400

Tableau 13. Score minimum entre la population de surface et la population cavernicole pour différentes tailles de population cavernicole et différents taux de migration et temps de divergence lorsque ce score est atteint.
Modèle démographique simplifié A. N_e en surface : 5 000. Simulations réalisées avec 10 générations de laboratoire, un temps de génération de 5 ans dans la population cavernicole et de 2 ans dans la population de surface.
%M : pourcentage de poissons migrants de la population de surface vers la population cavernicole, m : probabilité qu'un événement de migration ait lieu par an.
Vert : score inférieur à 3, orange : score compris entre 3 et 15, rouge : score supérieur ou égal à 15.

N_e CF	%M→	N_e SF 10 000 10%		1%		0,1%		0,01%	
	m↓	Score	Age	Score	Age	Score	Age	Score	Age
313	10^{-2}	96,13	37 700	101,07	43 200	25,77	35 700	10,72	5 200
	10^{-3}	9,50	29 500	6,19	99 900	0,47	50 700	13,25	4 000
	10^{-4}	5,00	41 000	0,60	32 100	5,63	34 900	9,95	5 200
	10^{-5}	2,96	44 900	10,91	4 300	11,40	4 300	11,54	4 800
625	10^{-2}	136,91	92 200	117,95	69 400	41,28	97 200	6,27	10 900
	10^{-3}	34,58	62 500	9,60	57 200	0,15	25 500	4,56	10 400
	10^{-4}	1,13	27 200	0,71	43 100	4,93	9 500	4,43	9 900
	10^{-5}	0,77	48 100	0,59	32 400	4,82	9 000	5,37	8 200
1 250	10^{-2}	141,39	22 200	120,72	13 900	41,23	48 100	1,39	20 400
	10^{-3}	28,66	16 300	32,18	51 100	1,69	22 900	1,05	18 600
	10^{-4}	1,36	18 200	0,88	24 900	0,97	20 200	2,35	18 100
	10^{-5}	1,36	16 200	0,69	17 100	1,54	19 400	1,66	19 500
2 500	10^{-2}	144,44	10 900	122,25	37 300	58,67	33 800	7,30	31 500
	10^{-3}	79,99	76 600	33,24	26 800	9,44	36 000	7,93	32 300
	10^{-4}	17,89	46 000	10,45	43 800	5,98	29 600	7,90	27 000
	10^{-5}	10,67	23 200	8,23	29 300	7,50	32 700	8,33	29 600
5 000	10^{-2}	139,62	66 000	116,78	72 700	66,67	22 000	27,92	39 700
	10^{-3}	103,88	90 700	67,07	25 300	30,19	48 200	28,03	35 400
	10^{-4}	34,84	35 100	31,99	46 000	26,92	49 900	25,55	32 300
	10^{-5}	29,25	42 500	29,68	44 600	28,07	49 000	29,55	51 300
10 000	10^{-2}	140,55	79 300	115,35	99 000	75,54	17 100	58,89	54 200
	10^{-3}	107,01	8 900	73,94	18 900	60,08	38 200	55,02	40 100
	10^{-4}	69,14	25 600	65,41	54 800	60,50	42 400	56,66	46 600
	10^{-5}	60,21	24 100	53,83	42 800	58,79	39 700	55,80	50 000

Tableau 14. Score minimum entre la population de surface et la population cavernicole pour différentes tailles de population cavernicole et différents taux de migration et temps de divergence lorsque ce score est atteint.
Modèle démographique simplifié A. N_e en surface : 10 000. Simulations réalisées avec 10 générations de laboratoire, un temps de génération de 5 ans dans la population cavernicole et de 2 ans dans la population de surface.
%M : pourcentage de poissons migrants de la population de surface vers la population cavernicole, m : probabilité qu'un événement de migration ait lieu par an.
Vert : score inférieur à 3, orange : score compris entre 3 et 15, rouge : score supérieur ou égal à 15.

N_e CF	%M→	N_e SF 5 000 10%		1%		0,1%		0,01%	
	m↓	Score	Age	Score	Age	Score	Age	Score	Age
313	10^{-2}	34,05	180	16,43	17 500	0,38	7 400	14,02	360
	10^{-3}	5,30	12 760	0,37	7 600	3,83	2 520	10,44	380
	10^{-4}	1,08	15 120	12,80	360	10,42	380	10,51	380
	10^{-5}	11,53	400	13,39	360	13,41	460	9,40	460
625	10^{-2}	76,13	8 840	53,59	15 420	2,71	6 640	8,19	1 100
	10^{-3}	2,88	13 820	0,87	5 080	4,58	1 880	5,49	1 060
	10^{-4}	1,51	17 740	5,48	900	7,55	660	6,76	1 060
	10^{-5}	2,13	3 040	7,02	940	5,67	980	6,28	980
1 250	10^{-2}	109,30	19 280	61,75	15 260	8,68	3 160	3,47	1 640
	10^{-3}	13,55	15 100	6,03	3 100	4,04	2 600	6,55	1 880
	10^{-4}	4,51	1 740	5,99	2 040	5,97	1 580	4,61	1 780
	10^{-5}	6,42	1 780	3,87	2 060	4,33	1 880	7,79	2 100
2 500	10^{-2}	106,95	3 140	72,92	15 380	33,43	2 400	19,50	2 760
	10^{-3}	46,56	12 020	26,35	2 560	20,63	3 260	22,06	3 900
	10^{-4}	20,23	2 200	23,47	3 400	15,66	2 920	19,05	3 360
	10^{-5}	20,59	3 860	19,98	3 500	21,37	2 800	19,50	2 680
5 000	10^{-2}	121,41	8 840	92,43	1 240	61,24	2 640	53,74	3 040
	10^{-3}	90,22	3 480	59,51	3 140	52,57	2 720	52,99	2 820
	10^{-4}	61,08	2 920	52,62	1 800	49,93	3 880	56,38	5 280
	10^{-5}	59,15	4 260	49,50	2 840	60,77	5 300	55,38	5 220
10 000	10^{-2}	122,69	6 220	102,48	1 020	85,91	1 180	91,11	1 300
	10^{-3}	95,60	1 060	95,80	2 240	96,16	1 560	94,08	1 280
	10^{-4}	99,82	1 420	89,02	1 240	89,16	2 000	88,23	1 400
	10^{-5}	97,53	1 820	88,34	1 400	93,98	1 880	91,15	1 840

Tableau 15. Score minimum entre la population de surface et la population cavernicole pour différentes tailles de population cavernicole et différents taux de migration et temps de divergence lorsque ce score est atteint.
Modèle démographique simplifié A. N_e en surface : 5 000. Simulations réalisées avec 10 générations de laboratoire, un temps de génération de 2 ans dans la population cavernicole et de 2 ans dans la population de surface.
%M : pourcentage de poissons migrants de la population de surface vers la population cavernicole, m : probabilité qu'un événement de migration ait lieu par an.
Vert : score inférieur à 3, orange : score compris entre 3 et 15, rouge : score supérieur ou égal à 15.

N_e CF	%M→	N_e SF 10 000 10%		1%		0,1%		0,01%	
	m↓	Score	Age	Score	Age	Score	Age	Score	Age
313	10^{-2}	44,92	640	39,97	11 760	9,30	4 860	15,86	360
	10^{-3}	11,08	2 920	5,84	18 540	0,11	6 380	14,40	320
	10^{-4}	6,49	17 360	0,75	13 080	17,23	2 840	14,66	380
	10^{-5}	16,04	340	12,42	340	18,54	300	17,37	280
625	10^{-2}	90,23	8 180	70,04	5 180	11,11	8 460	11,50	900
	10^{-3}	7,16	3 200	1,13	17 640	2,59	4 560	9,50	760
	10^{-4}	1,11	19 460	2,69	2 620	11,54	860	12,15	640
	10^{-5}	13,29	780	12,33	680	12,34	700	12,25	820
1 250	10^{-2}	126,61	18 080	93,82	6 100	12,32	9 380	5,26	1 580
	10^{-3}	5,48	15 080	2,94	9 100	2,98	4 920	3,81	1 720
	10^{-4}	1,52	11 500	2,35	4 560	4,07	1 880	4,81	1 720
	10^{-5}	5,49	1 700	1,44	5 420	5,84	1 620	5,12	1 980
2 500	10^{-2}	139,11	160	100,76	840	21,86	4 180	4,11	3 760
	10^{-3}	50,60	8 820	10,76	5 080	5,88	3 700	5,14	2 920
	10^{-4}	7,69	6 200	4,20	3 500	6,10	3 620	3,76	3 080
	10^{-5}	6,07	3 600	4,37	3 700	5,41	3 640	4,74	3 700
5 000	10^{-2}	136,82	7 580	101,36	6 760	41,34	5 120	21,18	4 440
	10^{-3}	60,27	16 920	44,85	8 960	20,90	5 420	21,88	5 400
	10^{-4}	22,89	3 760	22,35	5 260	21,41	5 980	22,49	5 440
	10^{-5}	18,86	6 220	19,67	6 300	20,01	6 620	19,73	5 580
10 000	10^{-2}	142,21	16 620	106,28	2 020	71,34	5 340	59,15	5 240
	10^{-3}	94,91	2 500	68,95	3 400	56,07	5 280	56,66	11 280
	10^{-4}	70,40	2 940	61,20	9 400	58,40	12 260	63,00	6 500
	10^{-5}	62,96	10 700	62,97	7 460	57,73	5 220	57,75	5 440

Tableau 16. Score minimum entre la population de surface et la population cavernicole pour différentes tailles de population cavernicole et différents taux de migration et temps de divergence lorsque ce score est atteint.
Modèle démographique simplifié A. N_e en surface : 10 000. Simulations réalisées avec 10 générations de laboratoire, un temps de génération de 2 ans dans la population cavernicole et de 2 ans dans la population de surface.
%M : pourcentage de poissons migrants de la population de surface vers la population cavernicole, m : probabilité qu'un événement de migration ait lieu par an.
Vert : score inférieur à 3, orange : score compris entre 3 et 15, rouge : score supérieur ou égal à 15.

N_e CF	%M→	N_e SF 5 000 10%		1%		0,1%		0,01%	
	m↓	Score	Age	Score	Age	Score	Age	Score	Age
313	10^{-2}	50,95	23 600	36,88	4 800	3,90	74 300	14,45	1 700
	10^{-3}	9,60	48 600	3,88	41 700	8,48	10 100	14,17	2 000
	10^{-4}	3,64	56 600	2,98	15 900	15,64	1 900	13,10	1 500
	10^{-5}	16,77	2 200	3,19	13 800	13,44	2 200	17,12	2 200
625	10^{-2}	99,07	76 900	55,77	80 600	3,51	27 100	10,43	5 000
	10^{-3}	8,17	8 100	2,38	17 300	9,85	4 100	9,33	4 600
	10^{-4}	3,95	25 000	7,72	7 300	10,36	4 100	10,82	3 500
	10^{-5}	3,45	22 100	9,78	3 600	8,53	4 100	9,82	3 900
1 250	10^{-2}	110,51	99 500	66,72	53 400	12,66	28 200	11,14	11 500
	10^{-3}	17,25	65 900	7,79	25 000	7,43	8 600	9,12	11 100
	10^{-4}	7,46	8 300	11,28	11 600	9,79	13 000	7,86	9 600
	10^{-5}	9,94	10 800	7,80	8 500	8,04	8 900	9,81	10 700
2 500	10^{-2}	123,24	54 600	81,61	56 400	28,84	24 800	20,34	19 800
	10^{-3}	56,28	48 300	25,16	17 500	21,19	18 700	22,30	13 700
	10^{-4}	19,86	22 200	19,61	24 600	21,91	21 700	17,51	18 300
	10^{-5}	22,77	20 500	23,35	18 000	18,93	15 000	18,63	22 400
5 000	10^{-2}	124,01	83 800	100,48	5 600	66,68	18 100	58,79	16 400
	10^{-3}	93,95	4 300	63,30	14 900	57,96	18 800	61,93	53 700
	10^{-4}	63,49	18 800	63,15	40 300	58,19	40 000	58,30	20 300
	10^{-5}	56,06	37 700	59,45	23 200	54,18	45 000	61,83	41 900
10 000	10^{-2}	129,44	27 900	118,42	9 900	108,95	7 000	101,82	8 000
	10^{-3}	116,71	7 300	103,39	5 500	107,78	14 600	95,13	8 700
	10^{-4}	100,86	10 400	103,88	12 400	104,70	6 700	102,76	12 600
	10^{-5}	105,41	8 600	109,96	7 300	98,55	9 900	104,71	6 400

Tableau 17. Score minimum entre la population de surface et la population cavernicole pour différentes tailles de population cavernicole et différents taux de migration et temps de divergence lorsque ce score est atteint.
Modèle démographique simplifié A. N_e en surface : 5 000. Simulations réalisées avec 10 générations de laboratoire, un temps de génération de 2 ans dans la population cavernicole et de 5 ans dans la population de surface.
%M : pourcentage de poissons migrants de la population de surface vers la population cavernicole, m : probabilité qu'un événement de migration ait lieu par an.
Vert : score inférieur à 3, orange : score compris entre 3 et 15, rouge : score supérieur ou égal à 15.

N_e CF	%M→	N_e SF 10 000 10%		1%		0,1%		0,01%	
	m↓	Score	Age	Score	Age	Score	Age	Score	Age
313	10^{-2}	61,93	27 500	58,45	8 200	14,51	28 700	17,11	1 500
	10^{-3}	14,18	24 600	11,47	40 200	4,79	67 000	19,42	1 700
	10^{-4}	11,14	58 300	2,76	47 900	16,05	1 800	19,78	2 000
	10^{-5}	11,26	26 400	19,82	1 800	17,67	1 800	15,32	1 900
625	10^{-2}	110,14	25 000	78,02	45 300	12,94	14 900	14,94	4 000
	10^{-3}	14,62	71 400	10,59	80 300	8,34	31 500	15,50	3 200
	10^{-4}	9,68	57 600	5,35	44 200	14,71	4 200	13,69	3 200
	10^{-5}	13,66	3 900	4,18	40 800	15,45	3 500	13,68	3 900
1 250	10^{-2}	126,15	97 100	96,77	88 300	13,92	65 400	10,29	8 100
	10^{-3}	13,85	12 700	7,50	56 500	5,54	30 700	12,37	9 400
	10^{-4}	4,10	84 900	6,31	35 700	10,52	8 300	10,42	8 300
	10^{-5}	9,77	7 200	5,48	34 700	8,72	8 000	11,00	8 500
2 500	10^{-2}	137,81	80 700	109,27	96 100	24,09	32 800	8,32	18 600
	10^{-3}	48,47	58 500	12,71	28 300	7,57	25 000	8,25	16 200
	10^{-4}	8,69	80 100	7,92	18 200	10,01	14 000	9,84	16 200
	10^{-5}	9,32	21 100	10,54	22 200	8,17	18 200	9,02	15 900
5 000	10^{-2}	142,33	2 600	108,86	30 700	44,82	32 500	22,07	33 800
	10^{-3}	94,45	84 200	26,48	32 500	23,08	43 100	25,13	36 900
	10^{-4}	27,74	61 600	20,92	32 400	20,84	43 300	22,83	43 400
	10^{-5}	20,37	35 900	18,52	32 600	21,50	39 500	22,43	32 600
10 000	10^{-2}	139,38	3 900	114,57	43 500	80,48	33 300	61,93	92 500
	10^{-3}	96,26	15 800	77,30	21 500	67,23	74 800	63,87	96 700
	10^{-4}	74,37	78 300	63,67	29 900	63,17	50 300	60,44	46 300
	10^{-5}	60,63	38 200	61,04	76 700	64,90	86 000	64,15	75 200

Tableau 18. Score minimum entre la population de surface et la population cavernicole pour différentes tailles de population cavernicole et différents taux de migration et temps de divergence lorsque ce score est atteint.
Modèle démographique simplifié A. N_e en surface : 10 000. Simulations réalisées avec 10 générations de laboratoire, un temps de génération de 2 ans dans la population cavernicole et de 5 ans dans la population de surface.
%M : pourcentage de poissons migrants de la population de surface vers la population cavernicole, m : probabilité qu'un événement de migration ait lieu par an.
Vert : score inférieur à 3, orange : score compris entre 3 et 15, rouge : score supérieur ou égal à 15.

		N_e SF 5 000		N_e SF 10 000	
N_e CF	t_2	Score	Age	Score	Age
313	0	12,71	6 710	22,28	27 210
	10 000	4,15	14 710	11,07	14 510
	20 000	9,74	25 110	6,36	24 610
	40 000	31,04	42 610	8,80	44 210
	80 000	95,23	82 110	33,41	83 610
625	0	5,00	10 910	13,20	11 010
	10 000	5,63	18 610	6,45	19 110
	20 000	9,51	27 610	7,38	29 510
	40 000	30,91	46 510	9,34	48 910
	80 000	104,19	82 710	29,97	87 710
1 250	0	8,86	16 610	5,54	26 210
	10 000	15,03	28 610	6,46	34 510
	20 000	23,16	36 610	6,57	39 810
	40 000	45,64	53 610	13,36	58 810
	80 000	97,02	84 110	36,89	93 310
2 500	0	32,10	20 510	10,79	39 110
	10 000	37,49	39 210	11,77	45 710
	20 000	45,34	55 410	14,72	55 110
	40 000	59,58	55 510	21,10	67 310
	80 000	115,67	81 810	44,63	98 710
5 000	0	72,40	24 810	34,73	56 310
	10 000	67,25	26 510	33,87	65 110
	20 000	68,93	34 910	38,15	78 510
	40 000	80,60	46 610	47,11	86 510
	80 000	108,14	80 110	65,55	95 910
10 000	0	103,39	11 810	72,85	47 910
	10 000	91,00	17 610	71,38	47 210
	20 000	79,19	21 810	72,08	90 610
	40 000	86,29	40 710	75,00	86 410
	80 000	109,68	81 010	81,11	98 110

Tableau 19. Score minimum et temps de divergence entre la population de surface et cavernicole lorsque ce score est atteint pour différentes tailles de population cavernicole et différents temps t_2 en utilisant le modèle démographique simplifié B. Simulations réalisées avec 5 générations de laboratoire, un temps de génération de 5 ans dans la population cavernicole et de 2 ans dans la population de surface. Vert : score inférieur à 3, orange : score compris entre 3 et 15, rouge : score supérieur ou égal à 15.

		N_e SF 5 000		N_e SF 10 000	
N_e CF	t_2	Score	Age	Score	Age
313	0	21,75	17 100	29,94	32 300
	10 000	14,09	12 000	22,37	11 700
	20 000	8,61	22 000	18,39	22 000
	40 000	4,29	41 700	10,08	42 000
	80 000	21,34	81 600	6,27	81 900
625	0	16,92	17 400	22,88	34 400
	10 000	11,38	14 300	19,67	14 300
	20 000	4,83	24 100	13,64	23 400
	40 000	5,45	45 600	8,42	43 900
	80 000	14,54	83 600	6,29	84 000
1 250	0	13,45	21 900	17,21	41 700
	10 000	11,38	22 400	14,67	18 400
	20 000	8,40	30 900	12,19	27 100
	40 000	9,13	46 900	5,79	48 300
	80 000	19,89	89 500	5,99	87 600
2 500	0	25,12	24 800	15,82	23 700
	10 000	24,83	22 500	13,60	47 300
	20 000	23,89	48 600	10,95	48 100
	40 000	26,33	63 900	10,14	58 200
	80 000	33,35	99 600	10,70	95 400
5 000	0	68,49	77 500	28,34	76 100
	10 000	66,44	84 500	27,72	83 300
	20 000	70,91	98 800	25,00	87 600
	40 000	76,46	85 800	27,49	93 200
	80 000	75,83	85 100	35,88	98 800
10 000	0	143,32	14 400	83,74	97 800
	10 000	125,10	17 300	76,71	97 300
	20 000	108,73	21 700	79,97	96 500
	40 000	83,69	40 100	73,98	65 300
	80 000	92,30	81 400	78,95	90 700

Tableau 20. Score minimum et temps de divergence entre la population de surface et cavernicole lorsque ce score est atteint pour différentes tailles de population cavernicole et différents temps t_2 en utilisant le modèle démographique simplifié B. Simulations réalisées avec 10 générations de laboratoire, un temps de génération de 2 ans dans la population cavernicole et de 5 ans dans la population de surface. Vert : score inférieur à 3, orange : score compris entre 3 et 15, rouge : score supérieur ou égal à 15.

		N_e SF 5 000		N_e SF 10 000	
N_e CF	t_2	Score	Age	Score	Age
313	0	4,84	5 590	11,54	4 790
	10 000	4,44	3 290	2,02	4 890
	20 000	18,32	2 290	4,16	3 890
	40 000	52,83	1 490	16,32	2 990
	80 000	118,13	290	47,71	1 990
625	0	1,31	9 490	4,73	10 990
	10 000	7,95	5 690	1,53	7 590
	20 000	24,28	5 290	5,04	7 890
	40 000	54,95	4 290	18,49	6 690
	80 000	111,48	190	50,09	3 990
1 250	0	8,61	16 690	2,10	17 690
	10 000	16,28	13 090	2,72	17 390
	20 000	27,40	10 390	9,01	10 790
	40 000	58,72	7 790	24,22	9 590
	80 000	108,76	1 090	51,24	4 290
2 500	0	27,38	20 890	7,71	29 390
	10 000	38,08	17 790	12,98	29 090
	20 000	43,68	9 690	18,74	19 790
	40 000	65,51	8 590	29,08	20 190
	80 000	109,01	790	61,34	12 190
5 000	0	57,40	21 690	28,70	35 290
	10 000	56,39	4 090	31,32	34 890
	20 000	63,12	4 390	34,53	22 790
	40 000	64,59	1 490	47,23	20 690
	80 000	104,77	590	66,26	4 390
10 000	0	75,69	14 490	56,75	44 990
	10 000	60,77	2 390	57,96	30 990
	20 000	62,03	290	55,28	8 490
	40 000	67,86	590	60,64	3 190
	80 000	112,20	1 290	69,88	4 090

Tableau 21. Score minimum et temps de divergence entre la population de surface et cavernicole lorsque ce score est atteint pour différentes tailles de population cavernicole et différents temps t_2 en utilisant le modèle démographique simplifié B. Simulations réalisées avec 10 générations de laboratoire, un temps de génération de 5 ans dans la population cavernicole et de 2 ans dans la population de surface. Vert : score inférieur à 3, orange : score compris entre 3 et 15, rouge : score supérieur ou égal à 15.

		N_e SF 5 000		N_e SF 10 000	
N_e CF	t_2	Score	Age	Score	Age
313	0	12,87	2 790	19,47	1 690
	10 000	2,97	1 690	9,90	1 790
	20 000	2,43	1 490	5,80	1 490
	40 000	11,71	1 490	2,18	1 690
	80 000	34,46	790	9,25	1 190
625	0	9,95	4 590	17,09	2 890
	10 000	5,45	3 190	7,81	4 190
	20 000	4,45	3 290	4,05	3 690
	40 000	11,06	2 590	2,81	3 190
	80 000	35,85	2 790	10,93	3 190
1 250	0	10,31	10 690	12,61	7 890
	10 000	7,75	9 890	5,45	7 790
	20 000	9,80	6 690	3,31	7 190
	40 000	16,87	6 690	2,61	6 190
	80 000	37,72	5 590	11,44	5 590
2 500	0	18,87	17 190	9,68	19 290
	10 000	19,85	9 890	8,59	16 290
	20 000	23,28	15 890	6,91	17 490
	40 000	34,83	11 490	8,38	13 690
	80 000	46,39	10 790	14,97	10 390
5 000	0	62,40	52 890	20,54	38 890
	10 000	55,81	23 690	22,61	36 590
	20 000	62,15	51 590	20,29	33 390
	40 000	51,44	16 890	21,83	29 690
	80 000	71,94	2 790	31,87	16 890
10 000	0	102,89	7 490	63,29	71 490
	10 000	78,28	190	63,08	83 990
	20 000	81,65	190	60,83	69 290
	40 000	73,18	1 190	67,61	5 390
	80 000	67,39	290	56,90	9 090

Tableau 22. Score minimum et temps de divergence entre la population de surface et cavernicole lorsque ce score est atteint pour différentes tailles de population cavernicole et différents temps t_2 en utilisant le modèle démographique simplifié B. Simulations réalisées avec 10 générations de laboratoire, un temps de génération de 2 ans dans la population cavernicole et de 5 ans dans la population de surface. Vert : score inférieur à 3, orange : score compris entre 3 et 15, rouge : score supérieur ou égal à 15.

		N_e SF 5000		N_e SF 10000	
N_e CF	t_2	Score	Age	Score	Age
313	0	10,72	1 96	15,69	1 62
	10 000	5,18	1 76	10,28	2 18
	20 000	2,85	1 96	7,95	1 7
	40 000	3,59	1 6	4,34	1 68
	80 000	11,51	1 68	2,39	1 52
625	0	5,44	4 3	10,61	3 58
	10 000	3,53	4 06	7,78	3 98
	20 000	2,89	3 8	5,74	3 62
	40 000	4,91	3 16	2,85	3 78
	80 000	13,46	3 06	3,19	3 28
1250	0	5,24	7 74	5,46	8
	10 000	3,65	7 46	4,7	9 26
	20 000	5,01	6 6	3,18	7 18
	40 000	9,94	5 12	3,25	7 34
	80 000	19,53	5 4	3,99	6 8
2500	0	20,87	12 86	4,89	17 14
	10 000	22,52	13 14	4,78	16 8
	20 000	20,92	13 94	5,97	17 1
	40 000	27,62	14 26	6,29	15 08
	80 000	33,67	10 98	11,09	14 42
5000	0	63,32	10 54	21,72	31 06
	10 000	59,84	11 68	22,75	24 92
	20 000	61,51	15 18	22,39	23 5
	40 000	53,61	7 78	21,71	26 78
	80 000	61,15	5 04	25,77	25 1
10000	0	94,67	5 86	62,24	31 72
	10 000	87,01	3 98	59,66	39 16
	20 000	80,55	1 52	58,34	20 8
	40 000	69,52	3 04	64,48	16 4
	80 000	62,23	260	63,06	3 96

Tableau 23. Score minimum et temps de divergence entre la population de surface et cavernicole lorsque ce score est atteint pour différentes tailles de population cavernicole et différents temps t_2 en utilisant le modèle démographique simplifié B. Simulations réalisées avec 10 générations de laboratoire, un temps de génération de 2 ans dans la population cavernicole et de 2 ans dans la population de surface. Vert : score inférieur à 3, orange : score compris entre 3 et 15, rouge : score supérieur ou égal à 15.

3.4 Étude d'un éventuel biais d'échantillonnage

L'estimation de l'âge de la population cavernicole dépend fortement des estimations des fréquences des différentes catégories de SNP obtenues au chapitre précédent. On peut donc se demander si ces résultats sont biaisés. En particulier on peut se demander si l'échantillonnage des poissons réalisé ne biaise pas l'identification des SNP ni leur classification dans les différentes catégories. En effet, les embryons séquencés sont issus de différentes pontes indépendantes à l'élevage de Gif-sur-Yvette. Le nombre d'embryons séquencés varie de 50 à 200. Par ailleurs, il est possible que pour un locus donné le nombre de séquences par embryon soit très variable indépendamment de son génotype.

Pour évaluer l'effet possible de biais d'échantillonnage, nous avons échantillonné de différentes façons des allèles dans des populations simulées, puis nous avons comparé les fréquences calculées aux fréquences réelles des allèles de cess populations.

Les deux populations de départ sont issues d'une simulation utilisant le modèle simplifié A avec les paramètres suivants :

- Taille de la population de surface : 10 000
- Taille de la population cavernicole : 625
- Probabilité d'un événement de migration : 10^{-4} / an
- Pourcentage de poissons de surface migrant vers la grotte : 1%
- Temps de génération en surface : 2 ans
- Temps de génération cavernicole : 5 ans

La simulation est arrêtée après 26 200 ans. Les fréquences des allèles dérivés dans la population cavernicole et la population de surface ont été calculées. Le nombre de SNP dans les différentes catégories ainsi que la fréquence relative des différentes catégories pour cet état allélique sont présentés Tableau 24.

Nous avons ensuite échantillonné dans ces populations entre 1 et 50 individus. Pour tenir compte d'une éventuelle variabilité d'ARNm rétrotranscrit par individu, nous avons échantillonné à chaque locus et en tenant compte des génotypes, soit une seule séquence par individu, soit un nombre aléatoire de séquences compris entre 0 et 10, 100 ou 1 000. Nous avons ensuite calculé le nombre de SNP dans chacune des huit catégories de polymorphisme définies Tableau 1 et les fréquences relatives de ces catégories. Pour comparer ces données avec les données de départ, nous avons calculé un score de χ^2 (Équation 6) entre les fréquences relatives observées entre les deux types de données. Les scores calculés sont présentés graphiquement Figure 38.

On observe que si le nombre de poissons échantillonnés est petit, c'est-à-dire moins de 5 individus, le score entre les fréquences calculées et réelles est relativement élevé. Par contre, au delà de 5 individus échantillonnés,

Catégorie	Nombre de SNP	%
1	253	12,5 %
2	119	5,9 %
3	237	11,7 %
4	58	2,9 %
5	1 011	50,0 %
6	196	9,7 %
7	147	7,3 %
8	0	0 %
Total	**2 021**	100 %

Tableau 24. Répartition des SNP dans les huit catégories définies Tableau 1.

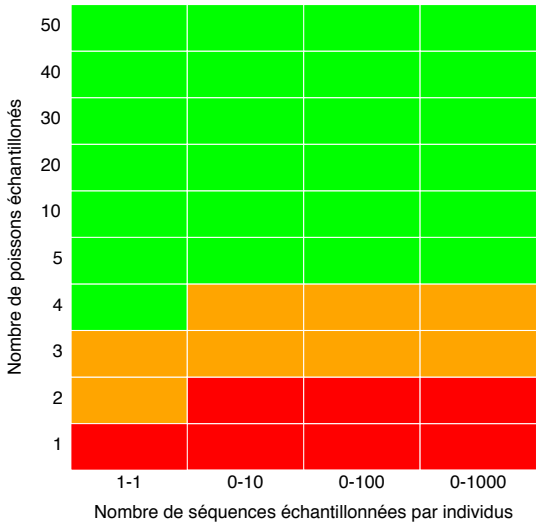


Figure 38. Score de la comparaison fréquences estimées / fréquences réelles. La couleur dépend du score : vert, score inférieur à 3 ; orange, score entre 3 et 15 ; rouge, score supérieur à 15.

même si la variance individuelle est très grande (entre 0 et 1000 séquences échantillonnées par individu), l'estimation des fréquences des différentes catégories est très bonne. Cela signifie qu'au delà de 5 poissons échantillonnés dans une population, les fréquences relatives dans les huit catégories de SNP seront très proches des fréquences réelles.

3.5 Analyse des simulations et âge de la population Pachón

Nous avons vu que les simulations avec des tailles de populations réduites dans la grotte Pachón donnaient les meilleurs ajustements entre données simulées et données observées.

Nous avons choisi d'analyser plus en détail trois simulations obtenues avec le modèle simplifié A (Figure 37) et avec une taille de population de surface de 10 000 individus, un temps de génération de 5 ans pour la population cavernicole et de 2 ans pour la population de surface et avec 10 générations en laboratoire.

3.5.1 *Simulation avec un mauvais ajustement*

La première simulation présente un mauvais ajustement entre les données réelles et les données simulées. Elle a été réalisée avec une taille de population cavernicole de 313 et une taille de population de surface de 10 000. Les migrations de la population de surface vers la population cavernicole étaient importantes : en moyenne tous les 100 ans (probabilité d'un événement de migration : 10^{-2}/an) avec 1 000 individus (pourcentage de poissons migrants : 10%). Les migrations dans cette simulation sont si fréquentes et si importantes, que la population cavernicole est en fait un duplicat de la population de surface.

Avec ces paramètres, l'ajustement entre les données simulées et les données réelles est fluctuant au cours du temps avec un score entre 100 et 150 (Figure 39 A), avec de forts changements après chaque événement de migration. Le meilleur score (96,13) est obtenu après 37 700 ans.

Le ratio de fixation des allèles dérivés est de 1 tout au long de la simulation (Figure 39 B), signifiant que dans les deux populations, on observe le même nombre de loci avec un allèle dérivé fixé. Le ratio de polymorphisme entre les deux populations est également de 1 (Figure 39 B). Le nombre de loci polymorphes est le même dans les deux populations malgré leur forte différence de taille (Figure 39 D). Avec ces paramètres il est donc impossible d'observer une différence de fixation entre la population de surface et la population cavernicole et il est également impossible d'observer une différence de niveau de polymorphisme entre ces deux populations.

Les fréquences relatives des 7 catégories de SNP sont constantes au cours du temps (Figure 39 C). Le polymorphisme partagé entre les populations est majoritaire ($\approx$ 40%) montrant une forte homogénéité des deux populations. La fréquence des catégories 1 et 2 (courbes cyan et bleue), où les deux populations ne présentent qu'un seul allèle, l'une des deux ayant fixé un allèle dérivé (Tableau 1), est très proche de 0 (Figure 39 C) : le fort taux de migration empêche la fixation dans la population cavernicole d'un allèle différent de celui de la population de surface. Les catégories 4 et 6 correspondant aux loci polymorphes dans une population et avec un allèle dérivé fixé dans l'autre population (Tableau 1) sont également à faible fréquence ($\approx$ 5%) (Figure 39 C). Cela s'explique, encore une fois, par le grand nombre d'événements de migration et le grand nombre de poissons migrants. Dans la population cavernicole, la fixation d'un allèle dérivé sera empêchée par l'arrivée massive du polymorphisme de surface. Lorsqu'un allèle se fixe dans la population de surface, les migrations vont apporter un

grand nombre de cet allèle dans la population cavernicole où il se fixera également rapidement.

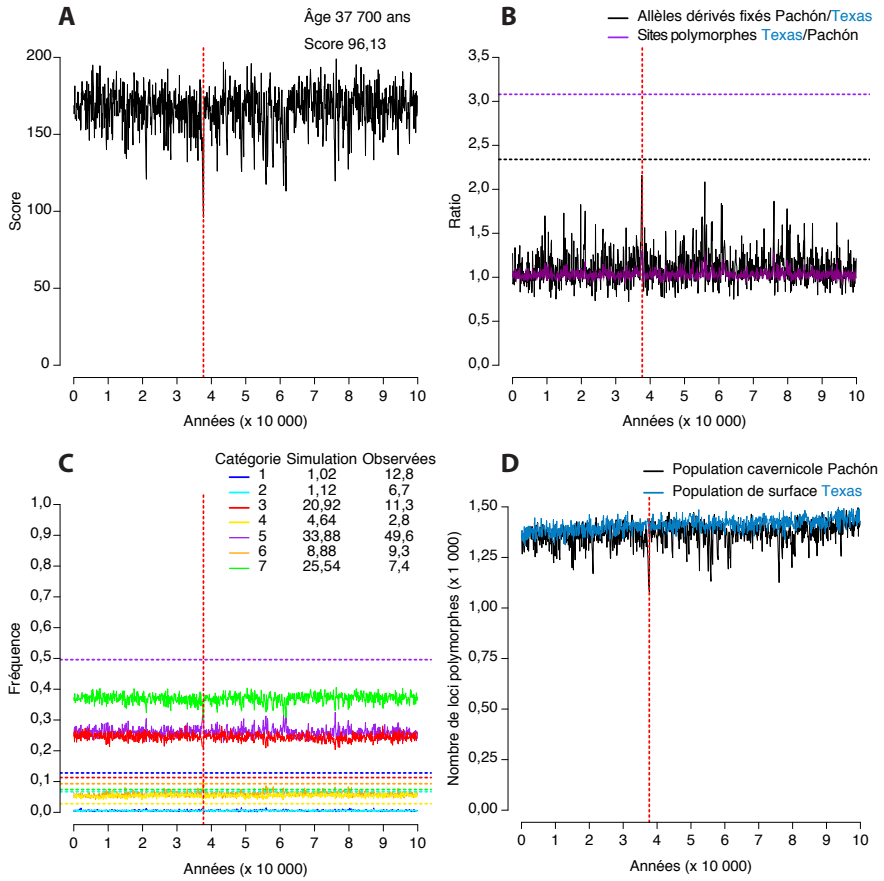


Figure 39. Résultat de la simulation avec le modèle démographique simplifié A et les paramètres suivants : taille de la population de surface (N_e SF = 10 000), taille de la population cavernicole (N_e CF = 313), probabilité d'un événement de migration : 10^{-2}/an, pourcentage de poissons migrants : 10%. La ligne verticale en pointillés rouges correspond à la génération présentant le meilleur ajustement (score le plus faible) entre données simulées et données observées. (A) Score de la simulation au cours du temps. (B) Ratios de polymorphisme au cours du temps entre la population de surface et la population cavernicole (en violet) et ratio de fixation d'allèles dérivés au cours du temps entre la population cavernicole et la population de surface (en noir). (C) Fréquence relative des 7 catégories de SNP dans la simulation (lignes pleines) et observées dans les populations (lignes horizontales en pointillés). (D) Nombre de loci polymorphes dans la population de surface (en bleu) et dans la population cavernicole (en noir).

3.5.2 *Simulation avec un ajustement moyen*

Pour cet exemple de simulation qui donne un ajustement moyen, nous avons choisi celle réalisée avec une taille de population cavernicole de 2 500 et une

taille de population de surface de 10 000. Les migrations de la population de surface vers la population cavernicole étaient peu fréquentes : en moyenne tous les 10 000 ans (probabilité d'un événement de migration : 10^{-5}) mais avec un nombre important de poissons migrant de la surface vers les grottes : 100 individus (pourcentage de poissons migrants : 1%).

Avec ces paramètres, l'ajustement entre les données simulées et les données réelles est mauvais en début de simulation (score $\approx$ 175) puisque les données simulées sont très différentes des données réelles. Puis, le score diminue fortement jusqu'à atteindre un minimum (score = 8,23) et réaugmente ensuite, les populations se différenciant trop dans la simulation par rapport aux données réelles (Figure 40 A). Le score minimum, correspondant au meilleur ajustement entre données réelles et simulées, est de 8,23 et est obtenu après 29 300 ans.

À l'âge du meilleur ajustement, le ratio de fixation des allèles dérivés entre la population cavernicole et la population de surface est d'environ 1,5 (Figure 40 B), signifiant que la population cavernicole a fixé légèrement plus d'allèles dérivés que la population de surface. Le ratio de polymorphisme entre les deux populations est de 2 (Figure 40 B). Le nombre de loci polymorphes est deux fois plus petit dans la population cavernicole que dans la population de surface au temps du meilleur ajustement (Figure 40 D). Avec ces paramètres, on peut donc observer une différence de niveau de polymorphisme entre les deux populations ainsi que dans la fixation des allèles dérivés, mais ces différences sont encore assez éloignées des différences observées dans les données réelles. Ce résultat pourrait être dû à la différence des tailles de populations ici peu importante.

À l'âge du meilleur ajustement entre les deux populations, les fréquences relatives des différentes catégories de SNP obtenues par simulation sont relativement éloignées de celles observées. Les fréquences relatives des 7 catégories de SNP sont constantes au cours du temps (Figure 40 C). Le polymorphisme partagé entre les populations est majoritaire ($\approx$ 40%) montrant une forte homogénéité des deux populations.

3.5.3 *Simulation avec un bon ajustement*

La simulation qui a donné le meilleur ajustement a été réalisée avec une taille de population cavernicole de 625 et de surface de 10 000. Des migrations de la population de surface vers la population cavernicole étaient possibles : en moyenne tous les 1 000 ans (probabilité d'un événement de migration : 10^{-3}) avec 10 individus (pourcentage de poissons migrants : 0,1%).

Avec ces paramètres, l'ajustement entre les données simulées et les données réelles est mauvais au début de la simulation, les données simulées ne ressemblant pas aux données réelles. Puis le score de la simulation diminue au cours du temps, signe d'un meilleur ajustement entre les données simulées et les données réelles, avant d'atteindre un mininum et de réaugmenter lorsque l'ajustement entre simulation et données réelles se dégrade

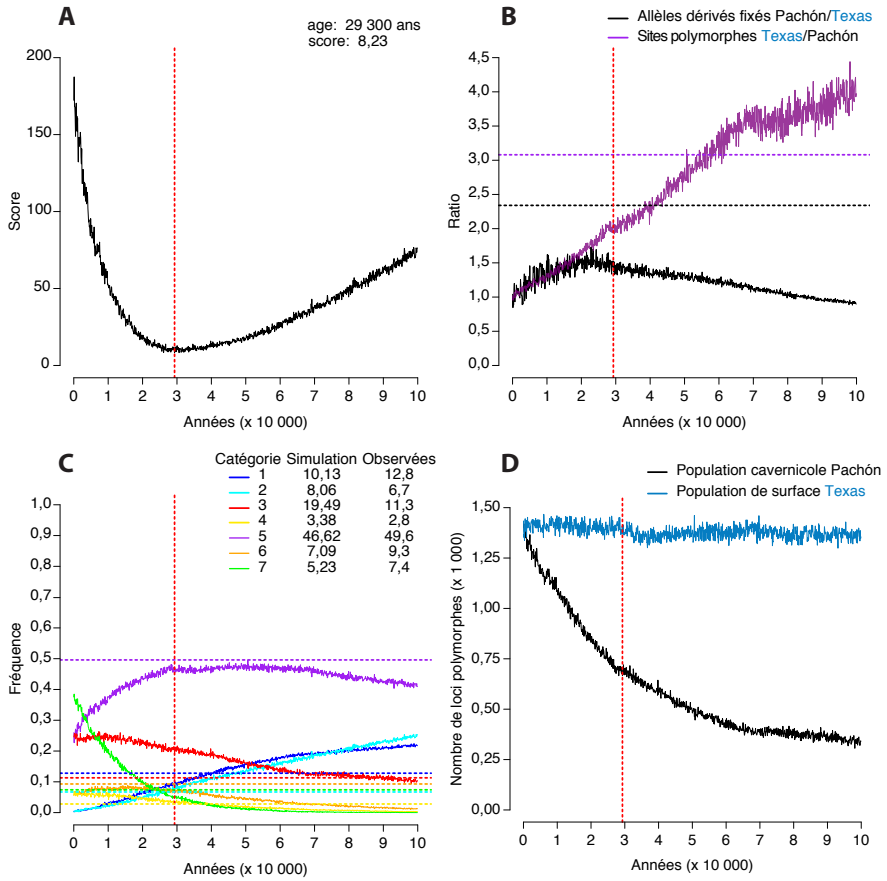


Figure 40. Résultat de la simulation avec le modèle démographique simplifié A et les paramètres suivants : taille de la population de surface (N_e SF = 10 000), taille de la population cavernicole (N_e CF = 2 500), probabilité d'un événement de migration : 10^{-5}, pourcentage de poissons migrants : 1%. La ligne verticale en pointillés rouges correspond à la génération présentant le meilleur ajustement (score le plus faible) entre données simulées et données observées. (A) Score de la simulation au cours du temps. (B) Ratios de polymorphisme au cours du temps entre la population de surface et la population cavernicole (en violet) et ratio de fixation d'allèles dérivés au cours du temps entre la population cavernicole et la population de surface (en noir). (C) Fréquence relative des 7 catégories de SNP dans la simulation (lignes pleines) et observées dans les populations (lignes horizontales en pointillés). (D) Nombre de loci polymorphes dans la population de surface (en bleu) et dans la population cavernicole (en noir).

(Figure 41 A). Le score minimum est de 0,15 et est atteint après 25 500 ans.

Le ratio de fixation des allèles dérivés atteint un pic de 4,5 avant de redescendre lentement vers 1 en fin de simulation (Figure 41 B). Le ratio de polymorphisme est fluctuant au cours du temps, mais les populations de surface possèdent toujours plus de loci à l'état polymorphe que les populations de surface. (Figure 41 B et D). À l'âge du meilleur ajustement, les ratios de

fixation des allèles dérivés et de polymorphisme dans les simulations sont très proches des valeurs observées.

Les fréquences relatives des 7 catégories de SNP sont également, à l'âge du meilleur ajustement, très proches des valeurs observées (Figure 41 C).

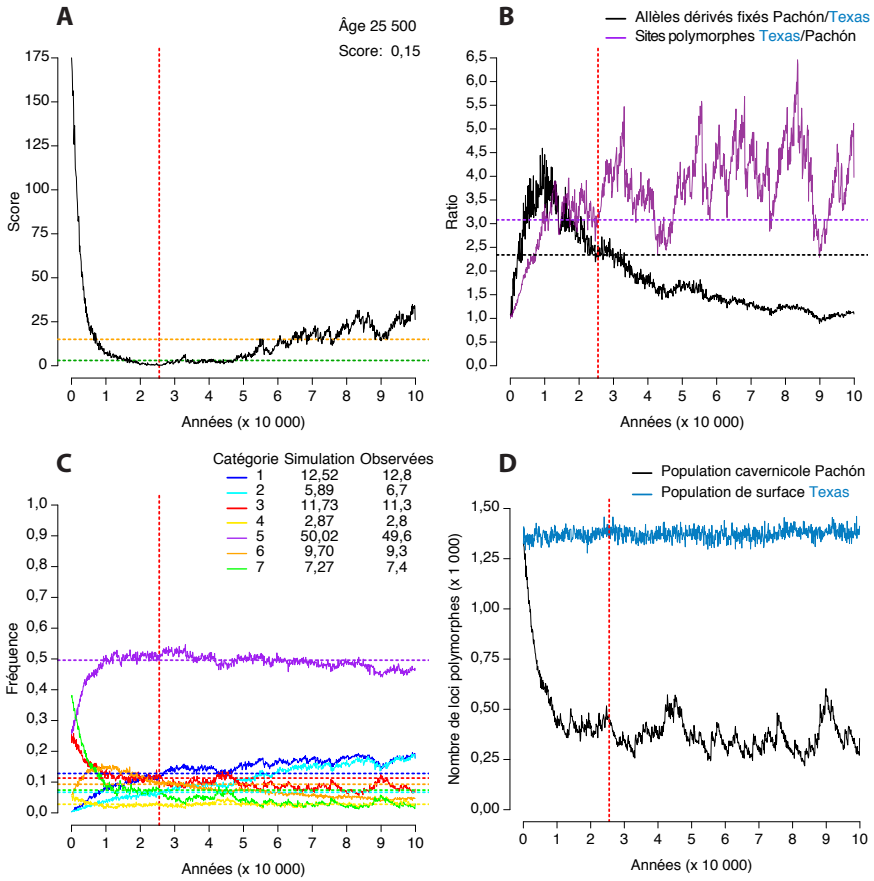


Figure 41. Résultat de la simulation avec le modèle démographique simplifié A et les paramètres suivants : taille de la population de surface (N_e SF = 10 000), taille de la population cavernicole (N_e CF = 625), probabilité d'un événement de migration : 10^{-3}, pourcentage de poissons migrants : 0,1%. La ligne verticale en pointillés rouges correspond à la génération présentant le meilleur ajustement (score le plus faible) entre données simulées et données observées. (A) Score de la simulation au cours du temps. Les lignes horizontales en pointillés représentent les limites de score de 3 (vert) et 15 (orange). (B) Ratios de polymorphisme au cours du temps entre la population de surface et la population cavernicole (en violet) et ratio de fixation d'allèles dérivés au cours du temps entre la population cavernicole et la population de surface (en noir). (C) Fréquence relative des 7 catégories de SNP dans la simulation (lignes pleines) et observées dans les populations (lignes horizontales en pointillés). (D) Nombre de loci polymorphes dans la population de surface (en bleu) et dans la population cavernicole (en noir).

3.5.4 *Résultat cyclique*

Quelques jeux de paramètres permettent d'obtenir un bon ajustement entre données simulées et observées avec un âge supérieur à 50 000 ans. Ces jeux de paramètres impliquent des migrations fréquentes et/ou avec beaucoup de migrants de la population de surface vers la population cavernicole. Il est toutefois difficile de donner un âge à la population cavernicole en utilisant ces simulations car, à chaque migration, les deux populations sont réhomogénéisées avant de diverger de nouveau. Ainsi, le score de l'ajustement entre les données simulées et observées va augmenter fortement après un événement de migration puis de nouveau diminuer avant d'atteindre un minimum et éventuellement réaugmenter d'un coup suite à un nouvel événement de migration (Figure 42). Le score de l'ajustement est alors cyclique : Le minimum global peut alors être ancien mais très peu différent des autres minimum locaux plus récents.

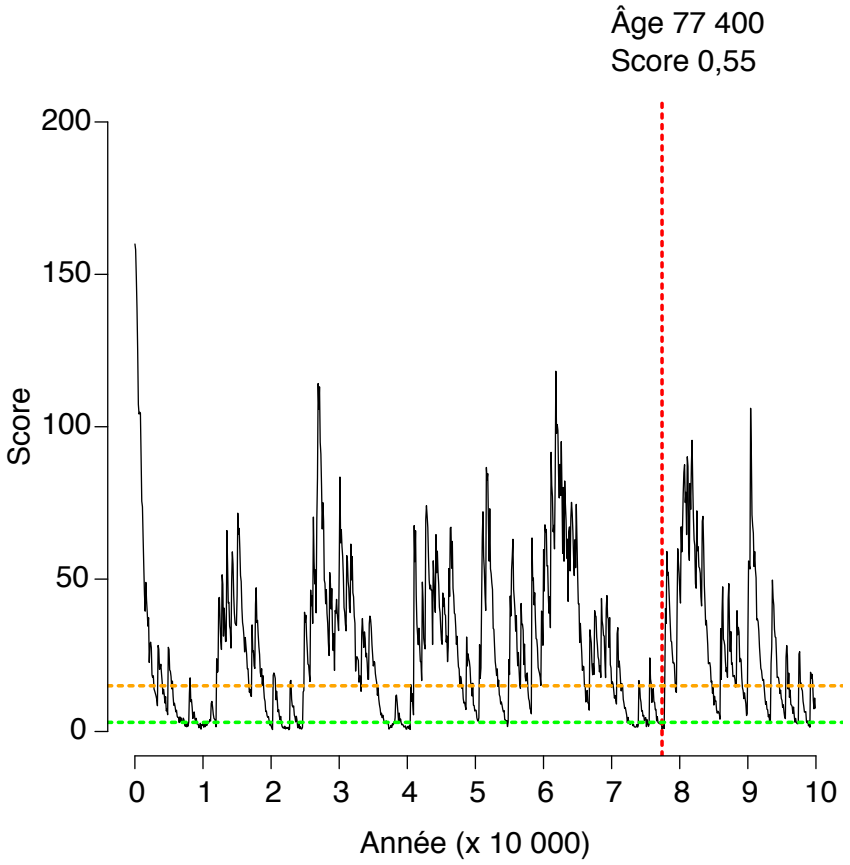


Figure 42. Score d'une simulation dans laquelle l'ajustement aux données est cyclique. Cette simulation a été réalisée avec les paramètres suivants : taille de population de surface 5 000, taille de la population cavernicole 313, probabilité d'un événement de migration 10^{-3}, pourcentage de poissons migrants 1%, temps de génération 5 ans dans la population cavernicole et 2 ans dans la population de surface.

De l'ensemble des simulations réalisées, nous observons que peu de jeux de paramètres permettent d'obtenir un bon ajustement entre les données réelles et les données simulées. La simulation ayant le meilleur ajustement, présentée ci-dessus, donne un temps de divergence entre la population de surface et la population cavernicole d'environ 25 000 ans. Ce temps de divergence suggère que la population Pachón est bien plus récente qu'actuellement décrit dans la littérature. Cette estimation de l'âge de la population Pachón a été réalisée de façon à surestimer le temps de divergence. Par exemple, la taille efficace de la population de surface utilisée (10 000) est deux fois plus grande que les estimations publiées pour des populations de surface [121] (aucune estimation n'a été publiée pour la population du Texas étudiée mais l'ensemble des populations semble correspondre à une seule population panmictique [130]). Le temps de génération de la population de surface a lui aussi été surestimé, 2 ans au lieu de 1 an dans des publications concernant d'autres espèces du genre *Astyanax*. L'utilisation de surestimations de ces deux paramètres ralentit l'évolution de la population de surface dans les simulations et donc entraîne une surestimation du temps de divergence entre la population de surface et la population cavernicole.

Nous verrons dans le chapitre suivant qu'en utilisant d'autres marqueurs, les microsatellites, disponibles dans la littérature, nous pouvons également obtenir une datation de la divergence entre populations de surface et cavernicole.

4 AUTRES ARGUMENTS ALLANT DANS LE SENS D'UNE ORIGINE RÉCENTE DE LA POPULATION PACHÓN

4.1 Datation du temps de divergence des populations en utilisant un autre marqueur moléculaire : les microsatellites

Les microsatellites sont des séquences courtes (deux à cinq nucléotides) répétées en tandem*. Ces séquences sont fortement mutables, et le nombre de répétitions en tandem d'un même motif est souvent très variable à l'intérieur d'une population. Le taux de mutation généralement observé sur les séquences microsatellites est de l'ordre de 10^{-4} mutation par locus et par génération [131]. Ce taux de mutation est bien plus élevé que celui observé pour les SNP, autour de 10^{-8} par locus et par génération [131]. Ainsi, ce fort taux de mutation va entraîner un fort polymorphisme dans les populations et une différenciation rapide de ces séquences microsatellites entre populations d'une même espèce, ce qui permet d'étudier la structuration génétique et géographique de ces espèces [132]. Il existe toutefois une forte homoplasie qui complique l'analyse quand les populations sont très divergentes. Il existe alors un grand nombre d'allèles à un locus et la plupart des individus sont hétérozygotes à ce site.

Répétition en tandem *: Répétitions adjacentes d'un motif.*

4.1.1 *Obtention des séquences microsatellites*

Nous avons récupéré les données d'une étude de la structuration des populations d'*Astyanax mexicanus* réalisée par Bradic *et al.* [121]. Des poissons de quatre populations de surface et de onze populations cavernicoles ont été échantillonnés (Figure 43) et pour chacune de ces populations, 26 locus ont été génotypés.

À partir des génotypes, nous avons calculé la fréquence des allèles observée dans chacune des populations pour chaque locus.

On peut alors calculer le nombre d'allèles observés par locus et par population, et à partir des fréquences alléliques, le nombre efficace d'allèles qui est une mesure de la diversité allélique au sein d'une population mais dépendant de la taille de l'échantillon. Le nombre efficace d'allèles correspond à la somme des carrés des fréquences alléliques [133] :

$$n_e = \frac{1}{\sum_i {p_i}^2} \tag{7}$$

Le nombre moyen d'allèles est de 12,7 dans les populations de surface et de 5,1 dans les populations cavernicoles. La moyenne du nombre effi-

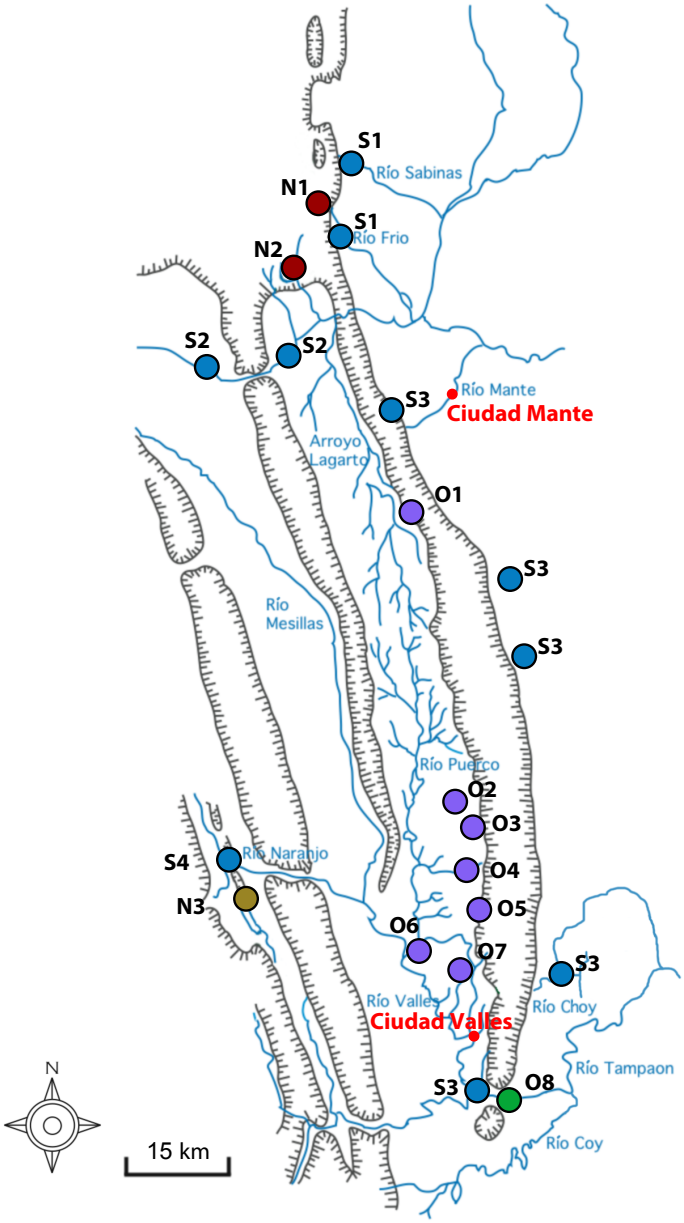


Figure 43. Carte de la Sierra de El Abra et des points d'échantillonnage des populations de surface et des populations cavernicoles. Les points bleus (S1-S4) indiquent les lieux d'échantillonnage des populations de surface. Les populations cavernicoles de la lignée dite "récente" sont représentées avec des points rouges (N1-N2) ou jaune (N3). Les populations de la lignée dite "ancienne" sont représentées avec des points violets (O1-O7) ou vert (O8). Modifié d'après [121].

cace d'allèles est de 7,0 dans les populations de surface et de 2,8 dans les populations cavernicoles (Tableau 25).

Population	Type	Id.	Nb moyen d'allèles	
			ne	n
Groupe de surface 2	**SF**	**S1**	7,1	12,6
Río Frio	*SF*	*S1*		
Arroyo Sacro	*SF*	*S1*		
Groupe de surface 3	**SF**	**S2**	6,3	10,6
Chamal	*SF*	*S2*		
Río Meco	*SF*	*S2*		
Groupe de surface 1	**SF**	**S3**	8,4	17,5
Río Tantaon	*SF*	*S3*		
Río Florido	*SF*	*S3*		
Río Tampaon	*SF*	*S3*		
Nacimiento del Río Santa Clara	*SF*	*S3*		
San Rafel Los Castros	*SF*	*S3*		
Río Subterráneo Valley	**SF**	**S4**	6,3	10,2
Moyenne populations de surface			7,0	12,7
Pachón	**CF**	**O1**	2,5	4,8
Yerbaniz	**CF**	**O2**	3,1	5,0
Japonés	**CF**	**O3**	2,8	4,1
Arroyo	**CF**	**O4**	2,6	3,7
Tinaja	**CF**	**O5**	2,4	3,0
Curva	**CF**	**O6**	2,5	3,8
Toro	**CF**	**O7**	2,7	3,0
Molino	**CF**	**N1**	1,9	3,0
Caballo Moro	**CF**	**N2**	3,0	5,7
Moyenne populations cavernicoles			2,6	4
Chica	**CF**	**O8**	3,2	9,3
Río Subterráneo	**CF**	**N3**	3,8	10,6
Moyenne populations cav. ayant des hybrides			3,5	9.95

Tableau 25. Nombre moyen d'allèles (n) et nombre efficace moyen d'allèles (ne) pour les différentes populations échantillonnées. Moyenne de ces valeurs pour les populations de surface, cavernicoles et cavernicoles ayant des poissons hybrides.

Nous avons ensuite calculé l'hétérozygotie (He) de chacun des 26 locus pour chacune des populations ainsi que l'hétérozygotie moyenne par population (Figure 44) :

$$H_e = 1 - \sum_{i=1}^{n} p_i^2 \quad (8)$$

.

La Figure 44 présente pour chaque population l'hétérozygotie moyenne ainsi que l'hétérozygotie pour chaque locus étudié. On peut y observer que les 4 populations de surface (S1 à S4) ont un niveau d'hétérozygotie moyen plus élevé que les populations cavernicoles (O1 à 08 et N1 à N3). Par exemple, l'hétérozygotie moyenne de la population S3, qui contient les populations de surface à proximité de la grotte Pachón, est de 0,85 et celle de la population O1 (la population cavernicole de la grotte Pachón), est de 0,49. Cette différence est significative (Test U de Mann-Whithney, W=622, p-value = $1{,}49.10^{-10}$). Les distributions par locus sont également plus resserrées autour de la moyenne dans les populations de surface. Dans les populations cavernicoles on observe des locus ayant une hétérozygotie de 0 ou très proche de 0, signifiant qu'ils ne possèdent qu'un seul allèle fixé ou à très haute fréquence.

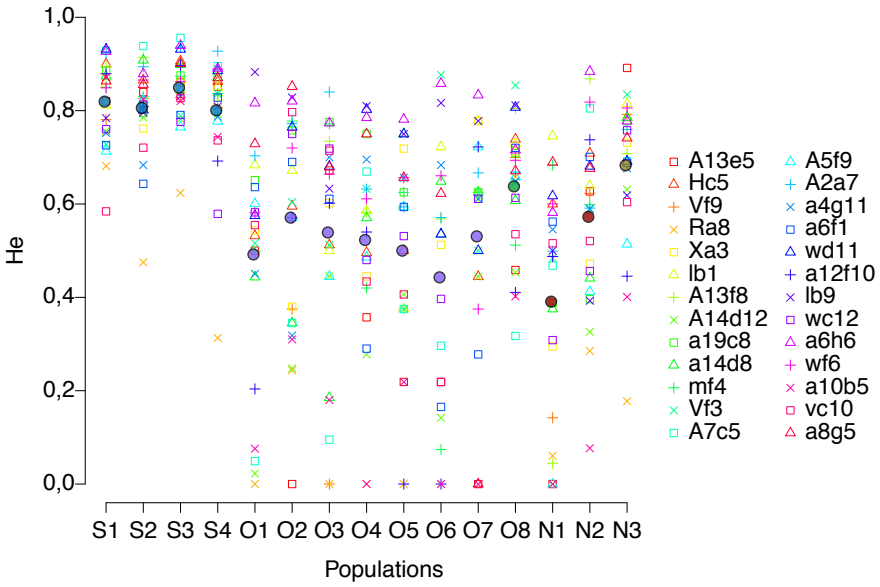


Figure 44. Hétérozygotie (He) de chaque locus (triangle, croix grecque et croix de saint André) pour chaque population et hétérozygotie moyenne de la population (cercle). Le nom des locus est donné dans la légende (à droite).

Deux populations cavernicoles (N3 et O8) sont plus polymorphes que les autres tout en l'étant moins que les populations de surface. Elles correspondent, respectivement, à la population de la grotte Río Subterráneo et à celle de Chica. Ces deux grottes sont connectées à la population de surface proche, et des hybridations peuvent avoir lieu entre poissons de surface et cavernicoles entraînant l'apparition de poissons ayant un phénotype intermédiaire.

Les populations de surface possèdent environ 2,5 fois plus d'allèles différents que les populations cavernicoles. Ce résultat est comparable au niveau de polymorphisme que nous avions observé entre la population de surface du Texas et la population cavernicole de Pachón pour les SNP (Chapitre 2).

La diversité allélique est donc plus grande dans les populations de surface que dans les populations cavernicoles. Ce résultat pourrait s'expliquer par la différence de taille des populations qui, comme nous l'avons vu, est un paramètre important dans le niveau de polymorphisme dans une population.

Nous avons représenté graphiquement les fréquences alléliques par population pour chaque locus Figure 45). On observe qu'à de nombreux loci les populations cavernicoles partagent les mêmes allèles fixés ou à des fréquences élevées. Ces allèles sont généralement présents dans la population de surface. Ces observations sont peu cohérentes avec l'hypothèse d'une divergence ancienne des populations. En effet, au cours du temps de nouveaux allèles auraient dû apparaître dans chaque population et devenir majoritaires. Nous devrions donc observer des allèles différents dans les populations cavernicoles et entre populations cavernicoles et de surface et donc un décalage des distributions.

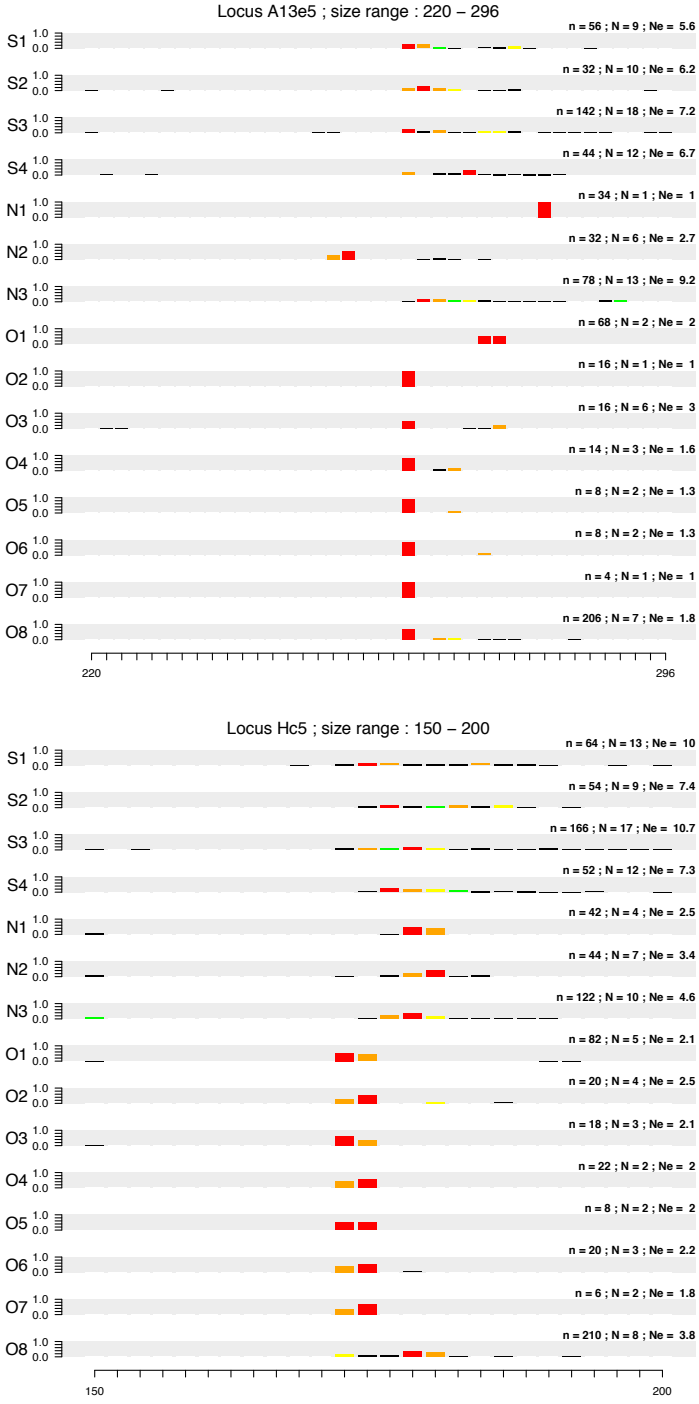


Figure 45. Fréquences alléliques dans les 16 populations étudiées pour les locus A13e5 et Hc5. n : nombre d'allèles échantillonnés, N : nombre d'allèles différents dans la population, N_e : nombre efficace d'allèles. Les allèles sont colorés par ordre de fréquence décroissante : rouge pour l'allèle le plus fréquent, puis orange, jaune et vert.

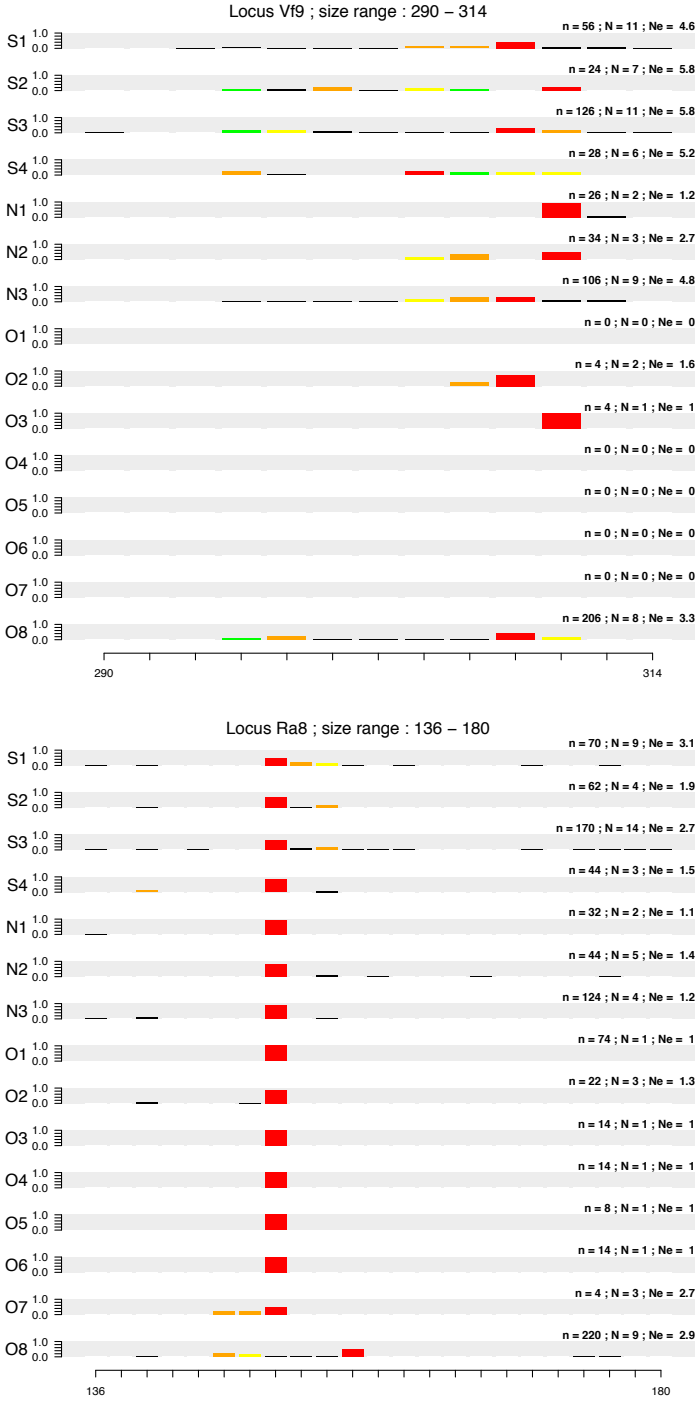


Figure 45. (suite) Fréquences alléliques dans les 16 populations étudiées pour les locus cf9 et Ra8. n : nombre d'allèles échantillonnés, N : nombre d'allèles différents dans la population, N_e : nombre efficace d'allèles. Les allèles sont colorés par ordre de fréquence décroissante : rouge pour l'allèle le plus fréquent, puis orange, jaune et vert.

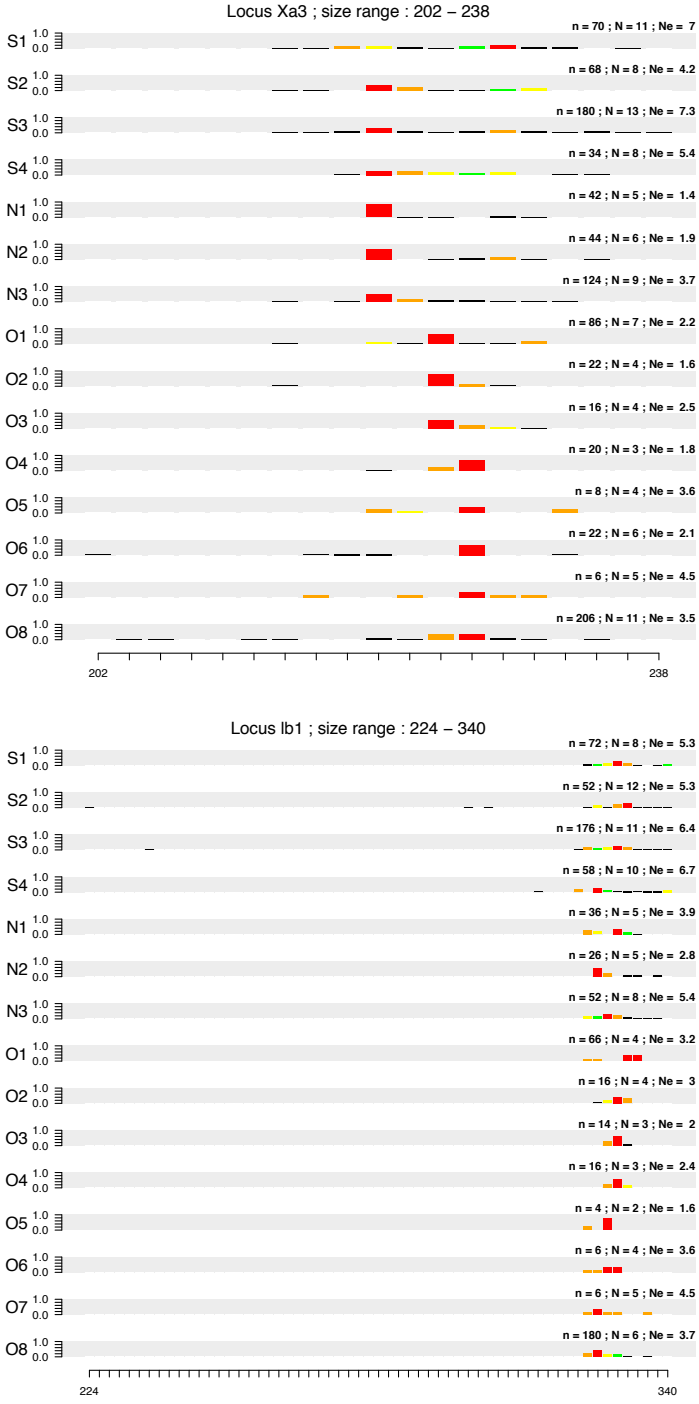


Figure 45. (suite) Fréquences alléliques dans les 16 populations étudiées pour les locus Xa3 et lb1. n : nombre d'allèles échantillonnés, N : nombre d'allèles différents dans la population, N_e : nombre efficace d'allèles. Les allèles sont colorés par ordre de fréquence décroissante : rouge pour l'allèle le plus fréquent, puis orange, jaune et vert.

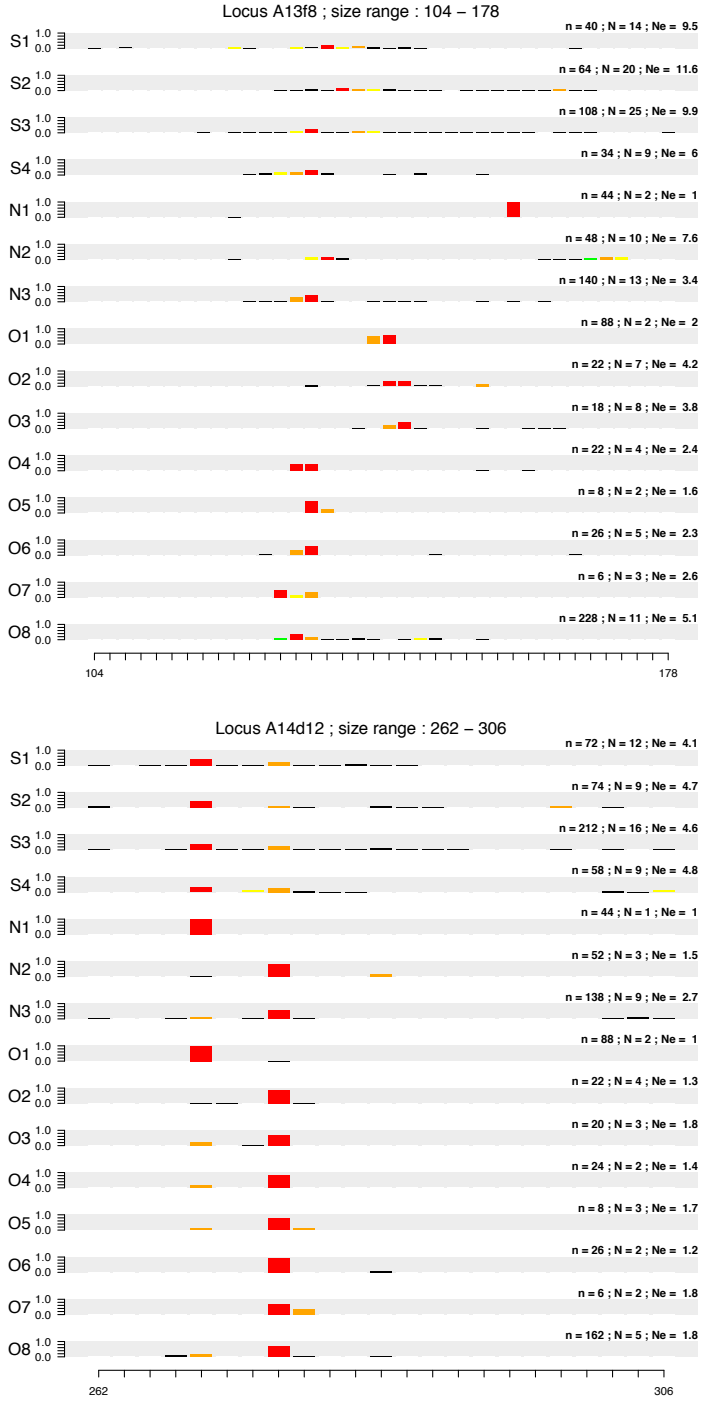


Figure 45. (suite) Fréquences alléliques dans les 16 populations étudiées pour les locus A13f8 et A14d12. n : nombre d'allèles échantillonnés, N : nombre d'allèles différents dans la population, N_e : nombre efficace d'allèles. Les allèles sont colorés par ordre de fréquence décroissante : rouge pour l'allèle le plus fréquent, puis orange, jaune et vert.

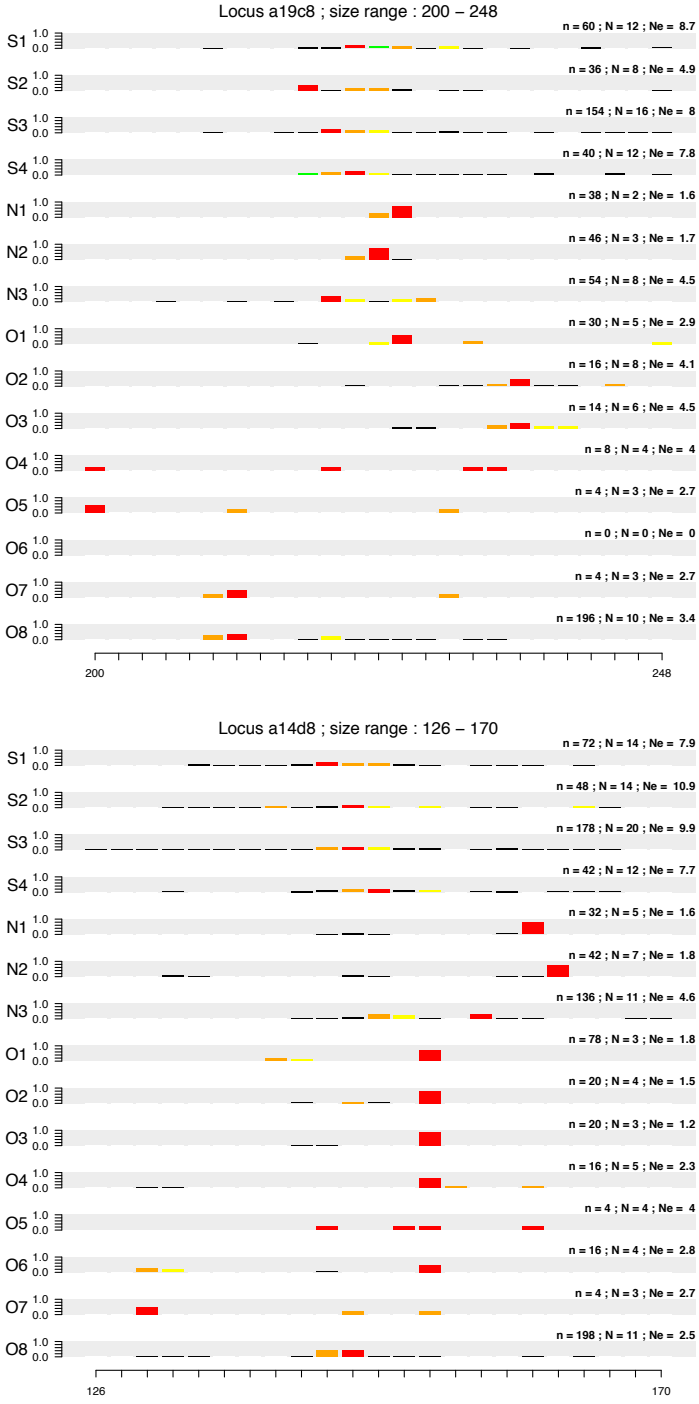


Figure 45. (suite) Fréquences alléliques dans les 16 populations étudiées pour les locus a19c8 et a14d8. n : nombre d'allèles échantillonnés, N : nombre d'allèles différents dans la population, N_e : nombre efficace d'allèles. Les allèles sont colorés par ordre de fréquence décroissante : rouge pour l'allèle le plus fréquent, puis orange, jaune et vert.

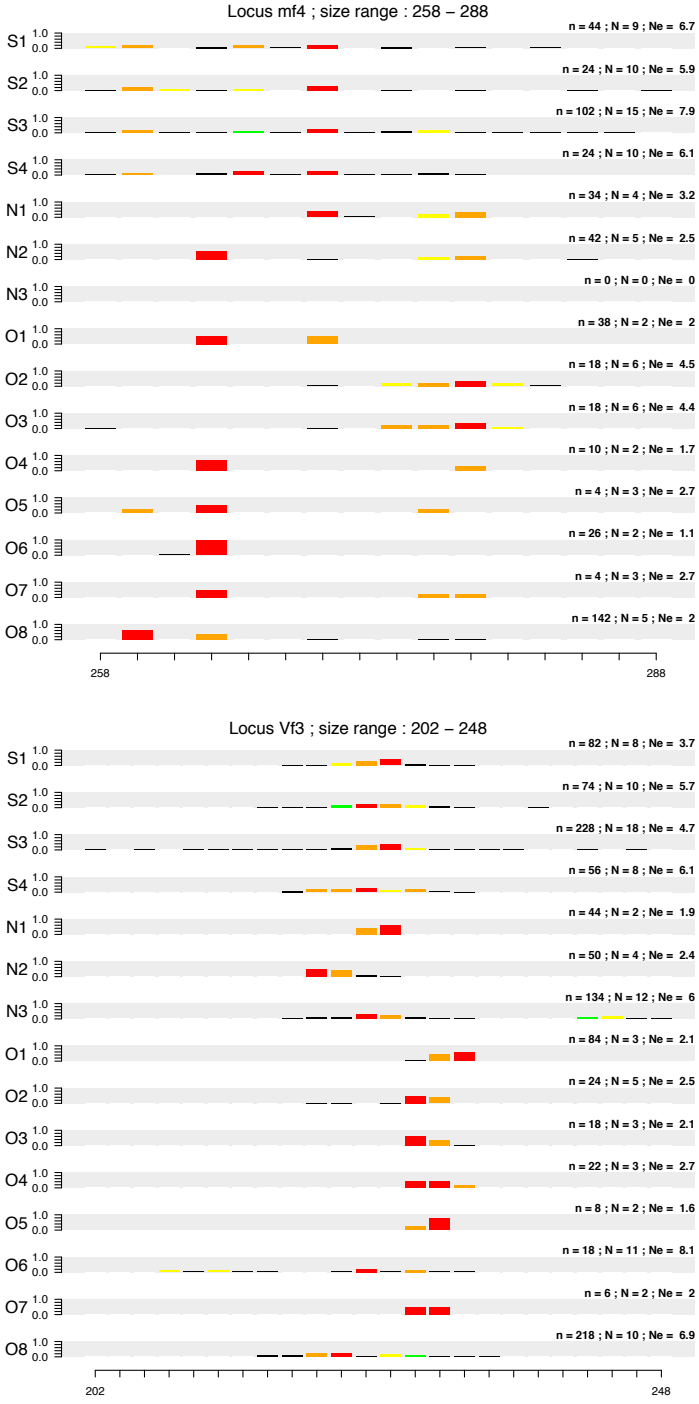


Figure 45. (suite) Fréquences alléliques dans les 16 populations étudiées pour les locus mf4 et Vf3. n : nombre d'allèles échantillonnés, N : nombre d'allèles différents dans la population, N_e : nombre efficace d'allèles. Les allèles sont colorés par ordre de fréquence décroissante : rouge pour l'allèle le plus fréquent, puis orange, jaune et vert.

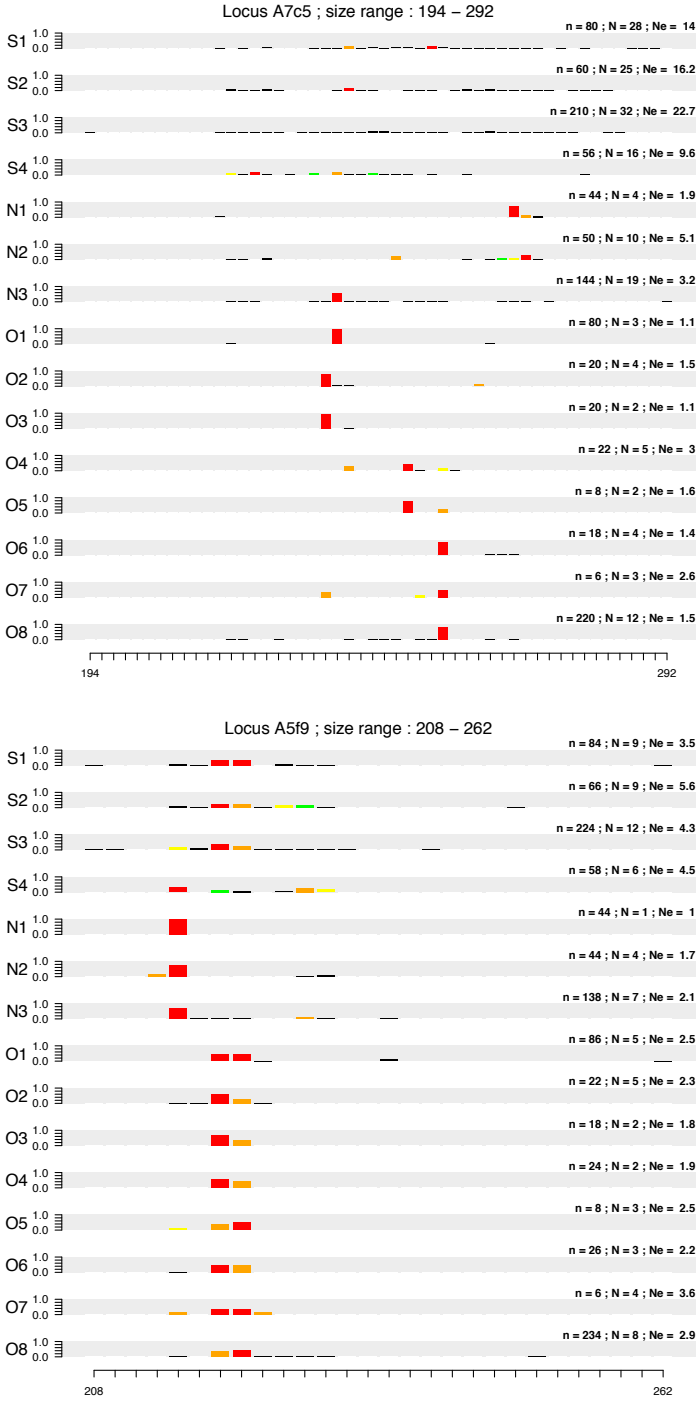


Figure 45. (suite) Fréquences alléliques dans les 16 populations étudiées pour les locus A7c5 et A5f9. n : nombre d'allèles échantillonnés, N : nombre d'allèles différents dans la population, N_e : nombre efficace d'allèles. Les allèles sont colorés par ordre de fréquence décroissante : rouge pour l'allèle le plus fréquent, puis orange, jaune et vert.

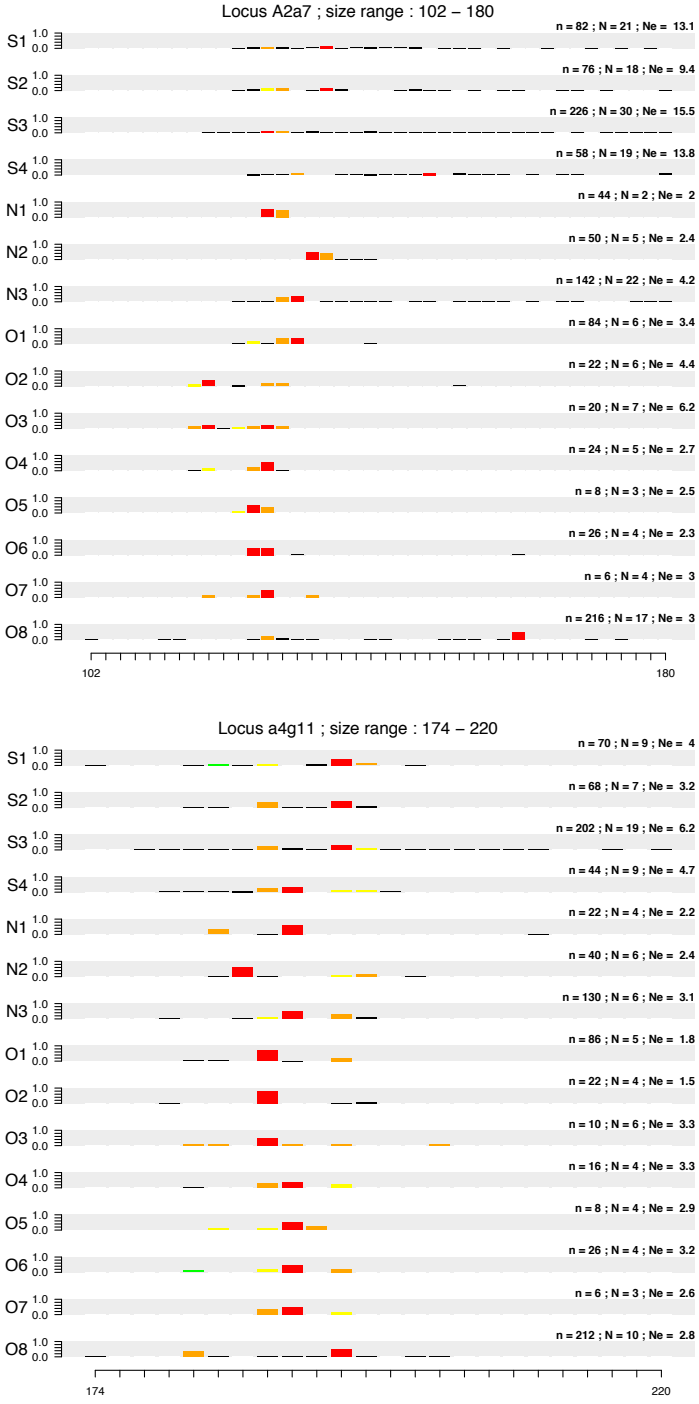


Figure 45. (suite) Fréquences alléliques dans les 16 populations étudiées pour les locus A2a7 et a4g11. n : nombre d'allèles échantillonnés, N : nombre d'allèles différents dans la population, N_e : nombre efficace d'allèles. Les allèles sont colorés par ordre de fréquence décroissante : rouge pour l'allèle le plus fréquent, puis orange, jaune et vert.

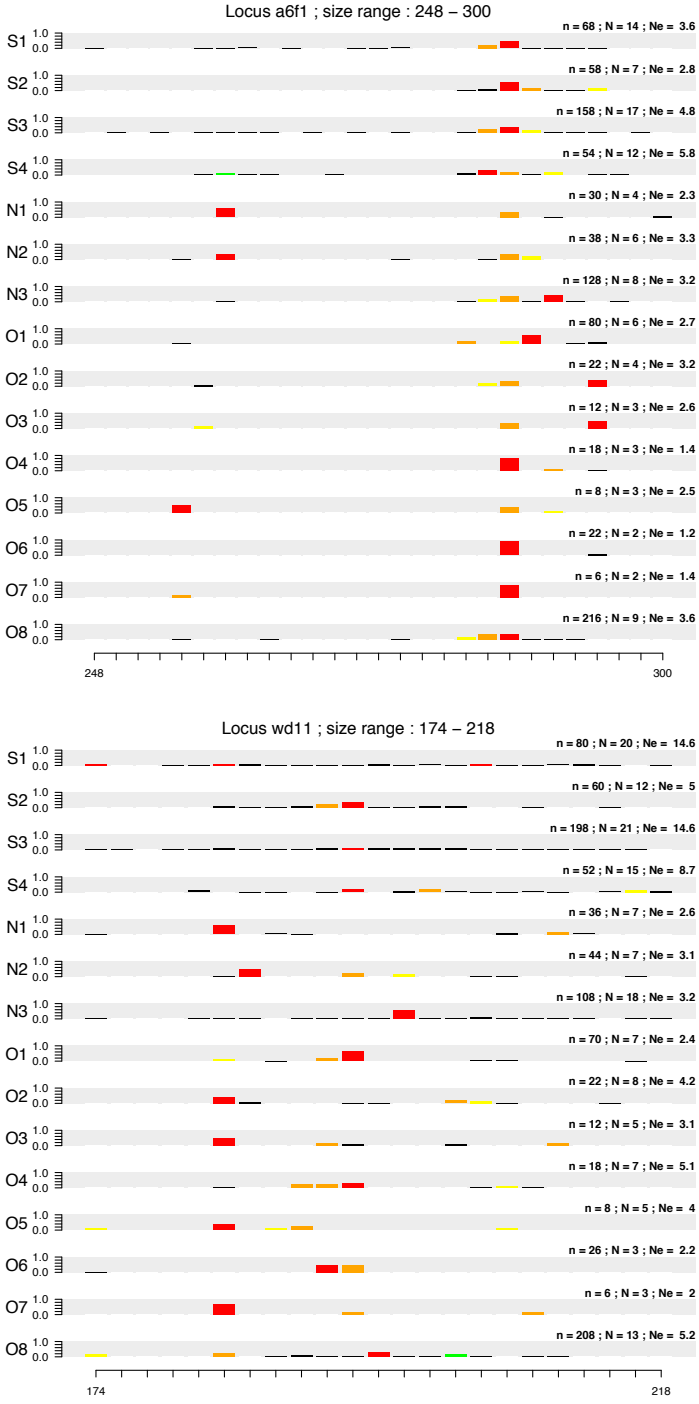


Figure 45. (suite) Fréquences alléliques dans les 16 populations étudiées pour les locus a6f1 et wd11. n : nombre d'allèles échantillonnés, N : nombre d'allèles différents dans la population, N_e : nombre efficace d'allèles. Les allèles sont colorés par ordre de fréquence décroissante : rouge pour l'allèle le plus fréquent, puis orange, jaune et vert.

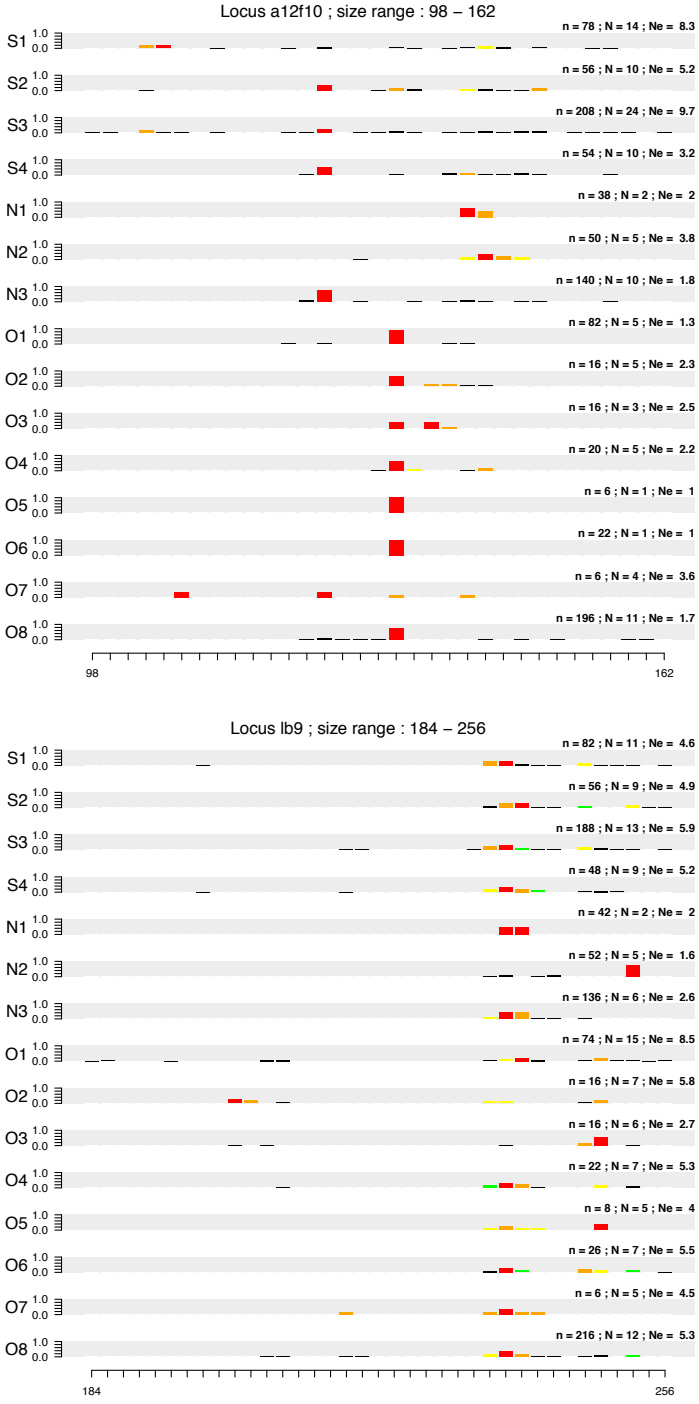


Figure 45. (suite) Fréquences alléliques dans les 16 populations étudiées pour les locus a12f10 et lb9. n : nombre d'allèles échantillonnés, N : nombre d'allèles différents dans la population, N_e : nombre efficace d'allèles. Les allèles sont colorés par ordre de fréquence décroissante : rouge pour l'allèle le plus fréquent, puis orange, jaune et vert.

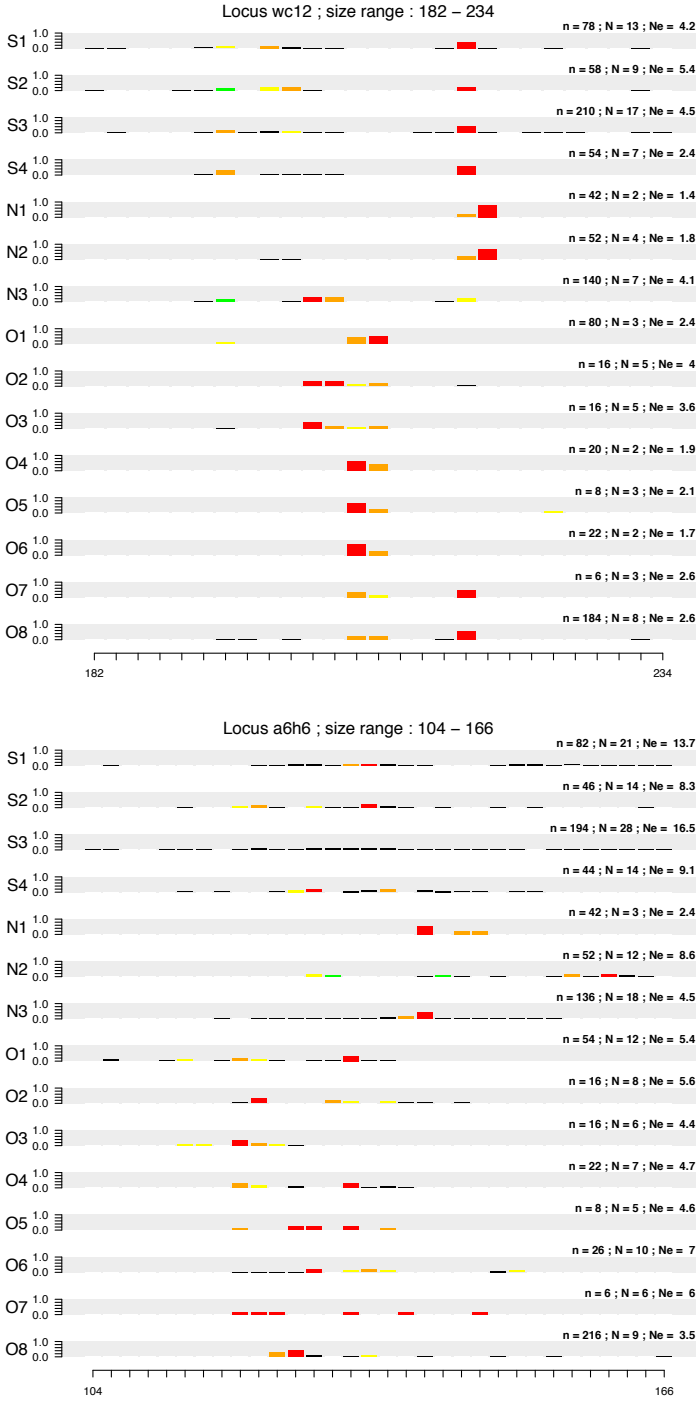


Figure 45. (suite) Fréquences alléliques dans les 16 populations étudiées pour les locus wc12 et a6h6. n : nombre d'allèles échantillonnés, N : nombre d'allèles différents dans la population, N_e : nombre efficace d'allèles. Les allèles sont colorés par ordre de fréquence décroissante : rouge pour l'allèle le plus fréquent, puis orange, jaune et vert.

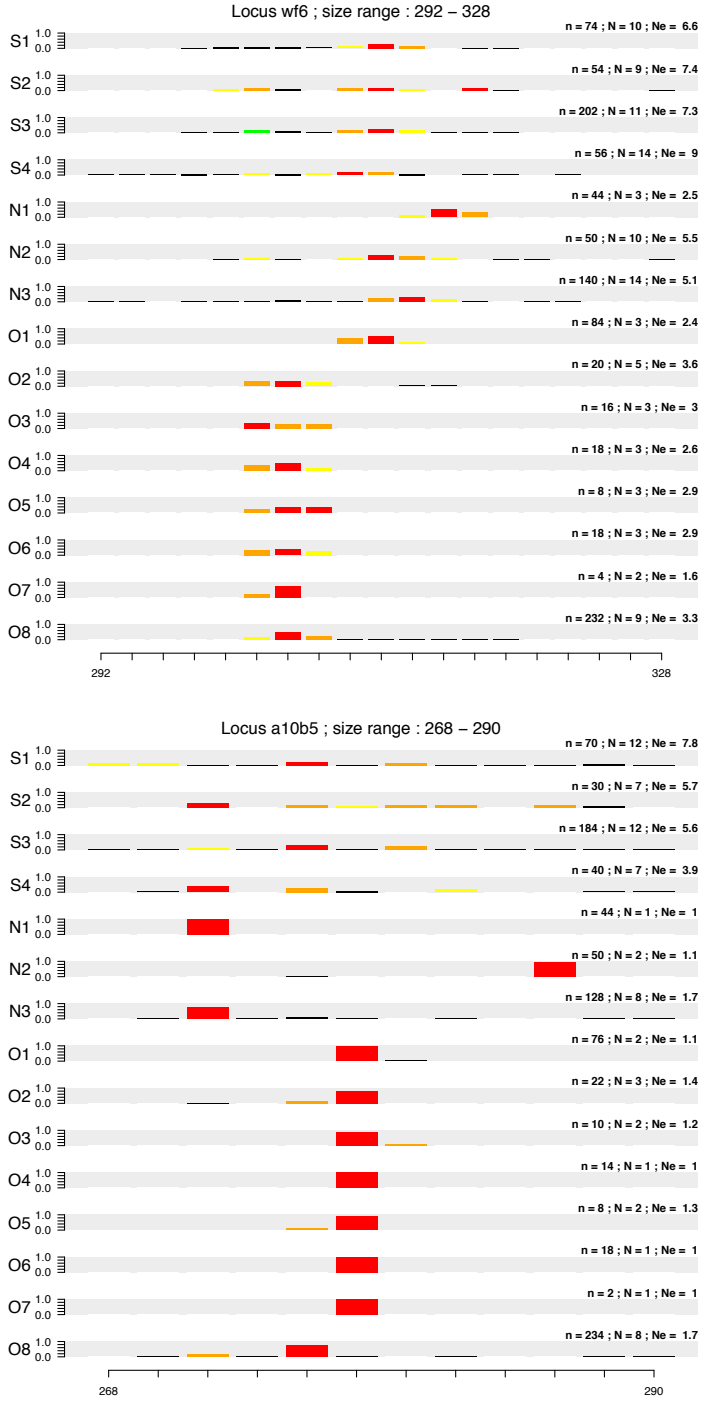


Figure 45. (suite) Fréquences alléliques dans les 16 populations étudiées pour les locus wf6 et a10b5. n : nombre d'allèles échantillonnés, N : nombre d'allèles différents dans la population, N_e : nombre efficace d'allèles. Les allèles sont colorés par ordre de fréquence décroissante : rouge pour l'allèle le plus fréquent, puis orange, jaune et vert.

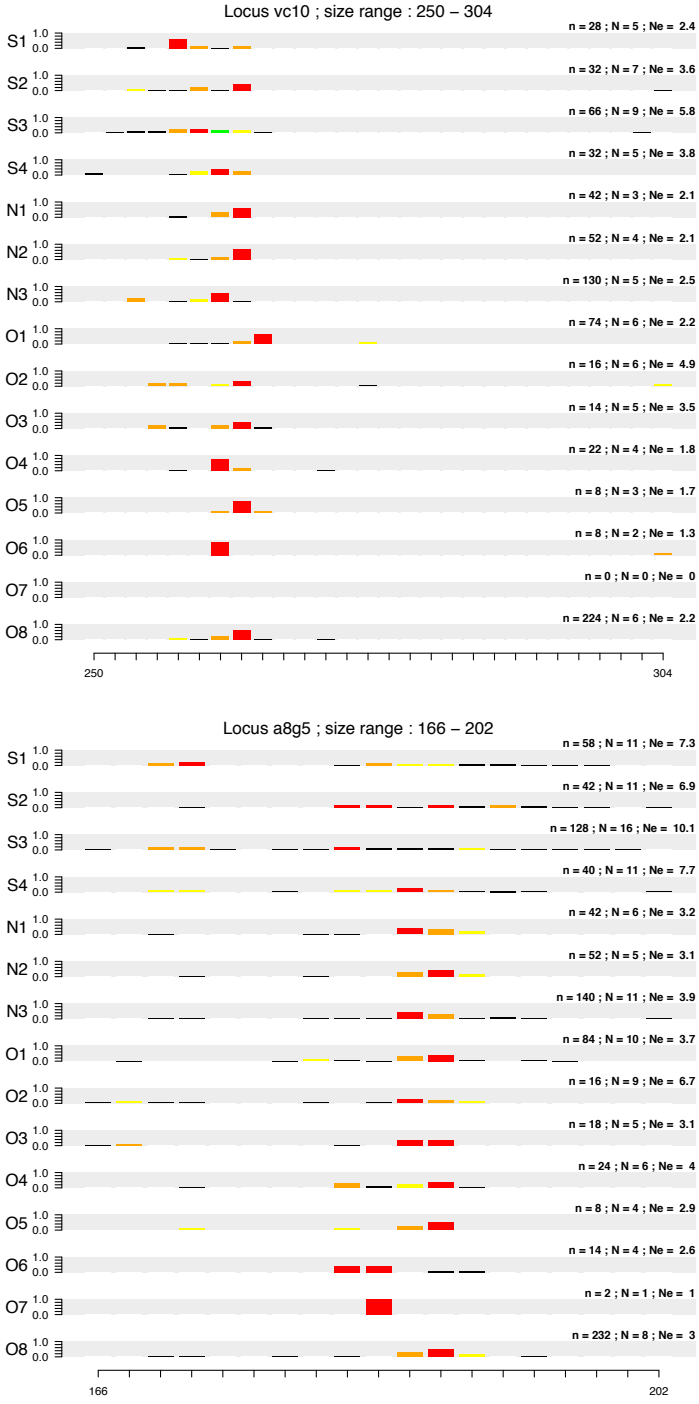


Figure 45. (suite) Fréquences alléliques dans les 16 populations étudiées pour les locus vc10 et a8g5. n : nombre d'allèles échantillonnés, N : nombre d'allèles différents dans la population, N_e : nombre efficace d'allèles. Les allèles sont colorés par ordre de fréquence décroissante : rouge pour l'allèle le plus fréquent, puis orange, jaune et vert.

4.1.2 *Estimation de l'âge des populations à partir de la diversité microsatellites*

Afin d'estimer l'âge des populations à partir des données microsatellites, nous avons utilisé la méthode introduite par Goldstein *et al.* [134] pour calculer la séparation entre populations humaines. Cette méthode a l'avantage d'être indépendante de la taille des populations. Elle permet ainsi de réaliser une datation en utilisant uniquement le taux de mutation observé dans les populations.

Nous avons choisi d'appliquer cette méthode à la population Pachón (O1) et au groupe de populations dans lequel se trouve la population de surface vivant à proximité de cette grotte (S3).

La première étape consiste à calculer la taille moyenne des fragments amplifiés à chaque locus dans chacune des deux populations. Cela est fait en utilisant la formule :

$$m = \sum i \times f_i \tag{9}$$

où f_i est la fréquence de l'allèle de taille i.

Les tailles moyennes des fragments amplifiés pour chaque locus dans les deux populations sont données Tableau 26.

On peut alors calculer le carré de la différence du nombre moyen de répétitions entre les deux populations grâce à la formule :

$$(\delta_\mu)^2 = \left(\frac{\mu_A - \mu_B}{l_r} \right)^2 \tag{10}$$

où μ_A est la taille moyenne du fragment dans la population A, μ_B dans la population B et l_r la taille d'une répétition.

Les $(\delta_\mu)^2$ calculés pour chaque locus sont présentés Tableau 26. On peut ensuite calculer le $(\delta_\mu)^2$ moyen sur tous les loci. À partir de cette valeur moyenne, il est possible de calculer le nombre de générations depuis la divergence des populations A et B. En effet, on a également la relation [134] :

$$(\delta_\mu)^2 = 2\mu \times t \tag{11}$$

où μ est le taux de mutation et t le nombre de générations.

On en déduit :

$$t = \frac{(\delta_\mu)^2}{2\mu} \tag{12}$$

On a aussi :

$$T = t \times G \tag{13}$$

où T est le temps en nombre d'années, t le temps en nombre de générations et G le temps de génération.

Locus	Taille moyenne d'un fragment		$(\delta\mu)^2$
	S3	O1	
A13e5	268,24	273	5,67
Hc5	180,29	173,07	13,02
Vf9	304,19	NA	NA
Ra8	152	150	1
Xa3	224,06	224,51	0,05
lb1	328,77	331,06	1,31
A13f8	137,76	141,07	2,74
A14d12	276,02	270,07	8,85
a19c8	226,48	229,40	2,13
a14d8	147,46	148,85	0,48
mf4	268,92	266,84	1,08
Vf3	224,61	231,05	10,38
A7c5	245,58	236,10	22,47
A5f9	220,43	222,91	1,54
A2a7	138,47	128,10	26,90
a4g11	191,90	188,95	2,17
a6f1	284,09	285,90	0,82
wd11	196,18	193,63	1,63
a12f10	128,52	132,20	3,38
lb9	239,12	233,03	9,27
wc12	206,31	205,78	0,07
a6h6	135,81	123,19	39,88
wf6	307,73	309,45	0,74
a10b5	278,60	278,08	0,07
vc10	260,15	266,14	8,95
a8g5	182,28	186	3,46
Moyenne			**6,72**

Tableau 26. Taille moyenne des fragments amplifiés pour les différents loci microsatellites étudiés pour la population de surface (S3) et la population de la grotte Pachón (O1) et carré de la différence du nombre moyen de répétitions entre ces deux populations ($(\delta_\mu)^2$)

.

Chez *Homo sapiens*, entre populations humaines africaines et non-africaines, la valeur moyenne de $(\delta_\mu)^2$ est 6,47 [134]. À partir de cette valeur et du taux de mutations ($5{,}6.10^{-4}$), le nombre de générations séparant Africains

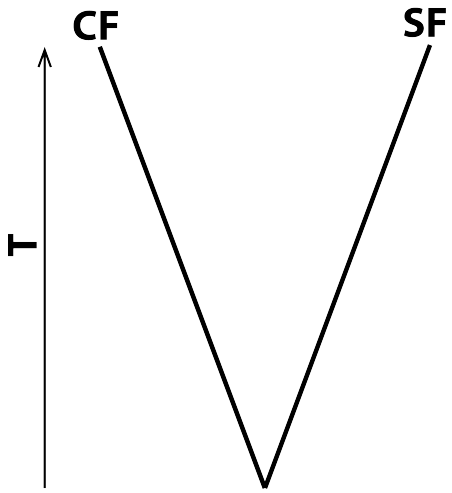


Figure 46. Le temps de divergence t en années est identique dans la population de surface et dans la population cavernicole.

et non-Africains a été estimé à 577,6. En tenant compte du temps de génération chez *Homo sapiens*, 27 ans, on trouve T = 156 000 ans [134]. Cette estimation est relativement proche des dernières réalisées avec d'autres méthodes [135].

Chez *Astyanax mexicanus*, nous avons trouvé un $(\delta_\mu)^2$ moyen de 6,72.

En considérant un taux de mutation de 5.10^{-4}, on obtient un temps de divergence de :

$$t = \frac{6,72}{2 * 5.10^{-4}} = 6\,720 \text{ générations.} \quad (14)$$

On obtient donc une divergence très récente, moins de 10 000 générations, de la population cavernicole Pachón en utilisant ce type de marqueur. Afin d'estimer le temps de divergence en années, nous avons tenu compte des différences de temps de génération des populations cavernicoles et de surface.

La même durée T en années s'est écoulée dans la population de surface et la population cavernicole (Figure 46). On a donc :

$$2t = t_{CF} + t_{SF} \quad (15)$$

où t_{CF} est le nombre de générations dans la population cavernicole et t_{SF} le nombre de générations dans la population de surface.

Le temps écoulé dans chaque branche dépend du temps de génération G_{CF} et G_{SF} des populations. On a donc :

$$T = t_{CF} \times G_{CF} = t_{SF} \times G_{SF} \quad (16)$$

(17)

et :

$$t_{CF} = \frac{t_{SF} \times G_{SF}}{G_{CF}} \tag{18}$$

On peut remplacer dans l'Équation 15 :

$$2t = \frac{t_{SF} \times G_{SF}}{G_{CF}} + t_{SF} \tag{19}$$

$$2t = t_{SF} \times \frac{G_{SF} + G_{CF}}{G_{CF}} \tag{20}$$

D'après l'Équation 20 et l'Équation 12, on obtient :

$$2 \times \frac{(\delta_\mu)^2}{2\mu} = t_{SF} \times \frac{G_{SF} + G_{CF}}{G_{CF}} \tag{21}$$

Au final on obtient :

$$t_{SF} = \frac{(\delta_\mu)^2}{\mu} \times \frac{G_{CF}}{G_{SF} + G_{CF}} \tag{22}$$

$$t_{CF} = \frac{(\delta_\mu)^2}{\mu} \times \frac{G_{SF}}{G_{SF} + G_{CF}} \tag{23}$$

On obtient le temps de divergence entre les deux populations en multipliant le nombre de générations dans une population par le temps de génération dans cette population Équation 17 :

$$T = \frac{(\delta_\mu)^2}{\mu} \times \frac{G_{CF}}{G_{SF} + G_{CF}} \times G_{SF} \tag{24}$$

$$T = \frac{(\delta_\mu)^2}{\mu} \times \frac{G_{SF}}{G_{SF} + G_{CF}} \times G_{CF} \tag{25}$$

En prenant comme temps de génération 5 ans pour la population Pachón et 2 ans pour la population de surface, on obtient :

$$T = \frac{6,72}{5.10^{-4}} \times \frac{5}{7} \times 2 = 19\ 205\ ans. \tag{26}$$

On obtient donc, avec les marqueurs microsatellites, un âge très proche de celui obtenu avec les SNP (25 500 ans). Ces deux estimations vont dans le même sens, celui d'une origine récente de la population Pachón.

4.2 Faible polymorphisme dans des gènes dispensables

Lorsqu'un gène n'est plus nécessaire, par exemple suite à un événement de duplication ou parce que la structure dans laquelle il est exprimé n'est plus

présente, les pressions de sélection sur ce gène sont relâchées. Le nombre de mutations observables sur ce gène pourra alors être grand, en particulier si les pressions de sélections sont relâchées depuis longtemps. On pourra alors observer un grand nombre de mutations dans ce gène et ces mutations pourront être synonymes, non-synonymes ou entraîner l'apparition d'un codon stop prématuré.

Nous nous sommes intéressés, chez *Astyanax mexicanus*, à des gènes spécifiques des yeux, qui devraient, si les populations sont anciennes, avoir accumulé de nombreuses mutations puisque ces gènes sont dispensables dans les grottes. Ces gènes sont les cristallines et les opsines.

Les cristallines spécifiques des yeux ont un rôle dans la maintien des propriétés optiques du cristallin [136].

Les cristallines appartiennent à trois familles : α, β et γ. Les cristallines de la famille α font partie de la grande famille des protéines de choc thermique (HSP, *Heat Shock Protein*) et ont donc une fonction proche de celles de chaperonnes. Les cristallines des familles β et γ sont plus proches de la famille des protéines AIM (Absent In Melanoma).

Onze séquences correspondant à des gènes de cristallines ont été retrouvées dans le transcriptome que nous avons généré (Chapitre 2). Dans ces 11 séquences nous avons pu identifier 29 SNP dont 16 synonymes et 13 non-synonymes [136].

Nous avons utilisé notre groupe externe *Hyphessobrycon anisitsi* afin d'orienter les mutations non-synonymes observées, et nous les avons classées (Tableau 27) dans les huit catégories définies précédemment (Tableau 1 Chapitre 2).

Cat.	Allèle SF	Allèle CF	Non-synonymes		Synonymes	
(1)	ancestral fixé	dérivé fixé	3	23 %	0	0 %
(2)	dérivé fixé	ancestral fixé	1	8 %	0	0 %
(3)	ancestral fixé	polymorphe	0	0 %	2	13 %
(4)	dérivé fixé	polymorphe	0	0 %	0	0 %
(5)	polymorphe	ancestral fixé	8	62 %	11	69 %
(6)	polymorphe	dérivé fixé	0	0 %	1	6 %
(7)	Polymorphisme partagé		1	7 %	2	13 %
(8)	Polymorphisme divergent		0	0 %	0	0 %
	Total		**13**	100	16	100

Tableau 27. Répartition dans les huit catégories définies Tableau 1 des SNP non-synonymes et synonymes identifiés dans les cristallines spécifiques de l'œil.

Comme observé dans l'ensemble du transcriptome (Chapitre 2), les poissons de surface sont plus polymorphes que les poissons cavernicoles alors que ces derniers ont fixé plus d'allèles dérivés. Le nombre de mutations

observées reste néanmoins très faible et aucun codon stop prématuré n'a été identifié. De plus, à l'exception d'une mutation observée chez les poissons de surface, les changements non-synonymes ne se font qu'entre acides aminés ayant des propriétés physico-chimiques proches [136].

Les opsines sont des protéines photosensibles que l'on retrouve dans des cellules spécialisées de la rétine : les cônes et les bâtonnets. Chaque opsine est sensible à une longueur d'onde spécifique.

Chez *Astyanax mexicanus*, trois gènes d'opsines (une rouge et deux vertes) ont été étudiés. Dans ces trois gènes, deux mutations non-synonymes et fixées ont été identifiées chez la population Pachón : une dans le gène codant la protéine *r007* (pigment rouge) et une dans le gène codant la protéine *g103* (pigment vert) [54]. Quelques mutations synonymes ont été identifiées dans ces mêmes gènes.

Chez une autre espèce cavernicole, *Sinocyclocheilus anshuiensis*, vivant sur le plateau Tibétain, en Chine, et divergeant depuis plusieurs millions d'années d'une espèce de surface proche *Sinocyclocheilus grahami*, de nombreux gènes de cristallines ne sont plus exprimés (24 sur 36 étudiés) [137]. Certains sont pseudogénéisés suite à l'apparition de codon stop prématurés [137]. Chez ce poisson, les opsines, d'autres protéines spécifiques des yeux, ont également un niveau d'expression beaucoup plus faible que chez l'espèce de surface.

Le nombre de mutations observées dans des gènes devenus dispensables dans l'environnement cavernicole, cristallines et opsines, est ainsi relativement faible. De plus aucun codon stop prématuré n'est présent. Ces résultats semblent contradictoires avec une origine ancienne de la population Pachón.

5 DISCUSSION

5.1 Âge de la population cavernicole de la grotte Pachón

L'adaptation à un nouvel environnement peut se faire principalement par la fixation d'allèles préexistants, provenant de la *standing genetic variation* ou par l'apparition et la fixation de mutation *de novo*. L'âge des populations est alors un paramètre important puisque, dans une population récente il n'y aura pas eu le temps suffisant pour que beaucoup de mutations *de novo* apparaissent, se fixent et permettent ainsi l'adaptation.

Le poisson *Astyanax mexicanus*, et en particulier le morphotype cavernicole, est un modèle très utilisé dans l'étude de l'adaptation à l'environnement. Et pourtant, l'âge des populations cavernicoles de cette espèce n'est pas connu précisément.

L'objet de cette thèse était donc d'estimer avec plus de précisions que les estimations actuelles, l'âge des populations cavernicoles de l'espèce *Astyanax mexicanus*, et en particulier celui de la grotte Pachón.

Dans une première partie (Chapitre 2), nous avons étudié le polymorphisme génétique entre une population de surface et la population cavernicole de la grotte Pachón. Nous avons vu que la population cavernicole présentait environ trois fois moins de polymorphismes que la population de surface, ce qui pourrait être expliqué par la plus petit taille de population dans les grottes. Parmi les sites polymorphes, environ 7% sont partagés entre ces deux populations. Ce niveau relativement élevé de polymorphisme partagé est explicable par une divergence récente ou par des flux migratoires entre les populations. Enfin, la population cavernicole a fixé environ deux fois plus d'allèles dérivés que la population de surface. Cette observation est, comme nous l'avons vu, explicable uniquement par une divergence récente des deux populations.

Dans une seconde partie (Chapitre 3), nous avons cherché à tester l'hypothèse de l'origine récente de la population Pachón en simulant des données de polymorphisme que nous avons ensuite comparées aux données observées. Lorsqu'un ajustement correct est trouvé entre les données observées et les données simulées le temps de divergence est toujours petit : moins de 30 000 ans.

Enfin, dans une dernière partie, nous avons réalisé une datation en utilisant un autre type de données, des marqueurs microsatellites provenant d'une publication et que nous avons réanalysé. L'estimation de l'âge de la population Pachón avec ces données est d'environ 20 000 ans.

Nous avons également étudié des gènes devenus dispensables dans l'environnement cavernicole. Dans des populations anciennes on attend qu'un

grand nombre de mutations apparaissent dans ces gènes en raison d'un relâchement de la pression de sélection qui s'exercent sur eux. Dans ces gènes, nous avons observé très peu de mutations. De plus, nous n'avons observé dans les population cavernicoles, aucune apparition de codons stop prématurés, de pseudogénéisation de ces gènes ou de mutations clairement délétères pour la fonction du gène.

5.2 Conséquence de cet âge sur l'origine du phénotype des populations cavernicoles d'*Astyanax mexicanus*

La population de la grotte Pachón semble donc avoir une origine récente, de l'ordre de quelques dizaines de milliers d'années. Cependant l'estimation obtenue avec les simulations réalisées au Chapitre 3 tout comme celle obtenue avec les marqueurs microsatellites au Chapitre 4 ont été volontairement surestimées. En effet, nous avons utilisé un temps de génération dans les populations de surface de deux ans au lieu de un an, valeur couramment citée dans la littérature pour d'autres espèces du genre *Astyanax*. Cette différence entraîne une sous-évaluation de la vitesse d'évolution de la population de surface et donc une surestimation du temps de divergence de la population de surface et de la population Pachón.

L'âge réel de la population de la grotte Pachón pourrait donc être bien inférieur à 30 000 ans et pourrait être de l'ordre de quelques milliers d'années.

Puisque l'origine de la population Pachón est récente, cela signifie que l'acquisition du phénotype cavernicole s'est fait en très peu de temps. Des mutations préexistantes à la colonisation de l'environnement cavernicole sont ainsi probablement à l'origine de ce phénotype. Les allèles impliqués étaient donc probablement présents dans la population de surface à basse fréquence. Ces allèles ont alors pu être fixés par dérive génétique et rapidement, en raison de la taille réduite des populations cavernicoles. Il est également possible qu'il y ait eu sélection positive de certains allèles délétères en surface, mais avantageux dans l'environnement cavernicole.

L'acquisition d'un si grand nombre de différences en si peu de temps pourrait paraître surprenant, mais la mise en place de modifications phénotypiques importantes en un temps court a été observée chez d'autres espèces de poissons.

5.2.1 *Évolution du poisson cavernicole de la Death Valley*

Le *Cyprinodon diabolis* est un cyprinodontidé vivant dans le Trou du Diable, une grotte située dans le désert de Mojave à cheval entre le Nevada et la Californie. Il vit isolé dans un bassin de 50m^2. La taille réelle de population de cette espèce varie entre 35 et 548 individus depuis qu'elle est surveillée depuis 1972 [138, 139].

La grotte n'est pas dans l'obscurité totale, puisqu'elle reçoit de la lumière du soleil durant toute l'année, sauf pendant 2 mois où les sources lumineuses

sont indirectes. La grotte est totalement isolée de la surface, ainsi il n'y a pas de flux de gènes d'espèces proches entrant dans la grotte [138].

Les poissons de cette espèce n'ont pas perdu leur pigmentation. Leurs yeux sont plus grands que chez d'autres espèces du même genre relativement à la taille du corps. Mais ils présentent d'autres modifications, certaines également observées dans les populations cavernicoles d'*Astyanax mexicanus* comme, par exemple, la réduction de l'agressivité.

Des premières études ont estimé, à partir de données mitochondriales, que cette population était isolée depuis 2 à 3 millions d'années [138]. Récemment, une étude utilisant des données de génomique a réévalué l'âge de la population en le réduisant de façon très impressionnante : entre 105 et 830 ans [138]. Une nouvelle étude a repoussé l'âge de la population à 10 000 à 60 000 ans [139].

5.2.2 *Cas de l'épinoche Gasterosteus aculeatus*

L'épinoche à trois épines (*Gasterosteus aculeatus*) est un poisson téléostéen vivant dans les zones tempérées et froides de l'hémisphère Nord. Des sous-espèces d'épinoches vivent dans des environnements variés. Certaines sous-espèces sont marines, mais peuvent également vivre en eau douce, alors que d'autres ne peuvent vivre qu'en eau douce. Les formes marines et non marines présentent des modifications morphologiques.

Le 27 mars 1964, la région d'Anchorage et de la Baie du Prince William en Alaska (États-Unis) est touchée par le séisme le plus important jamais enregistré en Amérique du Nord et le deuxième sur le globe (9,2 sur l'échelle de Richter). Ce séisme a entraîné des mouvements de subsidence* ou de soulèvement tectonique. Les îles de Montague, Middleton et Danger, à l'entrée de la Baie du Prince William ont ainsi été soulevées de plusieurs mètres. Ce soulèvement a entraîné l'isolement, en quelques minutes, de bassins qui étaient auparavant des habitats marins [140]. Les populations d'épinoches retrouvées dans ces bassins sont-elles issues de populations d'eau douce ou de populations marines ? Une étude [140] a montré que les populations d'épinoches d'eau douce retrouvées sur ces îles avaient une origine océanique. Ces populations ont évolué indépendamment des épinoches marines en une cinquantaine d'années. Des différences phénotypiques sont observées entre les populations marines et les populations d'eau douce, comme par exemple l'absence d'armure de plaques osseuses chez les épinoches d'eau douce. Ces différences ont des origines génétiques connues. Les différences phénotypiques sont donc dues à l'évolution et non pas à de la plasticité [140].

***Subsidence** : Affaissement de la lithosphère.*

5.2.3 *Les cichlidés du Lac Victoria*

Le Lac Victoria est un des plus grands lacs du monde, et le plus grand sur le continent Africain. Il constitue la source du Nil Blanc. Formé il y a environ 750 000 ans, le lac a été asséché à trois reprises et pour la dernière

fois il y a 17 000 ans. Il se serait ensuite de nouveau rempli il y a 14 000 ans [141, 142]. Et pourtant, il est possible d'observer plusieurs centaines d'espèces de cichlidés dans ce lac, qui auraient évoluées, depuis un même ancêtre commun, en seulement 14 000 ans [141].

5.3 Scénario évolutif des populations cavernicoles d'*Astyanax mexicanus*

Le scénario évolutif proposé jusqu'à présent repose sur l'existence de deux haplogroupes différents parmi les populations cavernicoles ainsi que parmi les populations de surface. Ces haplogroupes, A (Ia) et G (B, II), divergent depuis 1,8 à 4,5 millions d'années. En utilisant des séquences microsatellites la structuration des populations est à peu près similaire à celle obtenue avec les séquences mitochondriales. La population de la grotte Pachón n'ayant pas l'haplotype mitochondrial correspondant à ces séquences microsatellites, il a été proposé qu'une introgression d'ADN mitochondrial a eu lieu secondairement dans cette grotte.

Les résultats de cette thèse montrent que les populations cavernicoles sont récentes, et en particulier, pour la grotte Pachón de moins de 30 000 ans. Cet âge correspond, de façon intéressante, approximativement à la dernière période glaciaire, qui s'est terminée il y a environ 10 000 ans. L'études de spéléothèmes* a montré qu'il y a eu une période froide dans l'État de San Luis Potosi, où se trouve la Sierra de El Abra, il y a 55 000 à 20 000 ans, suivie d'une période chaude il y a 10 000 ans [143]. Durant cet épisode glaciaire, la calotte n'atteignait pas le Mexique, mais les températures y étaient bien plus faibles qu'actuellement.

***Spéléothème** : Dépôt minéral dans une grotte sous l'action de l'eau (par exemple, les stalagmites).*

Nous pouvons alors établir un nouveau scénario évolutif plus simple (Figure 47) des populations d'*Astyanax mexicanus* dans la région de la Sierra de El Abra. Des populations de poissons ont pu être séparées pendant une longue période conduisant à la mise en place des haplogroupes A et G. Lors de la dernière période glaciaire, le refroidissement aurait entraîné la migration de ces populations vers le Sud où elles auraient pu être en contact secondaire et ainsi se mélanger. À la fin de cette période, les populations auraient migré de nouveau vers le Nord (Figure 47). Des poissons appartenant à l'haplogroupe A ou G auraient alors pu être isolés dans les grottes. La faible fréquence de l'haplogroupe G dans la Sierra de El Abra pourrait expliquer pourquoi il n'a été, pour l'instant, retrouvé que dans une seule population à Rascón [52]. Les haplogroupes A et G sont retrouvés en sympatrie au Nord-Ouest du Mexique. Les deux haplogroupes sont donc retrouvés dans différentes régions du Mexique et parfois en sympatrie. Cette observation va dans le sens d'un contact secondaire récent des deux haplogroupes.

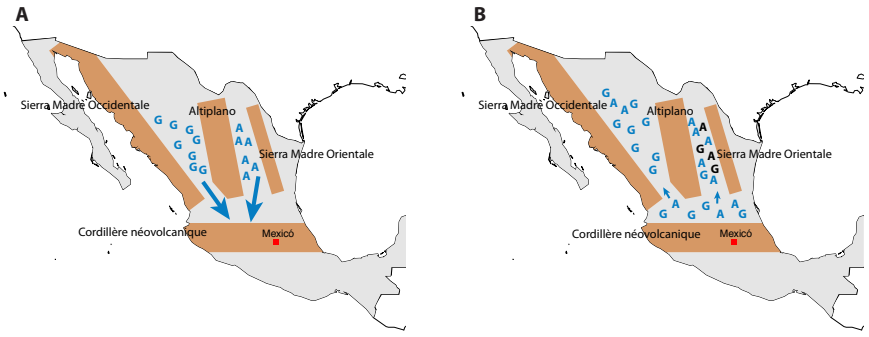


Figure 47. Schéma du scénario évolutif proposé. Deux populations ont été séparées pendant une longue période entraînant l'apparition de deux haplotypes (A et G). Ces deux haplotypes ont pu être en contact secondaire et se mélanger lors de la dernière période glaciaire pendant laquelle les populations ont pu migrer vers le Sud. À la fin de cette période glaciaire les populations auraient alors pu migrer de nouveau vers le Nord. Dans la Sierra de El Abra, des populations cavernicoles ont alors pu se mettre en place (en noir sur le schéma) et par le simple fait du hasard, ces populations ont pu fixer l'haplotype A ou l'haplotype G.

BIBLIOGRAPHIE

1. Gibert, J. & Daharveng, L. Subterranean Ecosystems : A Truncated Functional Biodiversity. *BioScience* **52,** 473–481 (2009).
2. Juan, C. *et al.* Evolution in caves : Darwin's 'wrecks of ancient life' in the molecular era. *Molecular ecology* **19,** 3865–3880 (2010).
3. Culver, D. C. & Holsinger, J. R. *Culver : How many species of troglobites are there* (National Speleological Society Bulletin, 1992).
4. Darwin, C. *On the Origin of the Species by Means of Natural Selection Or, The Preservation of Favoured Races in the Struggle for Life* 1859.
5. Romero, A & Green, S. M. The end of regressive evolution : examining and interpreting the evidence from cave fishes. *Journal of Fish Biology* **67,** 3–32 (2005).
6. Mayr, E. Cause and effect in biology. *Science (New York, N.Y.)* **134,** 1501–1506 (1961).
7. Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge University Press, 1984).
8. Aubry, M.-P. in *Encyclopedia of Scientific Dating Methods* 1–35 (Springer Netherlands, Dordrecht, 2014).
9. Andreu-Hayles, L. & Leland, C. in *Encyclopedia of Scientific Dating Methods* 1–12 (Springer Netherlands, Dordrecht, 2014).
10. Zuckerkandl, E & Pauling, L. Evolutionary divergence and convergence in proteins. *Evolving genes and proteins* (1965).
11. Welch, J & Bromham, L. Molecular dating when rates vary. *Trends in ecology & evolution* **20,** 320–327 (2005).
12. Jukes, T. H. & Cantor, C. R. in *Mammalian Protein Metabolism* 21–132 (Elsevier, 1969).
13. Tavaré, S. Some Mathematical Questions in Biology : DNA Sequence Analysis. *Lectures on mathematics in the life sciences* (1986).
14. Romero, A. Scientists Prefer them Blind : The History of Hypogean Fish Research. *Environmental biology of fishes* **62,** 43–71 (2001).
15. Romero, A. *The biology of hypogean fishes* (Springer Science & Business Media, 2013).
16. Gross, J. B. The complex origin of Astyanax cavefish. *BMC evolutionary biology* **12,** 105 (2012).
17. Chakraborty, R & Nei, M. Dynamics of gene differentiation between incompletely isolated populations of unequal sizes. *Theoretical population biology* **5,** 460–469 (1974).

18. Avise, J. C. & Selander, R. K. Evolutionary Genetics of Cave-Dwelling Fishes of the Genus Astyanax. *Evolution* **26,** 1–19 (1972).

19. Rétaux, S., Pottin, K. & Alunni, A. Shh and forebrain evolution in the blind cavefish Astyanax mexicanus. *Biology of the cell* **100,** 139–147 (2008).

20. Menuet, A. *et al.* Expanded expression of Sonic Hedgehog in Astyanax cavefish : multiple consequences on forebrain development and evolution. *Development* **134,** 845–855 (2007).

21. Jeffery, W. R. Adaptive Evolution of Eye Degeneration in the Mexican Blind Cavefish. *The Journal of heredity* **96,** 185–196 (2005).

22. Yamamoto, Y., Stock, D. W. & Jeffery, W. R. Hedgehog signalling controls eye degeneration in blind cavefish. *Nature* **431,** 844–847 (2004).

23. Jeffery, W. R., Strickler, A. G. & Yamamoto, Y. To see or not to see : evolution of eye degeneration in mexican blind cavefish. *Integrative and comparative biology* **43,** 531–541 (2003).

24. Jeffery, W. R. Cavefish as a Model System in Evolutionary Developmental Biology. *Developmental biology* **231,** 1–12 (2001).

25. Porter, M. L., Dittmar, K & Pérez-Losada, M. How long does evolution of the troglomorphic form take ? Estimating divergence times in Astyanax mexicanus. *Acta Carsologica,* 173–182 (2007).

26. Kasumyan, A. O. & Marusov, E. A. Chemoorientation in the feeding behavior of the blind Mexican cavefish Astyanax fasciatus (Characidae, Teleostei). *Russian Journal of Ecology* **46,** 559–563 (2015).

27. Bradic, M., Teotonio, H. & Borowsky, R. L. The Population Genomics of Repeated Evolution in the Blind Cavefish Astyanax mexicanus. *Molecular biology and evolution* **30,** 2383–2400 (2013).

28. Beale, A. *et al.* Circadian rhythms in Mexican blind cavefish Astyanax mexicanus in the lab and in the field. *Nature Communications* **4,** 2769–2769 (2012).

29. Hüppop, K. & Wilkens, H. Bigger eggs in subterranean Astyanax fasciatus (Characidae, Pisces). *Journal of Zoological Systematics and Evolutionary Research* **29,** 280–288 (1991).

30. Elipot, Y. *et al.* Astyanax transgenesis and husbandry : how cavefish enters the laboratory. *Zebrafish* **11,** 291–299 (2014).

31. Gross, J. B. *et al.* Natural bone fragmentation in the blind cavedwelling fish, Astyanax mexicanus : candidate gene identification through integrative comparative genomics. *Evolution & development* **18,** 7–18 (2016).

32. Caballero-Hernández, O. *et al.* Circadian rhythms and photic entrainment of swimming activity in cave-dwelling fish Astyanax mexicanus(Actinopterygii : Characidae), from El Sotano La Tinaja, San Luis Potosi, Mexico. *Biological Rhythm Research* **46,** 579–586 (2015).

33. Carlson, B. M., Onusko, S. W. & Gross, J. B. A High-Density Linkage Map for Astyanax mexicanus Using Genotyping-by-Sequencing Technology. *G3 Genes|Genomes|Genetics* **5,** 241–251 (2015).

34. Yoshizawa, M. Behaviors of cavefish offer insight into developmental evolution. *Molecular Reproduction and Development* **82,** 268–280 (2015).

35. McGaugh, S. E. *et al.* The cavefish genome reveals candidate genes for eye loss. *Nature Communications* **5,** 5307 (2014).

36. Gross, J. B., Krutzler, A. J. & Carlson, B. M. Complex craniofacial changes in blind cave-dwelling fish are mediated by genetically symmetric and asymmetric loci. *Genetics* **196,** 1303–1319 (2014).

37. Yoshizawa, M. *et al.* The sensitivity of lateral line receptors and their role in the behavior of Mexican blind cavefish (Astyanax mexicanus). *The Journal of Experimental Biology* **217,** 886–895 (2014).

38. Gross, J. B. & Wilkens, H. Albinism in phylogenetically and geographically distinct populations of Astyanax cavefish arises through the same loss-of-function Oca2 allele. *Heredity* **111,** 122–130 (2013).

39. Atukorala, A. D. S. *et al.* Adaptive evolution of the lower jaw dentition in Mexican tetra (Astyanax mexicanus). *EvoDevo* **4,** 28 (2013).

40. Borowsky, R. L. & Cohen, D. Genomic Consequences of Ecological Speciation in Astyanax Cavefish. *PloS one* **8,** e79903 (2013).

41. Dufton, M., Hall, B. K. & Franz-Odendaal, T. A. Early Lens Ablation Causes Dramatic Long-Term Effects on the Shape of Bones in the Craniofacial Skeleton of Astyanax mexicanus. *PloS one* **7,** e50308– (2012).

42. Yoshizawa, M., Ashida, G. & Jeffery, W. R. Parental genetic effects in a cavefish adaptive behavior explain disparity between nuclear and mitochondrial DNA. *Evolution* **66,** 2975–2982 (2012).

43. Gallo, N. D. & Jeffery, W. R. Evolution of Space Dependent Growth in the Teleost Astyanax mexicanus. *PloS one* **7,** e41443 (2012).

44. Yoshizawa, M. *et al.* Evolution of an adaptive behavior and its sensory receptors promotes eye regression in blind cavefish. *BMC biology* **10,** 108 (2012).

45. Gallo, N. D. & Jeffery, W. R. Evolution of space dependent growth in the teleost Astyanax mexicanus. *PloS one* **7,** e41443 (2012).

46. Cavallari, N. *et al.* A blind circadian clock in cavefish reveals that opsins mediate peripheral clock photoreception. *PLoS biology* **9,** e1001142 (2011).

47. Yoshizawa, M. & Jeffery, W. R. Evolutionary tuning of an adaptive behavior requires enhancement of the neuromast sensory system. *Communicative & integrative biology* **4,** 89–91 (2011).

48. Jeffery, W. R. Regressive Evolution in *Astyanax* Cavefish. *Annual Review of Genetics* **43,** 25–47 (2009).

49. Jeffery, W. R. Chapter 8 Evolution and Development in the Cavefish Astyanax. *Current Topics in Developmental Biology,* 191–221 (2009).

50. Protas, M. *et al.* Multi-trait evolution in a cave fish, Astyanax mexicanus. *Evolution & development* **10,** 196–209 (2008).

51. Yoshizawa, M. & Jeffery, W. R. Shadow response in the blind cavefish Astyanax reveals conservation of a functional pineal eye. *The Journal of Experimental Biology* **211,** 292–299 (2008).

52. Ornelas-García, C. P., 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* **8,** 340 (2008).

53. Yamamoto, Y. & Jeffery, W. R. Central Role for the Lens in Cave Fish Eye Degeneration. *Science (New York, N.Y.)* **289,** 631–633 (2000).

54. Yokoyama, S *et al.* Initial mutational steps toward loss of opsin gene function in cavefish. *Molecular biology and evolution* **12,** 527–532 (1995).

55. Alunni, A. *et al.* Developmental mechanisms for retinal degeneration in the blind cavefish Astyanax mexicanus. *The Journal of comparative neurology* **505,** 221–233 (2007).

56. Pottin, K., Hinaux, H. & Rétaux, S. Restoring eye size in Astyanax mexicanus blind cavefish embryos through modulation of the Shh and Fgf8 forebrain organising centres. *Development* **138,** 2467–2476 (2011).

57. Hinaux, H. *et al.* A developmental staging table for Astyanax mexicanus surface fish and Pachón cavefish. *Zebrafish* **8,** 155–165 (2011).

58. Hinaux, H. *et al.* De Novo Sequencing of Astyanax mexicanus Surface Fish and Pachón Cavefish Transcriptomes Reveals Enrichment of Mutations in Cavefish Putative Eye Genes. *PloS one* **8,** e53553 (2013).

59. Elipot, Y. *et al.* Evolutionary shift from fighting to foraging in blind cavefish through changes in the serotonin network. *Current biology : CB* **23,** 1–10 (2013).

60. Elipot, Y. *et al.* A mutation in the enzyme monoamine oxidase explains part of the Astyanax cavefish behavioural syndrome. *Nature Communications* **5,** 3647 (2014).

61. Protas, M. E. *et al.* Genetic analysis of cavefish reveals molecular convergence in the evolution of albinism. *Nature genetics* **38,** 107–111 (2006).

62. Gross, J. B., Borowsky, R. L. & Tabin, C. J. A novel role for Mc1r in the parallel evolution of depigmentation in independent populations of the cavefish Astyanax mexicanus. *PLoS genetics* **5,** e1000326 (2009).

63. Schiöth, H. B. *et al.* Loss of function mutations of the human melanocortin 1 receptor are common and are associated with red hair. *Biochemical and biophysical research communications* **260,** 488–491 (1999).

64. Jeffery, W. R. & Martasian, D. P. Evolution of Eye Regression in the Cavefish Astyanax : Apoptosis and the Pax-6 Gene. *American Zoologist* **38,** 685–696 (1998).

65. Rétaux, S. & Casane, D. Evolution of eye development in the darkness of caves : adaptation, drift, or both ? *EvoDevo* **4,** 26 (2013).

66. Mitchell, R. W., Russel, W. H. & Elliott, W. R. *Mexican Eyeless Characin Fishes, Genus Astyanax : Environment, Distribution, and Evolution* (Special Publication The Museum Texas Tech University, 1977).

67. Hüppop, K. Food-finding ability in cave fish (Astyanax fasciatus). *International Journal of Speleology* **16,** 59–66 (1987).

68. Rose, F. L. & Mitchell, R. W. Comparative Lipid Values of Epigean and Cave-Adapted Astyanax. *The Southwestern Naturalist* **27,** 357–358 (1982).

69. Aspiras, A. C. *et al.* Melanocortin 4 receptor mutations contribute to the adaptation of cavefish to nutrient-poor conditions. *Proceedings of the National Academy of Sciences* **112,** 9668–9673 (2015).

70. Ames, A *et al.* Energy metabolism of rabbit retina as related to function : high cost of Na+ transport. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **12,** 840–853 (1992).

71. Moran, D., Softley, R. & Warrant, E. J. The energetic cost of vision and the evolution of eyeless Mexican cavefish. *Science advances* **1,** e1500363 (2015).

72. Moran, D., Softley, R. & Warrant, E. J. Eyeless Mexican cavefish save energy by eliminating the circadian rhythm in metabolism. *PloS one* **9,** e107877 (2014).

73. Hara, T. J. *Fish Chemoreception* (ed Hara, T. J.) (Springer Science & Business Media, Dordrecht, 2012).

74. Vieira, F. G. & Rozas, J. Comparative Genomics of the Odorant-Binding and Chemosensory Protein Gene Families across the Arthropoda : Origin and Evolutionary History of the Chemosensory System. *Genome biology and evolution* **3,** 476–490 (2011).

75. Protas, M. *et al.* Multi-trait evolution in a cave fish, Astyanax mexicanus. *Evolution & development* **10,** 196–209 (2008).

76. Bibliowicz, J. *et al.* Differences in chemosensory response between eyed and eyeless Astyanax mexicanus of the Rio Subterráneo cave. *EvoDevo* **4,** 25 (2013).

77. Hinaux, H. *et al.* Sensory evolution in blind cavefish is driven by early embryonic events during gastrulation and neurulation. *Development* **143,** 4521–4532 (Nov. 2016).

78. Schemmel, D. Vergleichende Untersuchungen an den Hautsinnesorganen ober-und unterirdisch lebender Astyanax-Formen. *Zeitschrift für Morphologie der Tiere* **61,** 255–316 (1967).

79. Varatharasan, N., Croll, R. P. & Franz-Odendaal, T. Taste bud development and patterning in sighted and blind morphs of Astyanax mexicanus. *Developmental dynamics* **238,** 3056–3064 (2009).

80. Montgomery, J. C., Coombs, S. & Baker, C. F. in *The biology of hypogean fishes* 87–96 (Springer Netherlands, Dordrecht, 2001).

81. Ghysen, A. & Dambly-Chaudière, C. Development of the zebrafish lateral line. *Current Opinion in Neurobiology* **14,** 67–73 (2004).

82. Sumi, K. *et al.* Innervation of the lateral line system in the blind cavefish Astyanax mexicanus (Characidae) and comparisons with the eyed surface-dwelling form. *Ichthyological Research* **62,** 420–430 (2015).

83. Teyke, T. Morphological Differences in Neuromasts of the Blind Cave Fish Astyanax hubbsi and the Sighted River Fish Astyanax mexicanus. *Brain, Behavior and Evolution* **35,** 23–30 (1990).

84. Yoshizawa, M. *et al.* Evolution of a behavioral shift mediated by superficial neuromasts helps cavefish find food in darkness. *Current biology : CB* **20,** 1631–1636 (2010).

85. Lang, H. H. Surface wave discrimination between prey and nonprey by the back swimmerNotonecta glauca L. (Hemiptera, Heteroptera). *Behavioral Ecology and Sociobiology* **6,** 233–246 (1980).

86. Holzman, R., Perkol-Finkel, S. & Zilman, G. Mexican blind cavefish use mouth suction to detect obstacles. *The Journal of Experimental Biology* **217,** 1955–1962 (2014).

87. Schemmel, C. Studies on the Genetics of Feeding Behaviour in the Cave Fish Astyanax mexicanus f. anoptichthys. *Zeitschrift für Tierpsychologie* **53,** 9–22 (1980).

88. Yamamoto, Y. *et al.* Pleiotropic functions of embryonic sonic hedgehog expression link jaw and taste bud amplification with eye loss during cavefish evolution. *Developmental biology* **330,** 200–211 (2009).

89. Kowalko, J. E. *et al.* Loss of schooling behavior in cavefish through sight-dependent and sight-independent mechanisms. *Current biology : CB* **23,** 1874–1883 (2013).

90. Camargo, J. The middle Cretaceous El Abra Limestone at its type locality (facies, diagenesis and oil emplacement), east-central Mexico. *Revista Mexicana de Ciencias Geológicas* **15,** 1–15 (1998).

91. Elliott, W. R. in *Biology and Evolution of the Mexican Cavefish* (Academic Press, 2015).

92. Kirby, M. X., Jones, D. S. & MacFadden, B. J. Lower Miocene Stratigraphy along the Panama Canal and Its Bearing on the Central American Peninsula. *PloS one* **3,** e2791 (2008).

93. Bacon, C. D. *et al.* Biological evidence supports an early and complex emergence of the Isthmus of Panama. *Proceedings of the National Academy of Sciences* **112,** 6110–6115 (2015).

94. Burton, K. W., Ling, H.-F. & O'Nions, R. K. Closure of the Central American Isthmus and its effect on deep-water formation in the North Atlantic. *Nature* **386,** 382–385 (1997).

95. Haug, G. H. & Tiedemann, R. Effect of the formation of the Isthmus of Panama on Atlantic Ocean thermohaline circulation. *Nature* **393,** 673–676 (1998).

96. Lessios, H. A. The Great American Schism : Divergence of Marine Organisms After the Rise of the Central American Isthmus. *Annual Review of Ecology, Evolution, and Systematics* **39,** 63–91 (2008).

97. Udvardy, M. D. F. *A Classification of the Biogeographical Provinces of the World* 1975.

98. Ortí, G. & Meyer, A. The Radiation of Characiform Fishes and the Limits of Resolution of Mitochondrial Ribosomal DNA Sequences. *Systematic Biology* **46,** 75–100 (1997).

99. Keene, A., Yoshizawa, M. & McGaugh, S. E. *Biology and Evolution of the Mexican Cavefish* 1st ed. (Elsevier Inc., 2016).

100. Hubbs, C. L. & Innes, W. T. The first known blind fish of the family Characidae : a new genus from Mexico. *Occasional Papers of the Museum of Zoology University of Michigan,* 1–10 (1936).

101. Alvarez, J. *Revisión del género Anoptichthys con descripción de una especie nueva (Pisc., Characidae)* (An Esc Nac Cienc Biol Mex, 1946).

102. Alvarez, J. *Descripción de Anoptichthys hubbsi caracınido ciego de la cueva de los Sabinos* (SLP Rev Soc Mex Hist Nat, 1947).

103. Dowling, T. E., Martasian, D. P. & Jeffery, W. R. Evidence for multiple genetic forms with similar eyeless phenotypes in the blind cavefish, Astyanax mexicanus. *Molecular biology and evolution* **19,** 446–455 (2002).

104. Strecker, U, Bernatchez, L. & Wilkens, H. Genetic divergence between cave and surface populations of Astyanax in Mexico (Characidae, Teleostei). *Molecular ecology* **12,** 699–710 (2003).

105. Strecker, U., Faúndez, V. H. & Wilkens, H. Phylogeography of surface and cave Astyanax (Teleostei) from Central and North America based on cytochrome b sequence data. *Molecular phylogenetics and evolution* **33,** 469–481 (2004).

106. Near, T. J. *et al.* Resolution of ray-finned fish phylogeny and timing of diversification. *Proceedings of the National Academy of Sciences of the United States of America* **109,** 13698–13703 (2012).

107. Javonillo, R. *et al.* Relationships among major lineages of characid fishes (Teleostei : Ostariophysi : Characiformes), based on molecular sequence data. *Molecular phylogenetics and evolution* **54,** 498–511 (2010).

108. Mariguela, T. C. *et al.* Phylogeny and biogeography of Triportheidae (Teleostei : Characiformes) based on molecular data. *Molecular Phylogenetics and Evolution* **96,** 130–139 (2016).

109. Vidal, N. & Hedges, S. B. The phylogeny of squamate reptiles (lizards, snakes, and amphisbaenians) inferred from nine nuclear protein-coding genes. *Comptes rendus biologies* **328,** 1000–1008 (2005).

110. Saitou, N & Nei, M. The neighbor-joining method : a new method for reconstructing phylogenetic trees. *Molecular biology and evolution* **4,** 406–425 (1987).

111. Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution* **16,** 111–120 (1980).

112. Kumar, S., Stecher, G. & Tamura, K. MEGA7 : Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular biology and evolution* **33,** 1870–1874 (2016).

113. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics (Oxford, England)* **25,** 1754–1760 (2009).

114. McKenna, A. *et al.* The Genome Analysis Toolkit : a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20,** 1297–1303 (2010).

115. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25,** 2078–2079 (2009).

116. Altschul, S. F. *et al.* Basic local alignment search tool. *Journal of molecular biology* **215,** 403–410 (1990).

117. Flicek, P. *et al.* Ensembl 2014. *Nucleic acids research* **42,** gkt1196–D755 (2013).

118. Kasprzyk, A. BioMart : driving a paradigm change in biological data management. *Database : the journal of biological databases and curation* **2011,** bar049 (2011).

119. Grantham, R. Amino acid difference formula to help explain protein evolution. *Science (New York, N.Y.)* **185,** 862–864 (1974).

120. Li, W. H., Wu, C. I. & Luo, C. C. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Molecular biology and evolution* **2,** 150–174 (1985).

121. Bradic, M. *et al.* Gene flow and population structure in the Mexican blind cavefish complex (*Astyanax mexicanus*). *BMC evolutionary biology* **12,** 9 (2012).

122. Kingman, J. F. C. The coalescent. *Stochastic Processes and their Applications* **13,** 235–248 (1982).

123. Hoban, S., Bertorelle, G. & Gaggiotti, O. E. Computer simulations : tools for population and evolutionary genetics. *Nature Reviews Genetics* **13,** 110–122 (2012).

124. Charlesworth, B. & Charlesworth, D. *Elements of evolutionary genetics* (Greenwood Village, Colo. : Roberts and Co. Publishers, 2010).

125. Lynch, M. Evolution of the mutation rate. *Trends in genetics : TIG* **26,** 345–352 (2010).

126. Lacey, E. A. *Life Underground* (University of Chicago Press, 2000).

127. Voituron, Y. *et al.* Extreme lifespan of the human fish (Proteus anguinus) : a challenge for ageing mechanisms. *Biology letters* **7,** 105–107 (2011).

128. Winemiller, K. O. Patterns of variation in life history among South American fishes in seasonal environments. *Oecologia* **81,** 225–241 (1989).

129. Kernighan, B. W. & Ritchie, D. M. *The C Programming Language* (1988).

130. Panaram, K. & Borowsky, R. L. Gene Flow and Genetic Variability in Cave and Surface Populations of the Mexican Tetra, *Astyanax mexicanus* (Teleostei : Characidae). *Copeia* **2005,** 409–416 (2005).

131. Weber, J. L. & Wong, C. Mutation of human short tandem repeats. *Human molecular genetics* **2,** 1123–1128 (1993).

132. Bowcock, A. M. *et al.* High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* **368,** 455–457 (1994).

133. Crow, J. F. & Kimura, M. Population Genetics. (Book Reviews : An Introduction to Population Genetics Theory). *Science (New York, N.Y.)* **171,** 666–667 (1971).

134. Goldstein, D. B. *et al.* Genetic absolute dating based on microsatellites and the origin of modern humans. *Proceedings of the National Academy of Sciences* **92,** 6723–6727 (1995).

135. Shriner, D. *et al.* Genome-wide genotype and sequence-based reconstruction of the 140,000 year history of modern human ancestry. *Scientific reports* **4,** 6055 (2014).

136. Hinaux, H. *et al.* Lens defects in Astyanax mexicanus Cavefish : evolution of crystallins and a role for alphaA-crystallin. *Developmental Neurobiology* **75,** 505–521 (2015).

137. Yang, J. *et al.* The Sinocyclocheilus cavefish genome provides insights into cave adaptation. *BMC biology* **14,** 1 (2016).

138. Martin, C. H. *et al.* Diabolical survival in Death Valley : recent pupfish colonization, gene flow and genetic assimilation in the smallest species range on earth. *Proceedings of the Royal Society B : Biological Sciences* **283,** 20152334 (2016).

139. Sağlam, İ. K. *et al.* Phylogenetics support an ancient common origin of two scientific icons : Devils Hole and Devils Hole pupfish. *Molecular ecology* **25,** 1–12 (2016).

140. Lescak, E. A. *et al.* Evolution of stickleback in 50 years on earthquake-uplifted islands. *Proceedings of the National Academy of Sciences of the United States of America* **112,** E7204–12 (2015).

141. Johnson, T. *et al.* Late Pleistocene Desiccation of Lake Victoria and Rapid Evolution of Cichlid Fishes. *Science (New York, N.Y.)* **273,** 1091–1093 (1996).

142. Seehausen, O. Patterns in fish radiation are compatible with Pleistocene desiccation of Lake Victoria and 14,600 year history for its cichlid species flock. *Proceedings of the Royal Society B : Biological Sciences* **269,** 491–497 (2002).

143. Harmon, R. S. *et al.* Late Pleistocene paleoclimates of North America as inferred from stable isotope studies of speleothems. *Quaternary Research* **9,** 54–70 (1978).

A MISSION DE TERRAIN

Étudier une espèce dans son milieu naturel permet de mieux la connaître. En effet, les individus nés en élevage et ceux vivant dans le milieu naturel sont soumis à des conditions environnementales très différentes.

Par exemple, chez *Astyanax mexicanus*, les poissons cavernicoles sont élevés dans un cycle jour/nuit de 12h/12h alors que dans les populations naturelles les poissons sont soumis à une obscurité permanente. À l'animalerie, l'apport en nourriture est régulier.

Dans le cadre de la collaboration avec l'équipe de Sylvie Rétaux à l'INAF, nous nous sommes rendus au Mexique en mars 2016. Cette expédition avait pour but de réaliser des prélèvements d'eau, d'échantillonner des poissons de surface et cavernicoles ainsi que de réaliser des expériences *in situ*. Le groupe était composé de différentes personnes ayant des rôles précis : Luis Espinasa du Marist College à New York (connaissance des grottes de la Sierra de El Abra, sécurité), Éric Queinnec et Karen Pottin (échantillonnage de la faune cavernicole), Stéphane Père (pêche, sécurité), Laurent Legendre (échantillonnage d'eau, pêche, sécurité, étude de la diversité aquatique dans les rivières), Lucie Devos (échantillonnage génétique des poissons, FIV *in situ*), Victor Simon (prélèvement d'écailles), Carole Hyacinthe (enregistrement audio), Maryline Blin (olfaction), Didier Casane (échantillonnage), Sylvie Rétaux (supervision, pêche, échantillonnage, olfaction, photo) et moi même (bioinformatique de terrain).

La première expérience réalisée consistait à reproduire les expériences permettant de tester les capacités olfactives des poissons cavernicoles qui sont réalisées au laboratoire et répliquer une qui a été réalisée *in situ* lors d'une précédente mission de terrain [76].

Nous voulions également enregistrer les sons produits par les poissons cavernicoles afin d'étudier les différences de communication orale entre poissons de surface et poissons cavernicoles.

A.1 Étude du comportement *in situ* et de la communication acoustique

A.1.1 *Grottes étudiées*

Les expériences d'olfaction et d'enregistrement audio et vidéo ont été réalisées dans trois grottes différentes : Pachón, Rio Subterraneo et Tinaja.

A.1.2 *Enregistrement des poissons*

A.1.2.1 *Éclairage*

La principale difficulté était d'enregistrer des poissons dans l'obscurité totale et durant une longue période de temps (24h). En effet, afin de filmer, il faut de la lumière qui pourrait perturber les poissons ainsi que les autres animaux vivant dans la grotte, comme les chauves-souris. Nous avons donc choisi d'éclairer la scène avec de la lumière non-visible. Il fallait alors choisir entre lumière noire (ultra-violets, UV) et lumière infra-rouge (IR). Bien que les infra-rouges pénètrent difficilement dans l'eau, nous avons choisi d'utiliser cette lumière car exposer des poissons dépigmentés à des rayonnements ultra-violets pourrait leur être préjudiciable.

Nous avons utilisé des spots de 36 LED infrarouges, émettant à une longueur d'onde de 850nm, de marque Polaroïd, alimentables avec une source d'énergie de 5V. De plus, ils sont adaptables sur les pieds à spot de lumière habituellement utilisés en photographie.

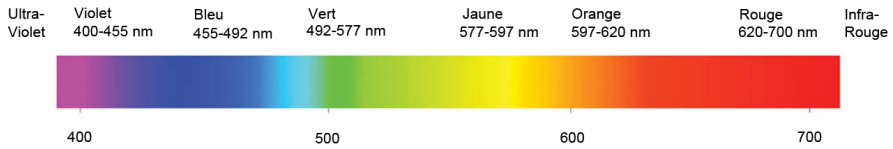


Figure 48. Spectre lumineux.

A.1.2.2 *Enregistrement vidéo*

La plupart des caméras disponibles dans le commerce sont insensibles aux infrarouges. En effet, leur capteur possède un filtre afin d'éliminer les infrarouges qui peuvent alterer la qualité des images, en particulier celle des couleurs (Figure 49).

Figure 49. Photo prise avec un appareil photo équipé d'un filtre IR (à gauche) et sans filtre IR (à droite). *Photos de Daniel Schwen*.

De plus, ces caméras sont équipées de batteries de faibles capacités, empêchant l'enregistrement sur de longues périodes de temps. Elles sont parfois également incapables de fonctionner branchées à une source d'alimentation.

Nous avons donc choisi de développer notre propre caméra infrarouge en utilisant un Raspberry Pi 2. Le Raspberry Pi est un nano-ordinateur possédant un processeur quadricoeur de type ARM cadencé à 900MHz et 1Go de RAM. Il est utilisé par de nombreux électroniciens en herbe pour réaliser des circuits électroniques contrôlés informatiquement, car il est facilement programmable en Python. Il est peu gourmand en électricité : il peut être alimenté par un simple câble USB en 5V. De nombreux accessoires sont disponibles pour cet ordinateur. Notamment, il est possible d'y ajouter un capteur photo sensible aux infrarouges (PiNoIR). L'ordinateur et la caméra ont été montés dans un boitier adaptable sur un pied d'appareil photo.

Le script permettant de contrôler la caméra a été écrit en Python, avec l'aide de la librairie , et est disponible sur Github : https://github.com/julienfumey/PiCaveRecord.git.

A.1.2.3 *Enregistrement sonore*

Les enregistrements sonores ont été réalisés avec des hydrophones Aquarian H2a. Les hydrophones étaient connectés à un enregistreur portable amplifié Zoom H4n.

A.1.2.4 *Ressources énergétiques*

Les lampes et la caméra nécessitent tous les deux une alimentation électrique. Les grottes n'étant évidemment pas équipées de prises électriques branchées sur un réseau électrique, nous avons dû apporter nos propres sources d'énergie. Nous avons choisi d'utiliser des banques d'énergie, couramment utilisées pour recharger un téléphone portable en USB en dépannage. Ces batteries développent un courant de 5V, qui est la tension nécessaire pour alimenter à la fois la caméra et les lampes infrarouges.

La capacité d'une batterie est exprimée en ampère-heure (Ah). Cette mesure indique le nombre d'ampères disponibles pendant une heure. Connaissant l'ampérage d'un appareil alimenté par une batterie, on peut connaître la temps de fonctionnement maximum.

Les batteries sélectionnées avaient une capacité de 16Ah. Branché sur une de ces batteries, le Raspberry Pi a fonctionné pendant 32h lors d'un test au laboratoire.

L'enregistreur sonore a été alimenté avec une batterie 9V de 1,2Ah permettant d'enregistrer pendant environ 20h.

Le dispositif expérimental (hydrophone, enregistreur sonore, caméra, lampe infrarouge) a été installé dans les trois grottes Figure 50 et laissé sur place, en fonctionnement, pendant une nuit complète.

Figure 50. Mise en place du dispositif expérimental d'enregistrement vidéo et audio longue durée dans la grotte Rio Subterraneo.

Les bandes sonores sont en cours d'analyse. Malheureusement, les films ne sont pas exploitables car après environ 30 min d'enregistrement l'image s'assombrit fortement. Des filtres corrigeant la luminosité et le contraste permettent de sauver quelques minutes d'enregistrement (Figure 51) mais sur une courte durée seulement. Sur une grande partie du film, aucun signal n'est détectable.

Cet assombrissement pourrait être dû au capteur IR ou plus généralement aux composants électroniques qui auraient pu ne pas supporter les conditions climatiques (chaud et humide) à l'intérieur des grottes. Les spots IR utilisés pourraient également avoir disfonctionné. En particulier, ils sont munis de petites batteries permettant de les utiliser en autonomie. Une fois chargée, les batteries auraient pu chauffer, entraînant de la condensation au niveau du circuit électronique. Enfin, les batteries utilisées pour alimenter l'ensemble du dispositif pourraient avoir délivré un potentiel électrique plus faible au fur et à mesure que sa charge décroissait.

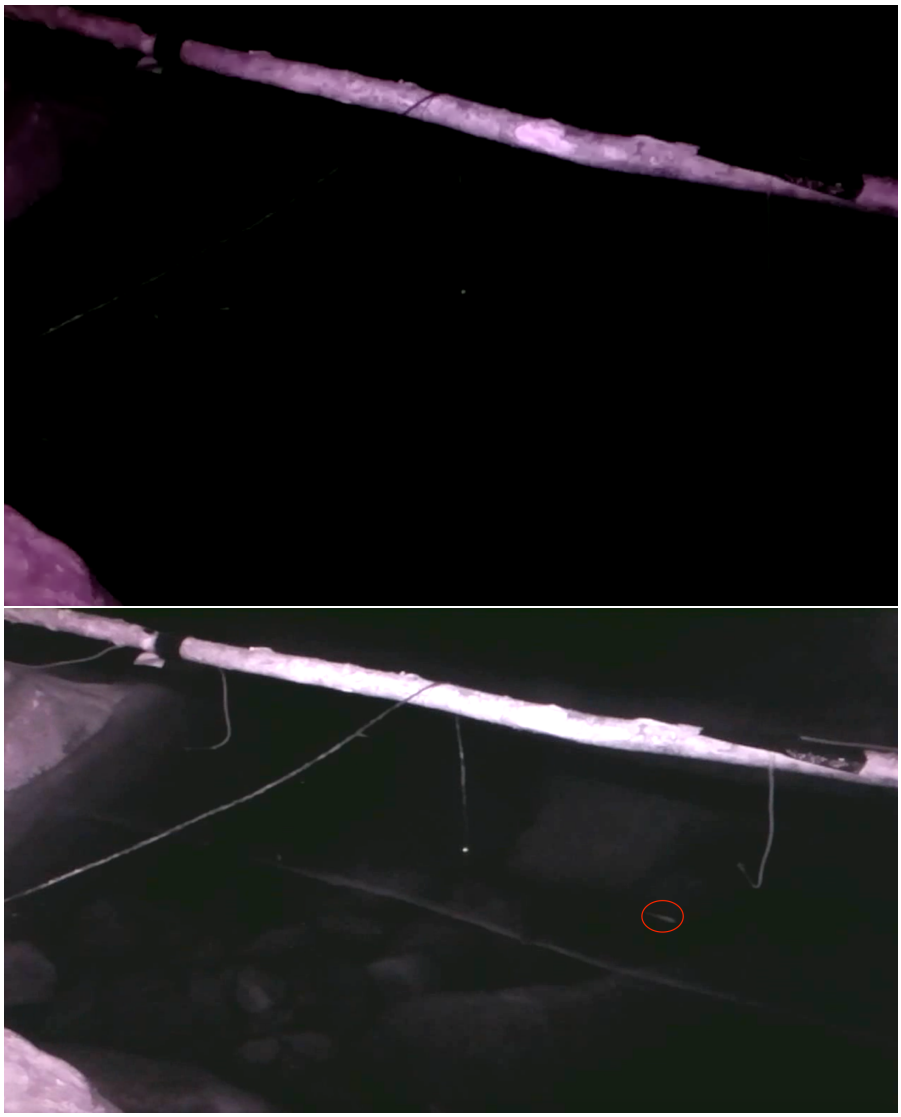


Figure 51. Capture d'écran du film réalisé à la grotte Tinaja avant (en haut) et après (en bas) retouche de la luminosité et du contraste. Après retouche, des poissons sont visibles dans le film (par exemple dans le cercle rouge).

A.2 Olfaction

Trois piscines gonflables ont été installées dans la grotte en même temps que le dispositif présenté auparavant. Ces piscines ont été remplies d'eau provenant des bassins de la grotte, et une dizaine de poissons y ont été placés pendant 24h pour habituation.

Le dispositif permettant l'enregistrement des poissons dans les bassins de la grotte a alors été réutilisé afin d'enregistrer les expériences d'olfaction (Figure 52). L'expérience est réalisée une fois par piscine gonflable dans le noir complet (Figure 53). Une expérience dure environ 30 minutes.

À l'aide de seringues et de tubulures de perfusion, différentes solutions sont injectées dans la piscine : eau, acide aminé à différentes concentrations.

Figure 52. Dispositif expérimental pour les expériences d'olfaction dans la grotte Pachón.

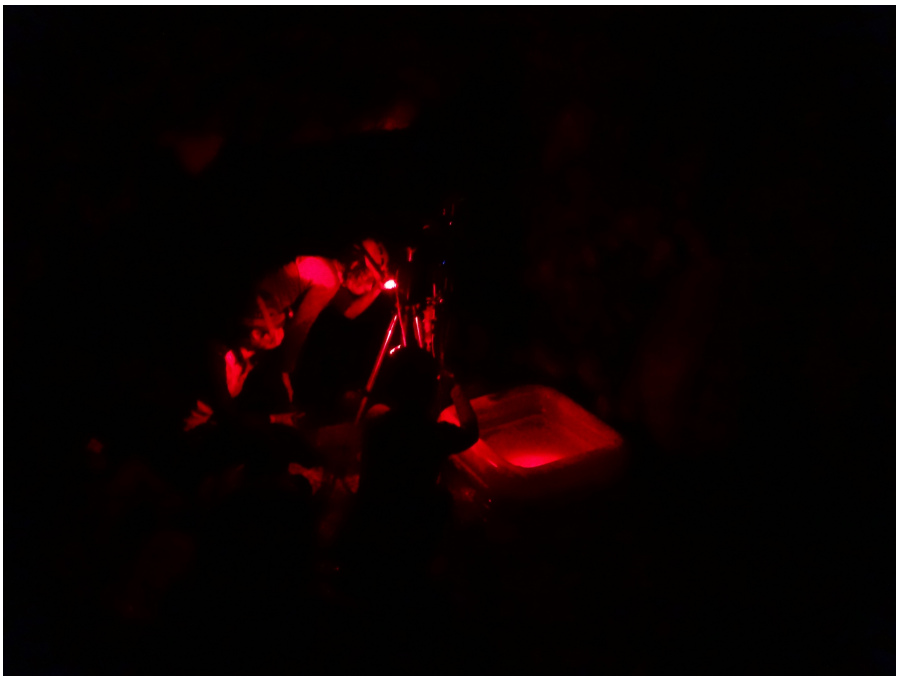


Figure 53. Lors d'une expérience d'olfaction dans la grotte Rio Subterraneo.

Les différentes concentrations d'acides aminés servent à tester le seuil de détections des poissons.

Les vidéos réalisées (Figure 54) sont en cours d'analyses, mais des résultats préliminaires semblent montrer une cohérence entre les observations faites au laboratoire et celles réalisées dans les différentes grottes.

Figure 54. Capture d'écran d'un film d'une expérience d'olfaction dans la grotte Rio Subterraneo

B ARTICLES PUBLIÉS

B.1 Lens Defects in *Astyanax mexicanus* Cavefish : Evolution of Crystallins and a Role for αA-Crystallin

Lens Defects in *Astyanax mexicanus* Cavefish: Evolution of Crystallins and a Role for alphaA-Crystallin

Hélène Hinaux,[1] Maryline Blin,[1] Julien Fumey,[2] Laurent Legendre,[3] Aurélie Heuzé,[4] Didier Casane,[2] Sylvie Rétaux[1]

[1] DECA group, Neurobiology and Development Laboratory, UPR3294, CNRS avenue de la terrasse, 91198, Gif sur Yvette, France

[2] MULTIGEN group, Evolution Genomes and Speciation Laboratory, UPR9034, CNRS avenue de la terrasse, 91198, Gif sur Yvette, France

[3] AMAGEN (UMS 3504 CNRS / UMS 1374 INRA), CNRS avenue de la terrasse, 91198, Gif sur Yvette, France

[4] CASBAH group, Neurobiology and Development Laboratory, UPR3294, CNRS avenue de la terrasse, 91198, Gif sur Yvette, France

Received 1 April 2014; accepted 24 October 2014

ABSTRACT: **The fish *Astyanax mexicanus* presents, within the same species, populations of river-dwelling surface fish (SF) and blind cave-living fish. In cavefish (CF), the eyes develop almost normally during embryogenesis. But 40 h after fertilization, the lens enters apoptosis, triggering the progressive degeneration of the entire eye. Before apoptosis, the CF lens expresses early differentiation factors correctly. Here, we searched for possible late differentiation defects that would be causal in CF lens degeneration. We reasoned that crystallins, the major lens structural proteins, could be defective or misregulated. We surveyed the CF and SF transcriptomes and uncovered 14 *Astyanax* crystallins from the beta, gamma, lambda, mu, and zeta families. These proteins are less polymorphic and accumulate more fixed mutations, some at highly conserved positions, in CF than in SF, suggesting relaxed selection at these loci in CF. *In situ* hybridizations and qPCR show that *crybb1c*, *crybgx*, *crygm5* are expressed at much lower levels or are not expressed in the CF lens. For the best crystallin candidates, we tested a potential causal role in CF lens apoptosis. *Crybgx*, *crybb1c* (not expressed in CF from very early on), and *cryaa* (previously shown to be faintly expressed in CF) failed to induce any defect when knocked-down in zebrafish embryos. However, the anti-apoptotic *cryaa* protected lens cells from apoptosis when reexpressed by transgenesis in CF, suggesting a cell-autonomous effect of cryaa on lens cell survival. Altogether, these data suggest that crystallin sequence evolution and expression defects may contribute to the loss of eyes in CF.** Develop Neurobiol 75: 505–521, 2015

Keywords: **molecular evolution; phylogeny; gene expression; cryaa; apoptosis**

INTRODUCTION

Astyanax mexicanus is a species of teleost fishes living in Mexico and comprising two types of populations or morphs: populations of eyed river-dwelling fishes

Additional Supporting Information may be found in the online version of this article.

Correspondence to: S. Rétaux (retaux@inaf.cnrs-gif.fr)

Contract grant sponsors: Retina France (to H.H.), Agence Nationale pour la Recherche ANR grants [ASTYCO] (to S.R.) and [BLINDTEST] (to S.R. and D.C.), and IDEEV (to S.R. and D.C.).

Published online 4 November 2014 in Wiley Online Library (wileyonlinelibrary.com).
DOI 10.1002/dneu.22239

(surface fish, SF) and 29 populations of blind cavefishes (CF) leaving in total and permanent darkness (Jeffery, 2008). This makes this fish an outstanding model species for microevolutionary studies. In particular, the mechanisms underlying the developmental regression of the eyes in CF have puzzled researchers for several decades (Jeffery, 2009; Rétaux and Casane, 2013).

The most studied CF population lives in the Pachón cave, and is the population studied in this article. In this fish, eyes first develop during embryogenesis, but they soon degenerate progressively during larval stages, so that only a cyst covered by skin remains in the fully eyeless and blind adult. The first eye structure that degenerates is the lens. It does so by undergoing massive apoptosis, starting at 40 h post fertilization (hpf, i.e., after hatching), and cell death subsequently spreads to the entire neural retina (Alunni et al., 2007). Importantly, the lens is responsible for the degeneration of the entire eye, as the transplantation of a CF apoptotic lens into a SF optic cup is sufficient to induce degeneration of the SF eye (Yamamoto and Jeffery, 2000). Conversely, the transplantation of a SF lens into a CF eye is sufficient to partially restore the CF eye. As these transplantation experiments were performed at 24 hpf, the authors concluded that CF lens apoptosis is a process autonomous to the lens (Yamamoto and Jeffery, 2000). Of note at 24 hpf in both SF and CF *Astyanax*, invagination of the lens mass is already completed and the first differentiation-driving transcription factors such as *Prox1* are expressed (Jeffery et al., 2000).

Vertebrate lenses are made of two cell types: fiber cells and epithelial cells. A lens epithelium forms correctly in CF, and continues to proliferate even after the onset of apoptosis (Alunni et al., 2007; Strickler et al., 2007a). Thus, defects in the CF lens could involve fiber cell differentiation deficiencies. The differentiation of these cells mostly involves the accumulation of particular proteins, denominated crystallins, at high concentrations in the central fiber cells of the lens (Fagerholm et al., 1981). These proteins maintain the refractive properties and optical clarity of the lens through their structural and chaperone-like roles. They belong to three major (and large) families: α, β, and γ. The classification of a protein as "crystallin" is arbitrary and does not have a phylogenetic basis: α-crystallins are related to small heat shock proteins (de Jong et al., 1993), whereas β and γ crystallins belong to the same superfamily as *aim* genes (absent in melanoma) (Wistow et al., 2005). In mammals, there are 14 crystallins (two alphas, seven betas, and eight gammas). In zebrafish, there are three alpha, 13 beta, and at least 37 gamma crystallins (Greiling et al., 2009). Besides, there are specific expansions of some crystallin subfamilies in fish, which therefore have no ortholog in mammals (Greiling and Clark, 2012). Functionally, α-crystallins act as chaperones, and their expression increases with age in several species including zebrafish (Greiling et al., 2009). The αA-crystallin *cryaa* in particular has an anti-apoptotic role in cell culture (Andley, 2000). In human, its loss or some mutations can lead to cataract (Litt, 1998; Mackay et al., 2003), and in zebrafish, the *cloche* mutant which has reduced levels of *cryaa* develops cataract very early (Goishi, 2006). However, *cryaa* knockdown by morpholino injection does not have a detectable effect in zebrafish (Posner et al., 2013). Crystallins from other families are less well known, in particular in fish, except for some expression data in zebrafish (Chen et al., 2001; Goishi, 2006; Wang et al., 2008). In short, the function of crystallins in fish lens development is poorly studied and understood.

More than 15 years ago, the expression of the anti-apoptotic *cryaa* was found to be lost in *Astyanax* Piedras CF (Behrens, 1998), opening novel directions in the understanding of CF eye degeneration. Since then, it was also shown that *cryaa* expression is only very transient in Pachón CF lens (Strickler et al., 2007b). *Cryaa* is expressed from 36 hpf at low levels in *Astyanax* lens, and there is a clear difference of expression between SF and CF from 48 hpf onward. We reasoned that other crystallins could be defective or misexpressed in CF, thus affecting lens differentiation and causing apoptosis. We decided to survey the *Astyanax* developmental transcriptome we recently generated (Hinaux et al., 2013) for crystallin sequences. On the 14 *Astyanax* crystallins we identified, we performed a molecular evolution analysis and an expression study during the "critical" developmental period. For three of them that we considered as best candidates [*crybgx*, not expressed in CF from very early on; *crybb1c*, almost not expressed in CF; and *cryaa*, with anti-apoptotic role (Andley, 2000)], we further tested a potential causal role in CF lens apoptosis either by morpholino knockdown in zebrafish or by expression rescue using transgenesis in CF. We discuss the evolutionary forces at work on the CF crystallin system and the crystallin-dependent mechanisms involved in CF eye degeneration.

MATERIAL AND METHODS

Fish Samples

Laboratory stocks of *A. mexicanus* SF and Pachón CF were obtained in 2004 from the Jeffery laboratory at the University

of Maryland, College Park, MD. They had been both lab-raised for some generations, and SF had initially been collected in San Solomon Spring, Balmorhea State Park, TX. In our facility, they were maintained and bred at 23°C (Pachón) and 26°C (surface) on a 12:12 h light/dark cycle in tap water. Embryos were collected after natural spawning, staged according to the developmental staging table (Hinaux et al., 2011) and fixed at various stages in 4% paraformaldehyde (PFA). After progressive dehydration in methanol, they were stored at −20°C.

Danio rerio AB zebrafish strain originating from the ZIRC (University of Oregon) were bred in UMS AMAGEN facilities. They were maintained and grown at 28°C with 14 h light per day. Embryos were obtained by pair mating for maximum efficiency for injection's goal.

Animals were treated according to the French and European regulations for handling of animals in research. SR's authorization for use of animals in research is number 91–116, and Paris Centre-Sud Ethic Committee authorization number is 2012-0052.

Phylogeny

All zebrafish crystallin protein sequences were retrieved from Ensembl database and BLASTed against the *Astyanax* developmental transcriptome contig sequences [available at http://genotoul-contigbrowser.toulouse.inra.fr:9099/index.html, (Hinaux et al., 2013)]. The putative *Astyanax* crystallin protein contigs were then aligned with Vertebrate crystallin proteins, retrieved from Ensembl by Perl API "one2one homolog" tool, using zebrafish gene IDs as queries. Evolutionary analyses were conducted with MEGA5 (Tamura et al., 2011), using the Neighbor-Joining method (Saitou and Nei, 1987). The evolutionary distances were computed using the Poisson correction method (Zuckerlandl and Pauling, 1965) and are in the units of the number of amino acid substitutions per site. All ambiguous positions were removed for each sequence pair.

Sequence Comparisons

The developmental transcriptome of *A. mexicanus* SF and Pachón CF from our animal facility, as well as the transcriptome of the closely related Characiform *Hyphessobrycon anisitsi* (Javonillo et al., 2010), were sequenced on Illumina HiSeq at the Imagif high throughput sequencing platform. Transcriptome reads were assembled at the Genotoul bioinformatics platform in Toulouse. The complete dataset will be published elsewhere (Fumey et al., in preparation).

The 14 *Astyanax* crystallin sequences uncovered by the phylogenetic analysis were BLASTed against the Illumina assembly. Eleven crystallins were retrieved and a SNP analysis was performed on these sequences (*crybgx*, *crygm5*, and *crygmxb* were missing: mapping of the reads against the Sanger contig revealed that there were almost no CF read for these crystallins, preventing any SNP analysis). First, variations in *Astyanax* sequences were detected using GATK (McKenna et al., 2010; DePristo et al., 2011). Then, a homemade program allowed to filter these variations: SNPs were defined as positions with at least two reads per allele and at least four reads per population [similar thresholds as in (Hinaux et al., 2013)]. SNPs were then classified as: (1) shared when the position was polymorphic in both SF and CF; (2) polymorphic in only one morph, the position being fixed in the other morph; or (3) fixed in both morphs.

The direction of the various changes was inferred using *H. anisitsi* as an outgroup: alleles were defined as ancestral when they were identical to the outgroup allele, and as derived otherwise.

Nonsynonymous mutations in CF were further analyzed by aligning Vertebrate crystallin protein sequences (retrieved from Ensembl) in CLC Sequence Viewer 7.0.2 (Qiagen) and looking for conservation. The position of secondary structures was found in the literature for crybb1 crystallins (Van Montfort et al., 2003) and in Molecular Modeling Database MMDB for cryba crystallins (Chaikuad et al., 2010; Madej et al., 2012).

Whole-Mount In Situ Hybridization

cDNAs were amplified by PCR from pCMV-Sport6 plasmids picked from our cDNA library using SP6 and T7 primers, and digoxygenin- riboprobes were synthesized from PCR templates. A protocol for automated whole-mount *in situ* hybridization (Intavis) was performed. Briefly, embryos were progressively rehydrated, permeabilized by proteinase K (Sigma) treatment before being incubated over night at 68° in hybridization buffer containing the appropriate crystallin probe. After stringent washes, the hybridized probes were detected by immunohistochemistry using an alkaline phosphatase-conjugated antibody against digoxygenin (Roche) and a NBT/BCIP chromogenic substrate (Roche).

After staining, embryos were photographed *in toto*, always in the same orientation, under a Nikon AZ100 stereomicroscope using agarose wells.

Quantitative PCR

Total RNA was extracted from 36 hpf CF or SF embryos with TRIzol reagent (Invitrogen) followed by purification and DNase treatment with the Macherey Nagel NucleoSpin® RNAII kit. RNA amounts were determined by the Nanodrop 2000c spectrophotometer (Thermo Scientific). Total RNA of 1 μg was reverse transcribed in a 20 μL final reaction volume using the High Capacity cDNA Reverse Transcription Kit (Life Technologies) with RNase inhibitor and random primers following the manufacturer's instructions. Quantitative PCR was performed on a QuantStudio™ 12K Flex Real-Time PCR System with a SYBR green detection protocol. cDNA of 3 ng were mixed with Fast SYBR® Green Master Mix and 500 n*M* of each primer in a final volume of 10 μL. The reaction mixture was submitted to 40 cycles of PCR (95°C/20 sec; [95°C/1 sec; 60°C/20 sec] X40) followed by a fusion cycle to analyze

the melting curve of the PCR products. Negative controls without the reverse transcriptase were introduced to verify the absence of genomic DNA contaminants. Primers were designed using the Primer-Blast tool from NCBI and the Primer Express 3.0 software (Life Technologies). Primers were defined either in one exon and one exon–exon junction or in two exons span by a large intron. The primers used are as followed:

0964-AMcrybgx-F1 CTGTGCTCCAACGTGCCTTT
0965-AMcrybgx-R1 GCACGGAGTTGCATCTTTCAG
0966-AMcrybb1c-F1 CGATTGCTTCATGTCCGTCA
0967-AMcrybb1c-R1 CACAGGCTGGGGATGTCTTC
0968-AMcrybb1a-F1 GGAGACCCCTTCATGGGAAA
0969-AMcrybb1a-R1 ACCACGGTCACACACGTTCA
0970-AMgapdh-F1 GTTGGCATCAACGGATTTGG
0971-AMgapdh-R1 CCAGGTCAATGAAGGGGTCA

Specificity and the absence of multilocus matching at the primer site were verified by BLAST analysis. The amplification efficiencies of primers were generated using the slopes of standard curves obtained by a fourfold dilution series. Amplification specificity for each real-time PCR reaction was confirmed by analysis of the dissociation curves. Determined Ct values were then exploited for further analysis, with the Gapdh gene as reference. Each sample measurement was made at least in duplicate. Statistical analysis of expression differences was performed using Mann–Whitney test.

Morpholino Injection

Three morpholinos targeting *D. rerio crybgx*, *crybb1c*, and *cryaa* mRNAs were designed and ordered from GeneTools: one targets the splice site at the junction of Exon 4 and Intron 4 of *crybgx* (5′ ACTATAACTGTGTGTCTGACCTGTT 3′), another targets the splice site at the junction of Exon 4 and Intron 4 of *crybb1c* (5′ ATAGAGCTAT TAACCCACCATTCGA 3′), and the last one blocks translation of *cryaa* by targeting ATG (5′GTGTTGGATCGCA ATATCCATAATG 3′) and was already published (Posner et al., 2013).

These morpholinos were injected at the one-cell stage in zebrafish AB strain eggs at 250–500 μ*M*. Zebrafish larvae were then raised at 28°C for 2, 3, or 4 days, and fixed in PFA 4% for TUNEL assay (Promega), or immersed in Trizol for RNA extraction. The form of *crybgx* mRNA was assessed with PCR on cDNA, using the following primers amplifying Exons 1 to 5: Fw 5′ AGTCCCCGGACTAGCCCAAC 3′ and Rv 5′ GGTTGGGAATGTCATCATGAAGC 3′; and the form of *crybb1c* mRNA was assessed with PCR on cDNA, using the following primers amplifying Exons 3 to 6 : Fw 5′ GGGCTCCATCAGAGTGGAG 3′ and Rv 5′ CTGGGTACTGATATCCAACCC 3′.

Transgenesis

A fish transgenesis vector pDEST_AMA12Hkaede from the AMAGEN transgenesis platform, with a CFP reporter gene under the control of zebrafish *crybb1a* promoter, was modified to allow the expression of *Astyanax* SF *cryaa* under the control of zebrafish *crybb1a* promoter. The vector was coinjected with Isce1 meganuclease mRNA or Tol2 transposase at the one-cell stage in *Astyanax* CF embryos (Elipot et al., 2014). F0 embryos were screened at 60 hpf for lens fluorescence. Positive larvae were raised until adulthood or fixed at 60 hpf. Fixed larvae were either subjected to *in situ* hybridization for *cryaa* or immuno-stained with chick anti-GFP antibody (1:1000, Aveslab, which efficiently recognizes CFP) and rabbit anti-activated caspase3 antibody (1:1000, BD Pharmingen). Secondary antibodies raised in goat and coupled to AlexaFluor dyes (Invitrogen) were used (1:500). Larvae were counterstained with DAPI (5 μg/mL, Sigma). All images were acquired on a Leica SP8 confocal microscope using 40X oil objective (ON 1.30), with the LAS AF software (Leica), in a sequential manner. Cell counts were carried out using Fiji Cell Counter plugin. Statistical analysis of apoptosis differences in transgenic and nontransgenic cells was performed using Fischer's exact test.

RESULTS

Identification of Astyanax Crystallins and Phylogeny

We took advantage of *Astyanax* developmental transcriptome data we obtained recently (Hinaux et al., 2013) and which corresponds to transcripts expressed between 6 hpf and 2 weeks of development, to search for crystallin ESTs in our model species. We BLASTed protein sequences of all zebrafish crystallins to retrieve all *Astyanax* crystallin contigs present in our transcriptome assembly. We then aligned them for phylogenetic analysis with several teleostean as well as some tetrapod vertebrate crystallin proteins. Below we report the identification of 14 *Astyanax* crystallins: six β-crystallins, three γ-crystallins as well as a large number of γM2d crystallins, two βγ crystallins, one λ-crystallin, one μ-crystallin and one ζ-crystallin (Genbank accession numbers KM873651to KM873664, Table 1).

The crystallin phylogenetic trees shown were built using the Neighbor-Joining method. Identical results concerning the orthology of the various *Astyanax* sequences were obtained when using Maximum Likelihood method (data not shown).

In the beta family, our phylogenetic tree confirms the existence of two major subgroups: βA and βB [Fig. 1(A)]. *Astyanax* transcripts of the β family clearly correspond to six different genes: *crybb1a*, *crybb1c*, *crybb1d*, *cryba2*, *cryba1l*, and *cryba1b*. This crystallin group is, therefore, well represented during *Astyanax* embryonic and larval lens development.

Table 1 A Summary Table for Crystallins Studied in This Article

Phylogenetic Annotation (Genbank Accession Number)	Transcript Contig name	Ensembl Gene Code	SF/CF Nonsynonymous Fixed Differences	Number of SF Sanger ESTs		Number of CF Sanger ESTs	
				Per contig	Total	Per contig	Total
cryba1b (KM873657)	ARA0AAA58YH21EM1.b.am.1	ENSAMXG00000019822	Mutation in CF 96R > H	25	25	1	1
cryba1l (KM873664)	ARA0AAA17YM13EM1.b.am.1	ENSAMXG00000013063		2	64	1	5
	ARA0AHA22YO11EM1.b.am.1			1		0	
	ARA0AAA20YO03EM1.b.am.1			61		4	
cryba2 (KM873663)	ARA0AHA19YM01EM1.b.am.1	ENSAMXG00000002203		1	26	0	1
	ARA0AAA86YO20EM4.b.am.1			25		1	
crybb1a (KM873660)	ARA0ABA13YJ01EM1.b.am.1	ENSAMXG00000009325?	Mutations in CF 113G > V and 180V > M	1	1	0	0
crybb1d (KM873653)	ARA0ABA19YN10EM1.b.am.1	ENSAMXG00000013053		9	9	0	0
crybb1c (KM873655)	ARA0ABA90YK17EM1.b.am.1	ENSAMXG00000017742		1	1	0	0
crygn2 (KM873662)	ARA0ABA26YL06EM1.b.am.1	ENSAMXG00000018024		1	33	0	1
	ARA0AAA74YI01EM1.b.am.1			32		1	
crybgx (KM873659)	ARA0ABA105YF08EM1.b.am.1	ENSAMXG00000006190	NA	11	11	0	0
Crygmxb (KM873652)	ARA0ABA58YE03EM1.b.am.1	ENSAMXG00000010864 (exons 5–6)	NA	2	2	0	0
crygm5 (KM873651)	ARA0ABA56YK03EM1.b.am.1	ENSAMXG00000010864 (exons 6–7-8)	NA	1	1	0	0
cryzl1 (KM873656)	ARA0ADA47YD17EM1.b.am.1	ENSAMXG00000012452		0	0	1	1
crym (KM873658)	ARA0AAA98YN19EM1.b.am.1	ENSAMXG00000008929		1	1	1	1
cryl1 (KM873661)	ARA0ABA18YD12EM1.b.am.1	ENSAMXG00000005498		3	4	0	1
	ARA0AAA39YP23EM1.b.am.1			1		1	
crybg3 (KM873654)	ARA0ABA21YF17EM1.b.am.1	ENSAMXG00000008917	Mutation in CF 601R > C	1	1	1	1

Successive columns show (1) the "phylogenetic name" given after phylogenetic analysis and corresponding Genbank Accession number, (2) the transcript contig name in our *Astyanax* transcriptome browser (Hinaux et al., 2013), (3) the corresponding Ensembl gene code in the preliminary Pachón genome released on 6 December 2013 (http://uswest.ensembl.org/Astyanax_mexicanus/Info/Index; McGaugh et al., 2014), (4) the position and nature of identified fixed mutations in CF protein sequences, and (5,6) the number of EST for each crystallin retrieved from our transcriptome sequencing (Sanger) in SF and CF. Note: according to the Pachón cavefish preliminary assembly, there would be a premature stop codon in the 5th exon of *crybgx*. However, the sequence is "unknown" only 28bp after this stop codon, and in a more recent transcriptome Illumina dataset (see Methods) obtained in the group, we do not find this premature stop codon.

The γ family is also subdivided into different subgroups [Fig. 1(B)]. On one hand, crystallins of the γA to γF subfamilies do not have orthologs in fish including *Astyanax*. On the other hand, crystallins of the γM subfamily (which appear as paraphyletic on the tree) are "fish-specific" (Wistow et al., 2005). In this group, we identified two *Astyanax* sequences: *crygm5* and *crygmxb*. Finally, γN and γS crystallins are present in all vertebrates, and one *Astyanax* member, *crygn2*, was retrieved from the developmental transcriptome.

Besides, we found a particularly high number ($n = 47$) of crystallin contigs belonging to the γM2d subfamily (Supporting Information FigS1A). The sequences of these contigs were BLASTed against the preliminary Pachón CF genome assembly, and several hits were retrieved on six different scaffolds (KB882270.1, KB882254.1, KB882235.1, KB882151.1, KB882173.1, and KB882147.1; McGaugh et al., 2014). On one of them (KB882151.1), there are many hits between 2.73 and 2.84 Mb (mostly unannotated in the current Ensembl version – release 74) that could correspond to tandem duplications. Only four crystallin genes are annotated in this region, ENSAMXG00000009606, ENSAMXG00000009646, ENSAMXG00000009725 and ENSAMXG00000009713, and would correspond to some of these paralogs (McGaugh et al., 2014).

Two crystallins of the βγ family, usually grouped with the gamma family, are present in the *Astyanax* developmental transcriptome: *crybgx* and *crybg3* [Fig. 1(B) and Supporting Information FigS2]. *Crybg3* is not shown on the γ family tree [Fig. 1(B)], as it does not align correctly with other γ crystallins. The phylogenetic analysis for this subfamily was performed by alignment with *aim* (absent in melanoma) genes, to which they seem to be more related (Supporting Information FigS2).

No tree is shown for the α-crystallin family, as there is no α-crystallin transcript in our dataset. Finally, some other crystallins are present in our transcriptome data: a λ-crystallin *cryl1*, a μ-crystallin *crym*, and a ζ-crystallin *cryzl1* (Supporting Information FigS2). In Supporting Information Table 1 (Columns 1–3), the correspondence between our developmental transcriptome contig name, the "phylogenetic" name, the Genbank accession number and the Ensembl code of the recent genome assembly is given (McGaugh et al., 2014). In sum, we have identified 14 crystallins of various groups, plus 47 γM2d crystallins, that are expressed during development in *Astyanax*, at the time when the eye first develops and then starts degenerating.

Sequence Comparison between *Astyanax* SF and CF Crystallins

Our developmental transcriptome was generated from hundreds of individuals born in our fish facility, therefore giving us access to the genetic diversity present in SF and CF populations (Hinaux et al., 2013). We compared SF and CF coding sequences for 11 of the 14 crystallins uncovered in *Astyanax*, looking for polymorphic sites and for fixed mutations [Fig. 2(A,B)]. Of note, *crygm2d* crystallins were not analyzed as we could not assign the various transcripts to genes; and data on CF sequences for *crybgx*, *crygm5*, and *crygmxb* were missing. Among the 11 sequences analyzed, we found 16 variable positions which did not affect the protein sequence (= synonymous changes), and 13 variable positions which affected the protein sequence (= nonsynonymous changes) [Fig. 2(C,D), Table 1]. To infer the direction of the changes, that is, whether mutations occurred in the SF or in the CF lineage, we compared the alleles found in the two *Astyanax* morphs with those found at the same position in an outgroup, *H. anisitsi*, a closely related Characiform for which we also sequenced the developmental transcriptome (see Methods): we reasoned that the most parsimonious scenario was that mutations occurred in the lineage carrying an allele differing from the *H. anisitsi* allele [Fig. 2(A)]. Among the nonsynonymous changes,

- One is a variable position in both SF and CF (= a shared polymorphism) in *crybb1a*.
- Eight positions in *cryba1l*, *crybg3*, *crym*, and *cryzl1* are polymorphic in SF only. Among them, a polymorphic allele found for SF *cryba1l* results in an amino acid with different physicochemical properties from the ancestral one and can thus be considered as a radical mutation. For these eight positions, the CF allele is the same as the allele in the outgroup *H. anisitsi*, and therefore probably corresponds to the ancestral allele.
- No position is polymorphic in CF only, which is reminiscent of what was found transcriptome wide at larger scale in our previous study (Hinaux et al., 2013): CF are globally less polymorphic than SF, probably because of their small population size. This "rule" applies without exception to the crystallins.
- Finally four positions (in *cryba1b*, *crybb1a*, and *crybg3*) correspond to fixed mutations in the CF lineage. Two of these mutations (in *crybb1a* and *crybg3*) can be considered radical, that is, the amino acid properties in SF/*Hyphessobrycon* and CF are different, and could affect the

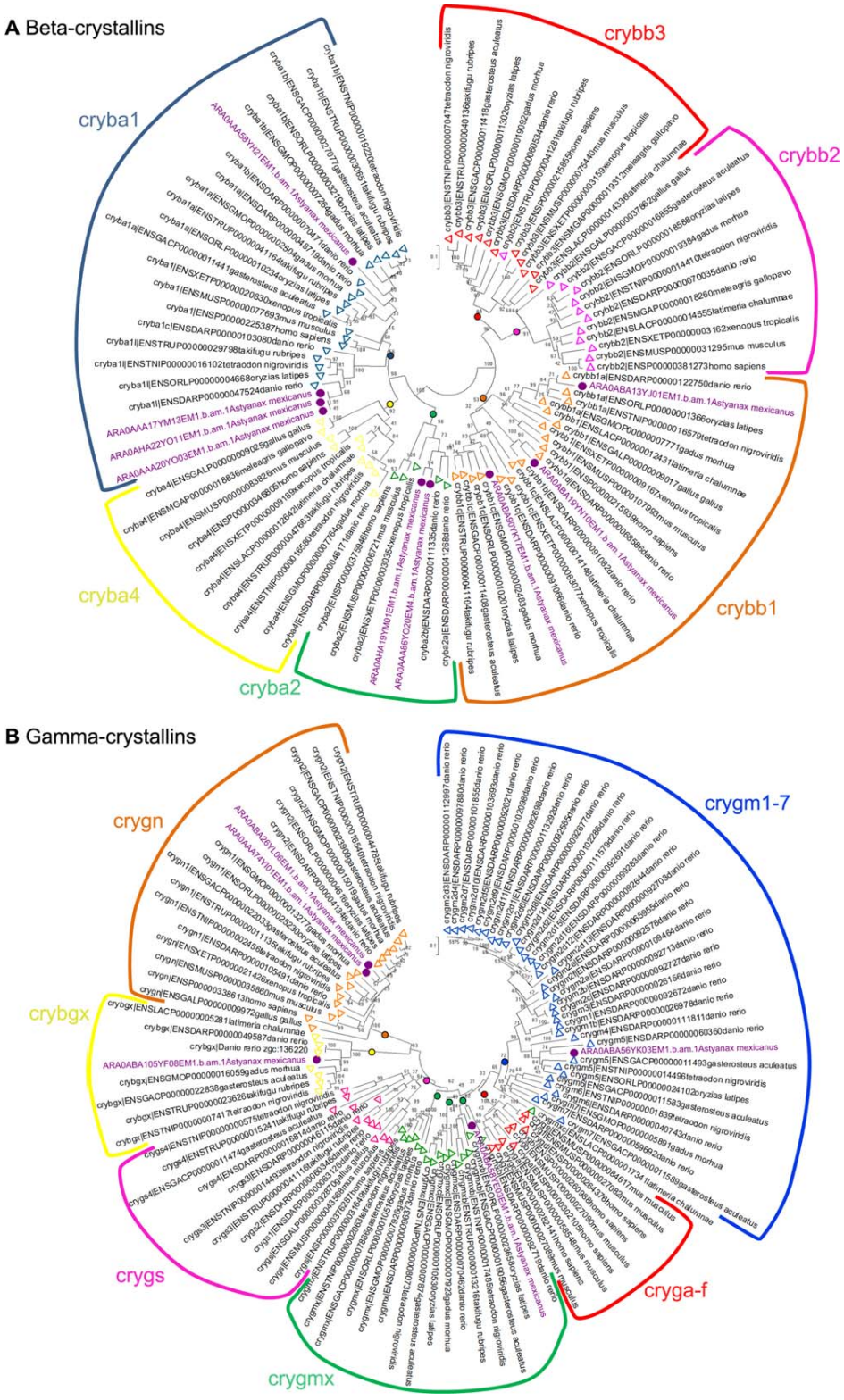


Figure 1 Phylogeny of Vertebrate crystallins of the beta family (A) and gamma family (B). Neighbor joining trees of vertebrate crystallin protein sequences. *Astyanax* sequences are landmarked with purple dots. Triangles indicate the orthology of the various sequences according to their names. *Takifugu rubripes crybb2* seems to belong to the *crybb3* subfamily. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

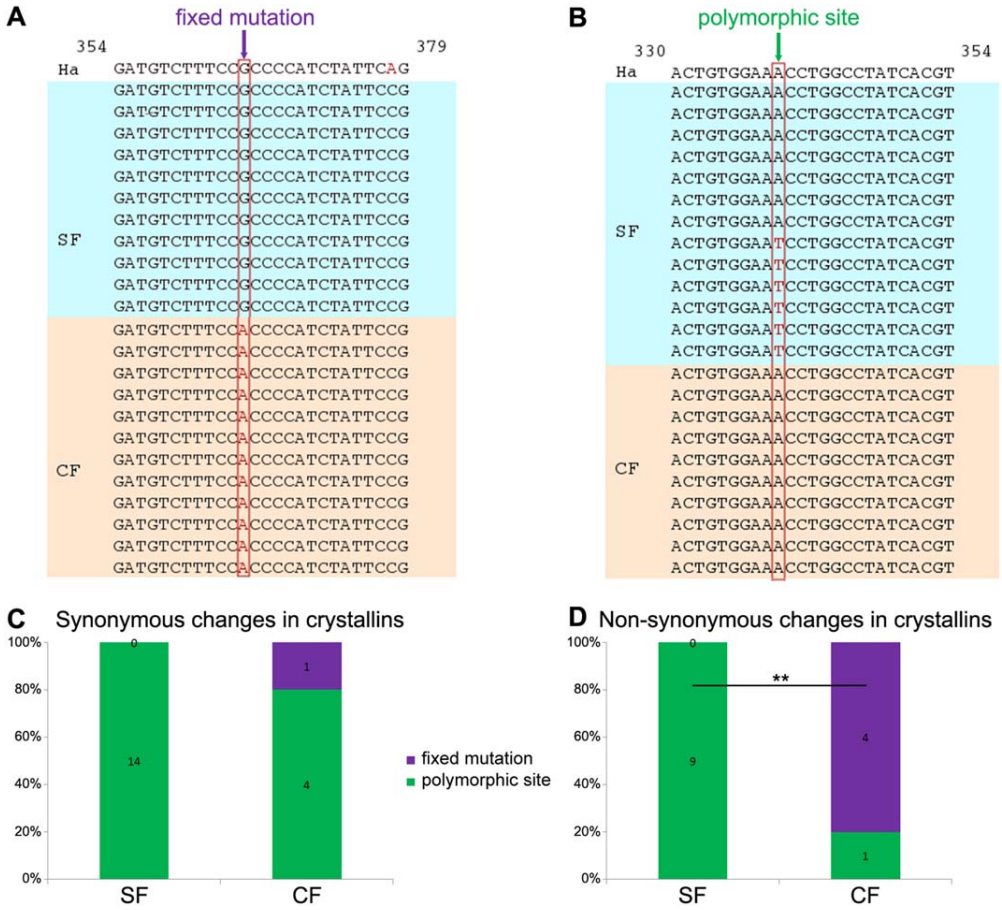


Figure 2 Sequence variability in *Astyanax* crystallins. (A) Alignment of nucleotidic sequences of a portion of *cryba1b* in *H. anisitsi* (Ha, white), SF (blue), and CF (orange) according to Illumina transcriptome sequencing, showing an example of a fixed mutation in CF. Nucleotides that differ from the consensus are shown in red. (B) Alignment of nucleotidic sequences of a portion of *cryba1l* in *H. anisitsi* (Ha, white), SF (blue), and CF (orange) according to Illumina transcriptome sequencing, showing an example of a polymorphic site in SF. Nucleotides that differ from the consensus are shown in red. (C) Within synonymous changes, proportion of fixed mutations (purple) and polymorphic sites (green) in each population. Numbers on the graphs indicate the number of sites in each category. (D) Within nonsynonymous changes, proportion of fixed mutations and polymorphic sites in each population. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

folding of the proteins. We did not find insertions/deletions in CF sequences.

- We did not detect any fixed mutation in SF crystallins.

In sum, SF have thus zero fixed nonsynonymous mutation and nine polymorphic positions, whereas CF have four fixed nonsynonymous mutations and only one polymorphic position [Fig. 2(D), Fisher test, $p = 0.0049$, **]. This pattern suggests that CF crystallins have a lower polymorphism level than SF (consistent with the small population size of CF), but that they accumulate more nonsynonymous fixed mutations than SF (Table 1, Column 4): this could be due either to relaxed purifying selection in CF on these genes (probably useless in the dark) or to a general decrease in selection intensity in the small CF population (genetic drift). However, the surprisingly low number of fixed changes detected in CF crystallins (only four nonsynonymous and one synonymous changes) makes it impossible to calculate a Ka/Ks ratio.

Finally, some of the mutations found in CF occur at highly conserved positions (Fig. 3): the R > H mutation in *cryba1b* is at a position conserved across Vertebrates for all beta A crystallins, and the G > V radical mutation in *crybb1a* is found at a position conserved across Vertebrates for all beta B

crystallins. The very high conservation at these positions, both found in beta sheets (Van Montfort et al., 2003; Chaikuad et al., 2010; Madej et al., 2012) is suggestive of strong selection on these amino acids. Changes at such positions in CF also point toward relaxed selection on crystallin genes.

In situ Hybridization Screen for Crystallin Expression in SF and CF Lens

We next examined expression patterns and levels of the crystallins identified above during early larval SF and CF development. Our transcriptome study, based on Sanger sequencing, was not intended to analyze expression differences between SF and CF. However, for some crystallins, the difference in the number of SF and CF reads was so striking that it gave a clue about expression differences (Table 1, Columns 5 and 6). Moreover, *cryaa*, previously reported to be expressed very transiently during a few hours around 30 hpf in CF (Strickler et al., 2007b), was cloned to be used as a control for *in situ* hybridization. The expression screen on the 14 identified crystallins was performed at 36 hpf, a stage slightly before the onset of lens apoptosis in CF (Alunni et al., 2007).

At this relevant and critical stage of lens development, we found that:

1. *cryaa* was lowly expressed in SF but not CF lens at 36 hpf [Fig. 4(A,B)]: *cryaa* is not correctly expressed in CF at a critical stage of CF lens development.
2. *cryba1b*, *cryba1l*, *crybb1a*, *crybb1d*, *cryba2*, and *crygn2* were strongly expressed in both SF and CF lenses at 36 hpf [Fig. 4(C–N)]. Importantly, the smaller expression spot in CF eyes is only due to the small size of the lens but not to an expression defect. These six crystallins are therefore probably not good candidates to explain the lens defect in CF.
3. *Crybg3, crygm5*, *crygmxb*, *cryl1*, *crym*, and *cryzl1* were not expressed, neither in SF nor in CF lens at 36 hpf (Supporting Information FigS3). We, therefore, searched for onset of expression of these six crystallins. One day later at 60 hpf, *crygm5* was strongly expressed in SF but not in CF [Fig. 5(A,B)]. This difference is very significant, yet it probably occurs too late to participate in the trigger of CF lens apoptosis, which starts at 40 hpf. The other five crystallins were not expressed in the lens at 60 hpf (Supporting Information FigS3); *crybg3* and *cryl1*, however, were expressed in the olfactory epithelium (Supporting Information FigS3M-P).
4. *crybb1c* and *crybgx* were strongly expressed in SF but not in CF lens at 36 hpf (or at extremely low level for *crybb1c*) [Fig. 4(O–R)]. In SF, these two crystallins are expressed in lens fibers cells but not in the lens epithelium [Fig. 4(O,Q), insets]. *Crybb1c* and *crybgx* are, therefore, newly identified crystallins that could account for the CF lens defect.

Characterization of *crybb1c* and *crybgx* Expression Defects in the CF Lens

As expression levels cannot be rigorously quantitatively assessed by *in situ* hybridization, we next performed RT-qPCR on RNA extracts of whole SF and CF 36 hpf larvae to confirm the absence/reduction of *crybb1c* and *crybgx* expression in CF. *Gapdh* was used as reference and *crybb1a*, which gave strong ISH signals in both morphs, served as a crystallin "control." *Crybb1a* expression was indeed easily detectable by qPCR, but was reduced by about five times in CF (Fig. 6). We suggest that this reduction is solely due to the smaller size of the lens in CF: a five-fold reduction in the volume of the lens would correspond to a 1.7-fold reduction of the radius, which fits with lens diameter measurements performed on *crybb1a* ISH at 36 hpf (SF: 42.6 μm $\pm$1.3 ($n = 12$); CF: 24.9 μm $\pm$ 2.1 ($n = 9$), which corresponds to a 1.71 ratio). Conversely, *crybgx* was hardly detectable by qPCR in CF, meaning that its expression is reduced by at least 60-fold compared to SF. This confirms that *crybgx* expression is lost in CF lens at 36 hpf. Finally, *crybb1c* expression was detected at very weak levels in CF, with an estimated 12-fold reduction compared to SF. This confirms that *crybb1c* regulation is strongly affected at 36 hpf in CF.

A complete developmental time-course of expression was then performed for *crybb1c* and *crybgx*, to assess the onset of differential expression between SF and CF, and to discern between the possibilities of an expression heterochrony or a true loss of expression. Neither of the two crystallins was expressed at 24 hpf (data not shown).

Crybb1c expression started in both morphs at 28 hpf, although *in situ* hybridization signals were very low in CF compared to SF [Fig. 7(A,B)]. Although its expression was almost lost at 36 hpf in CF [Fig. 4(P)], it was again observed at low levels at 60 hpf [Fig. 7(F)]. This suggests that the difference in *crybb1c* expression found at 36 hpf is transient.

Concerning *crybgx*, expression was detected in SF but not in CF at 28 and 32 hpf, and continued absence of expression was further observed in CF at 60 hpf

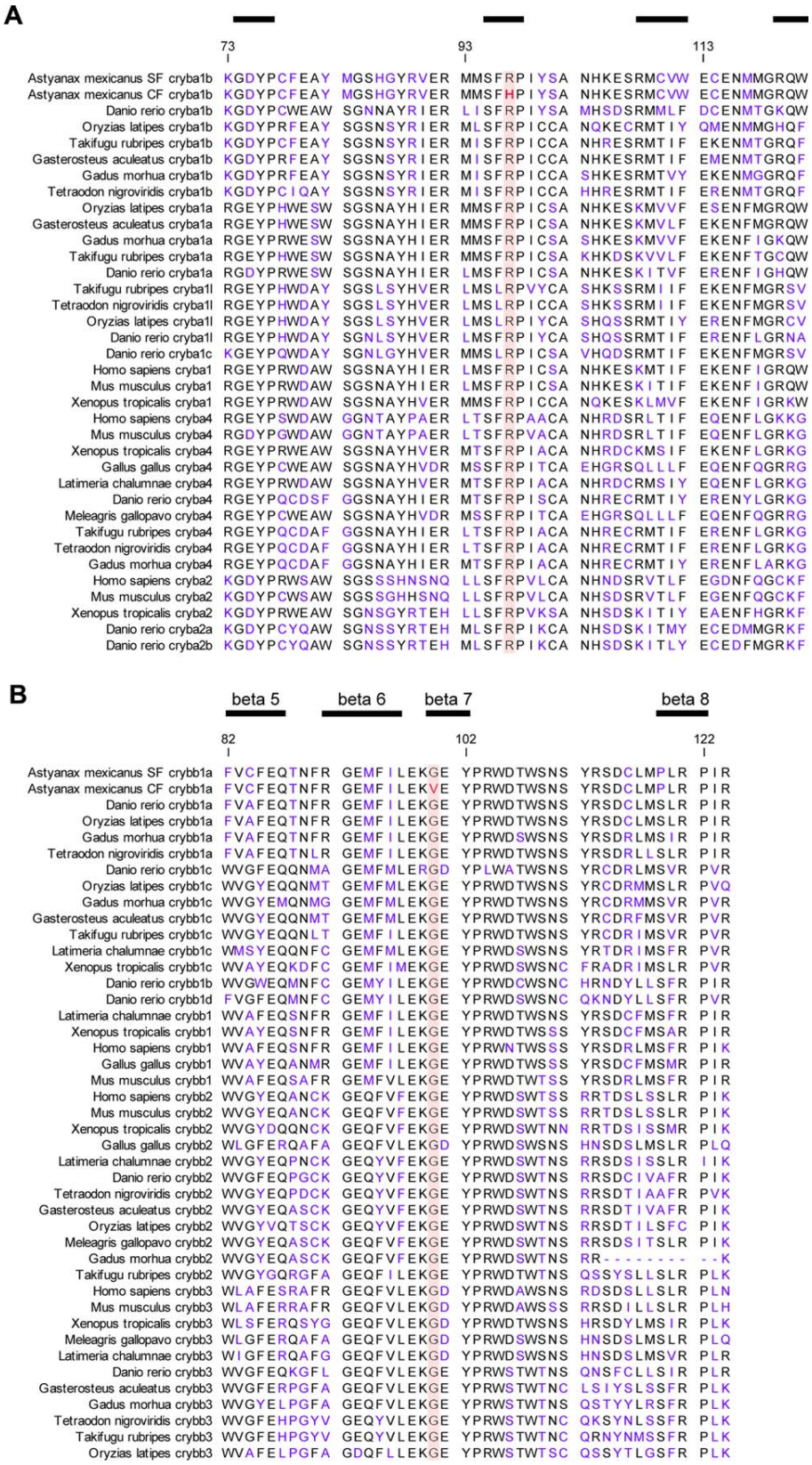


Figure 3 Mutations of crystallins at highly conserved positions. (A) Alignment of protein sequences of Vertebrate beta A crystallins, showing that CF cryba1b is mutated at a position that is otherwise 100% conserved across Vertebrates. (B) Alignment of protein sequences of Vertebrate beta B crystallins, showing that CF crybb1a is mutated at a position that is otherwise 100% conserved across Vertebrates. Black bars show the position of secondary structures (beta sheets), according to (Chaikuad et al., 2010; Madej et al., 2012; Van Montfort et al., 2003). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

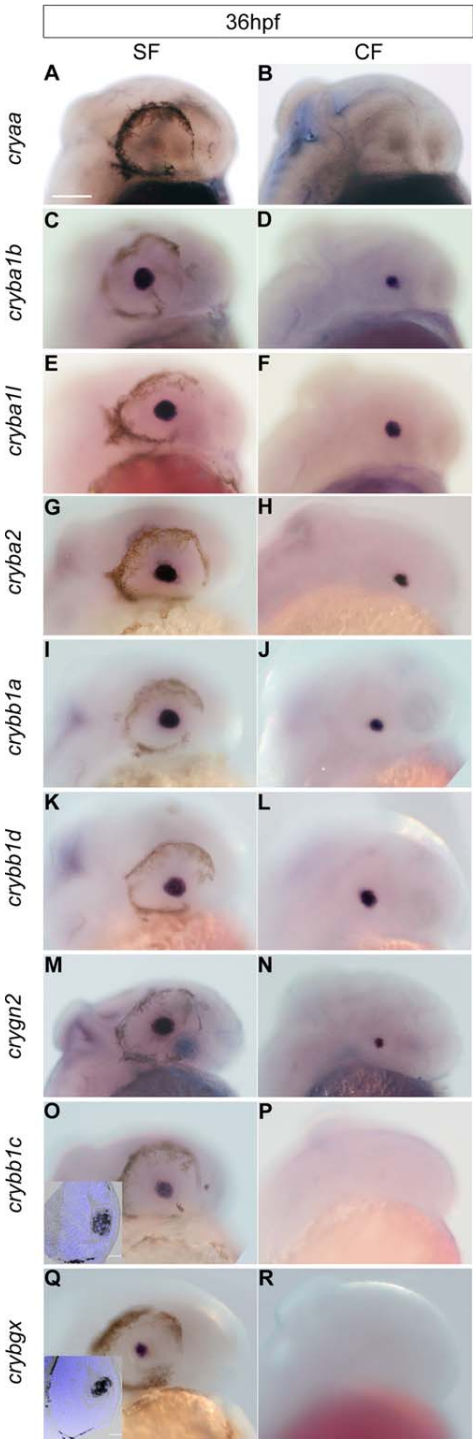


Figure 4 *In situ* hybridization screen of crystallins expression at 36 hpf. (A–R) Photographs *in toto* of whole-mount *in situ* hybridizations in lateral views for the nine indicated crystallin genes in SF and CF. Anterior is right and dorsal is up. No expression of *crybgx* is detected in CF. A very faint expression of *crybb1c* seems to be present in CF. For *crybb1c* and *crybgx*, insets show sections through SF eye, counterstained with DAPI. Scale bar in A: 100 μm. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

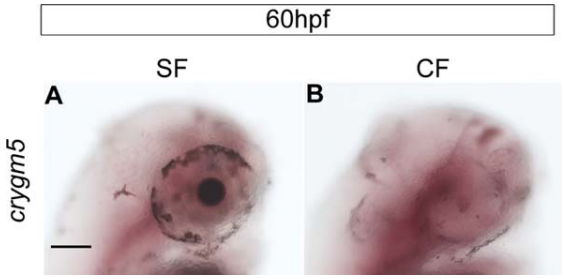


Figure 5 *In situ* hybridizations for *crygm5* at 60 hpf. Photographs *in toto* of whole-mount *in situ* hybridizations in lateral views for *crygm5* in SF and CF. Only a faint expression of *crygm5* seems to be present in CF. Scale bar in A: 100 μm. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

[Fig. 7(G–L)]. The absence of *crybgx* expression in CF is not due to a gene loss, as it is present as ENSAMXG00000006190 in the Pachón CF genome assembly (Table1). Thus, crybgx and crybb1c are novel and good candidates for having a role in CF lens defect.

Crybgx, cryaa, and *crybb1c* Loss of function Analysis

Crybgx, *cryaa*, and *crybb1c* expressions are absent in CF from very early on, before the onset of apoptosis. This makes these crystallins interesting candidates to be involved in the lens apoptosis phenotype of CF. Besides, to our knowledge, there is a complete lack of data about *crybgx* and *crybb1c* functional roles in any species. We thus decided to perform morpholino knock-down of *crybgx*, *cryaa*, and/or *crybb1c*

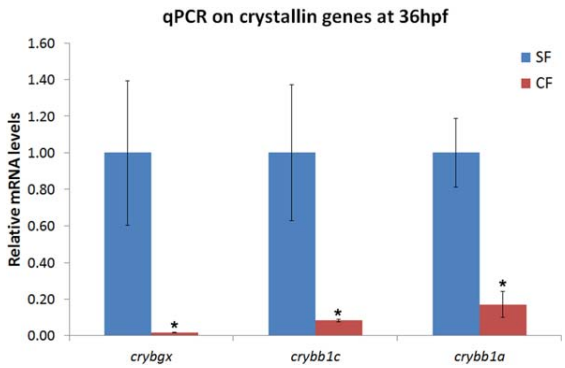


Figure 6 Expression levels of three crystallins assessed with qPCR. Relative mRNA levels of *crybgx*, *crybb1c*, and *crybb1a* in SF and CF were estimated by qPCR, using *GAPDH* as a reference gene, with triplicates. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

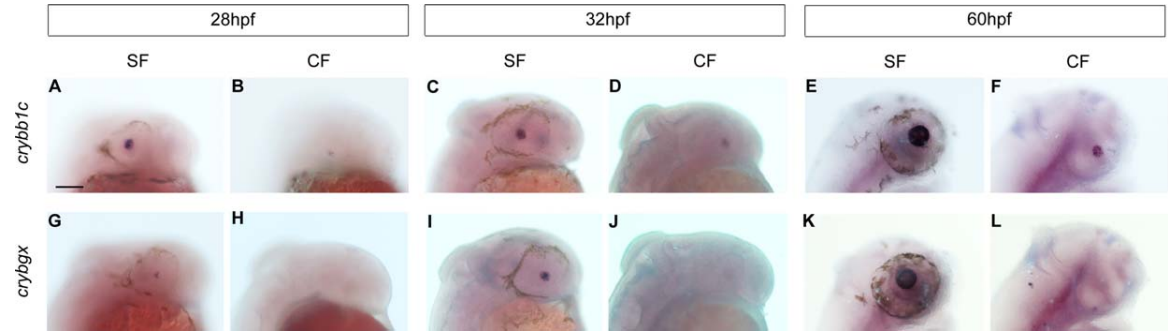


Figure 7 Time course of *crybb1c* and *crybgx* expression. Photographs *in toto* of whole-mount *in situ* hybridizations in lateral views for *crybb1c* and *crybgx* in SF and CF at 28, 32, and 60 hpf. *Crybgx* is never expressed in CF. Scale bar in A: 100 μm. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

expression in zebrafish [Fig. 8]. *D. rerio* is for now more amenable to functional studies than *Astyanax*, and some tools exist for *cryaa* loss of function study (Posner et al., 2013). We designed *crybgx* and *crybb1c* morpholinos to target the splice site at the junction of Exon 4 and Intron 4 of zebrafish *crybgx* and *crybb1c* mRNAs, respectively. They provoke the appearance of longer mRNA forms at 2 and 2.5 dpf [Fig. 8(G)], leading to a frame shift and a premature stop codon. We used a previously published *cryaa* morpholino that blocks the translation of *cryaa* mRNA and leads to complete loss of cryaa protein (Posner et al., 2013).

We investigated the effect of these knock-downs on lens apoptosis 3 and 4 days after injection, using TUNEL labeling [Fig. 8(A–D)]. There was no effect of morpholino injection at these two stages [Fig. 8(H)], showing that loss of expression of one single crystallin is not sufficient to trigger apoptosis in the zebrafish lens. We also performed triple morpholino injection, but did not detect any increase in apoptosis either [Fig. 8(E,H)]. It thus seems that the loss of three crystallins is not sufficient to trigger apoptosis in a healthy lens.

Cryaa Transgenic Lens Rescue Assay

αA-crystallin (*cryaa*) is known to have an anti-apoptotic role in cell culture (Andley, 2000) and QTL studies in *Astyanax* also suggest that *cryaa* gene is close to a QTL for eye size (Gross et al., 2008). Despite the results of the morpholino experiments, it thus seems that *cryaa* could have a role in CF eye loss. We, therefore, designed a transgenic strategy to restore *cryaa* expression in *Astyanax* CF lens, aiming to assess whether the reexpression of this single anti-apoptotic protein in the CF lens was sufficient to prevent apoptosis. The transgene included *Astyanax cryaa* coding sequence under the control of zebrafish *crybb1a* promoter, as well as a CFP reporter gene under the control of the same zebrafish *crybb1a* promoter [Fig. 9(A)]. We reasoned that as *crybb1a* is correctly expressed in *Astyanax* CF (Fig. 4), *crybb1a* promoter should be active in *Astyanax* CF lens (while a SF *cryaa* promoter could have been inactive, as *cryaa* loss of expression in CF might be due to upstream *trans* factors). Indeed, after injection of the transgene in CF eggs, CFP reporter fluorescence was observed in the lens of 2.5 dpf larvae [Fig. 9(B)], suggesting that the *crybb1a* promoter is indeed driving transgene expression. *In situ* hybridizations for *cryaa* on these F0 transgenic larvae, in which transgene expression is mosaic, show that CFP fluorescence is indeed a reliable reporter and marker for *cryaa* reexpression in CF lens cells [compare Fig. 9(C,D), Fig. 9(E,F)].

In these F0 transgenic larvae, we performed immuno-staining for activated caspase3, an apoptosis marker. This technique was preferred to TUNEL labeling because the latter may also mark the lens fiber cells losing their nuclei, which is part of their normal differentiation process (Dahm et al., 2007). Caspase3 staining appeared excluded from CF transgenic lens cells expressing *cryaa*, which are CFP-positive [Fig. 9(H–J)]. Indeed there is a very significantly reduced ratio of apoptosis among *cryaa*-positive lens fiber cells when compared to *cryaa*-negative cells in transgenic larvae, or when compared to wild-type CF lens cells [Fig. 9(K); Fisher's exact test, *p*-value : 1.45E-48 and 1.13E-41, respectively]. This shows that *cryaa* expression protects CF lens cells from apoptosis in a cell-autonomous manner.

Finally, in the positive F0 injected CF larvae, in which the expression of *cryaa* in lens cells is mosaic, the expression of the CFP reporter was progressively lost after a week of development, and there was no rescue of the eye phenotype in adults [Fig. 9(G)],

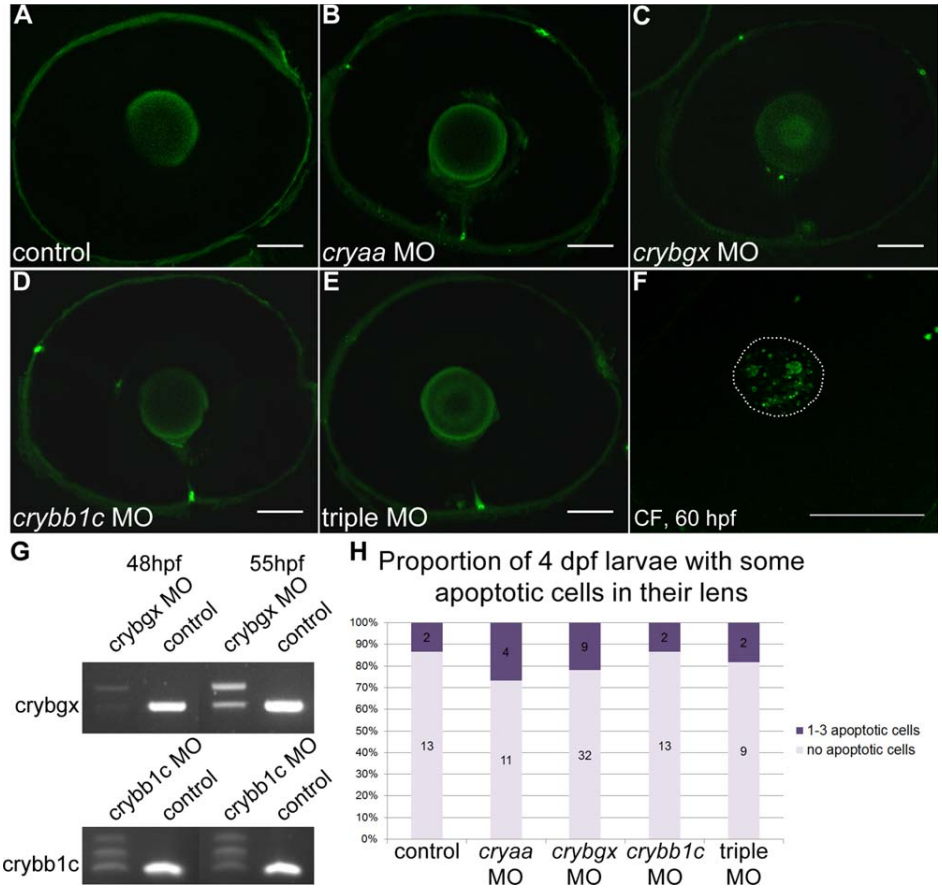


Figure 8 Effect of *crybgx*, *cryaa*, and *crybb1c* knockdown in zebrafish. Photograph of a zebrafish control eye (A), *cryaa* morphant eye (B), *crybgx* morphant eye (C), *crybb1c* morphant eye (D), or triple morphant eye (E) after TUNEL labeling at 4 dpf. For comparison, is shown a CF eye at 2.5 dpf after TUNEL labeling, undergoing massive apoptosis (F). Scale bar, 50 μm. (G) RT-PCR on mRNA extracted from control and *crybgx* or *crybb1c* MO-injected embryos shows the efficiency of the *crybgx* and *crybb1c* MOs at the two indicated stages. (H) Quantification of apoptotic cells in control and morphant lenses. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

even in those that showed strong expression in a large portion of the lens at 2.5 dpf ($n = 2$ with strong expression, $n = 8$ with medium expression). Experiments aiming at obtaining F1 individuals with ubiquitous expression of *cryaa* in the lens, and possibly a healthy lens and a rescued eye, are ongoing.

DISCUSSION

Crystallins and Their Evolution in *Astyanax* CF

We have identified 14 crystallins of various groups, plus 47 γM2d crystallins, that are expressed during development in *Astyanax*.

No α-crystallin sequence was identified, which is not so surprising, as our developmental transcriptome contained samples of young embryos and larvae (up to 2 weeks old), and in zebrafish, α crystallin content is low in young larvae and increases with age (Greiling et al., 2009). Conversely, six β-crystallins were identified from the *Astyanax* developmental transcriptome, in line with the strong representation and expression of this crystallin family during early development in zebrafish (Greiling et al., 2009). Three γ-crystallins, two βγ crystallins, one λ-crystallin, one μ-crystallin, and one ζ-crystallin were identified as well, therefore covering the large diversity of the crystallins.

Concerning the γM2d crystallins expansion, it is not clear whether this results from an expansion of this subfamily [independent from the one in zebrafish, see *crygm2d* expansion in zebrafish on Fig. 1(B)] or whether the *γM2d* crystallin contigs correspond to alternative splice variants of only a few

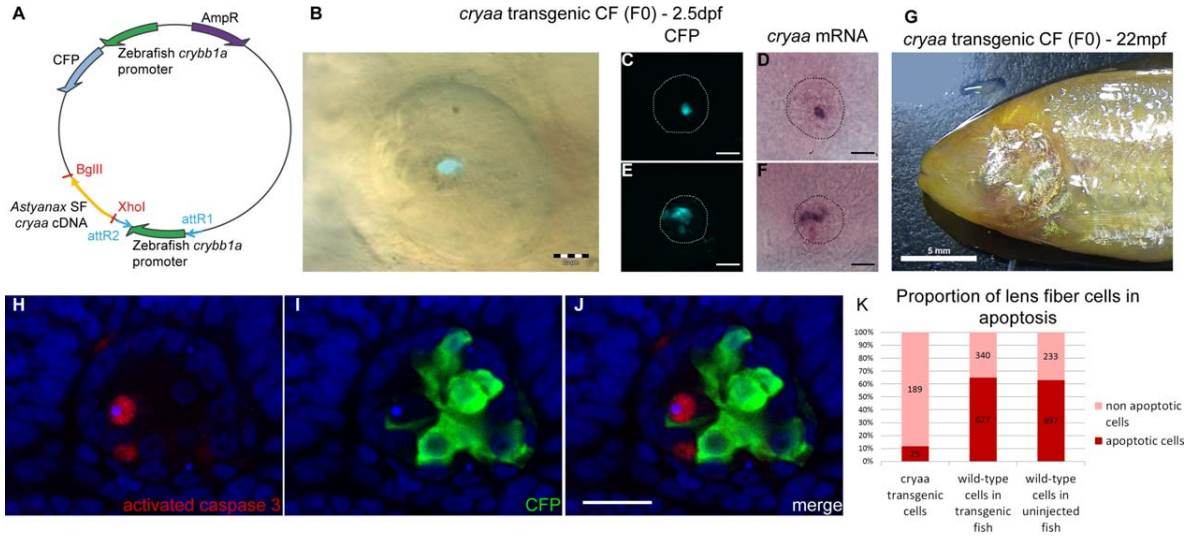


Figure 9 Rescue of *cryaa* expression in CF by transgenesis. (A) Scheme of the transgene. (B) Photograph of a transgenic CF larva, showing mosaic CFP fluorescence in the lens. (C and E) Examples of two larvae showing mosaic CFP fluorescence in the lens. (D and F) Corresponding *in situ* hybridization for *cryaa* on the same CF larvae, showing the same pattern as CFP fluorescence. (G) Photograph of a transgenic CF adult, showing the lack of eye rescue. The fish shown is the same as in B at larval stage. (H–J) Confocal images of the lens of a transgenic larva after immunohistochemistry anti-GFP and anti-activated caspase3 and DAPI staining, showing apoptosis (H), transgene expression (I) and merged image (J). This individual has a particularly high number of CFP+ cells and low number of apoptotic cells. (K) Quantification of apoptosis in transgenic versus wild-type lens fiber cells. Numbers on the graphs indicate the number of cells in each category. Scale bar 20 μm, except for G.

genes. At least some of them must be different paralogs as several hits were retrieved in the preliminary genome assembly. It is interesting to note that the expression of one *crygm2d* crystallin has already been studied in *Astyanax* (Jeffery et al., 2000) under the name « γM-crystallin », and revealed no differences between SF and CF.

From our molecular evolution analysis, we conclude that CF crystallins have a lower polymorphism level than SF but accumulate more nonsynonymous fixed mutations than SF, and some of the mutations found in CF occur at highly conserved positions. Moreover, three crystallins (*crybb1a*, *cryba1b*, *crybg3*) carry amino-acid changing point mutations in CF, some of them occuring at highly conserved positions across vertebrates. This pattern suggests relaxed purifying selection on CF crystallin genes, which are probably useless in the dark, or a general decrease in selection intensity in the small CF population (genetic drift).

Interestingly, similar—although less numerous—defects are reported for the crystallins of the naked mole rat *Heterocephalus glaber*, an underground living, microphtalmic, and blind rodent. In the genome of this animal, *CRYBA4* and *CRYBB3* present premature stop codons, *GRYGS* carries a mutation (Kim et al., 2011). In parallel to these molecular defects, the naked mole rat lens is not apoptotic, but it is indeed malformed (Nikitina et al., 2004). Although we did not find any premature stop codon in CF crystallins, this may suggest that the lens defects are less advanced in the mole rat than in the CF.

Crystallins Expression Defects in *Astyanax* CF

We have chosen to study expression of *Astyanax* crystallins at 36 hpf, that is, slightly before the onset of lens apoptosis in CF (Alunni et al., 2007). Although a study has suggested that lens apoptosis might even start earlier, at 24 hpf (i.e., when larvae hatch; Jeffery and Martasian, 1998), it appears that most apoptotic cells at this stage reside in the surface ectoderm (presumptive cornea), and not in the lens. Besides, there are also some apoptotic cells in the surface ectoderm of SF at this stage (Jeffery and Martasian, 1998). Indeed, in zebrafish, apoptosis also occurs in the surface ectoderm and could be involved in the reorganization of the lens epithelium in a single

cell layer (Greiling et al., 2010). The earliest time point for which specific apoptosis inside the CF lens is certain is thus 40 hpf (Alunni et al., 2007).

Around the onset of eye degeneration in *A. mexicanus* CF, at least four crystallins are not expressed correctly in the lens: *cryaa*, *crybb1c*, and *crybgx* show strongly reduced or no expression at 36 hpf; and *crygm5* is not expressed at 60 hpf. The CF lens, therefore, misses four of its major differentiation proteins.

The absence of *cryaa* expression in CF at 36 hpf is not in total agreement with Strickler's study, where *cryaa* was found to be expressed faintly in both SF and CF at 36 hpf, and not expressed anymore in CF at 48 hpf (Strickler et al., 2007b). This discrepancy can be due to the fact that embryos in this study were raised at 25°C instead of 23°C, which makes developmental stages difficult to compare (Hinaux et al., 2011). Of note, this previous study found the same result (equivalent expression in SF and CF) for the crystallin *crybb1d*, which they named *crybb1* in their article (Strickler et al., 2007b).

Very little is known about *crybb1c* and *crybgx* in other species. In zebrafish, *crybb1c* is one of the most abundantly expressed crystallins throughout development and from 4.5 dpf (days post fertilization) onward, and *crybgx* is expressed at low levels, starting around 4.5 dpf (Greiling et al., 2009).

For discussion purpose, it is interesting to note that *crybb1c* (but not the other beta B crystallins *crybb1a* and *crybb1d)* is subject to dysregulation in CF. This is despite the fact that the expression of these crystallins, very close phylogenetically, are probably under the control of the same transcription factors, such as Maf and Pitx3 (Ishibashi and Yasuda, 2001; Hooker et al., 2012; Templeton et al., 2013). Thus, it seems likely that *crybb1c* expression modification in CF is due to alterations in the regulatory sequences of *crybb1c*, but not to upstream factors.

Crystallin Functional Defects in *Astyanax* CF: A Role for Cryaa

The test of the effect of crystallin expression differences through zebrafish morpholino experiments show that *crybgx*, *crybb1c*, and *cryaa* knockdowns, either single or multiple, are not sufficient to trigger lens apoptosis. Maybe it would be necessary to reproduce both the CF losses of expression by morpholinos (*crybgx*, *cryaa*, and *crybb1c*) and the CF mutations by genome editing (*cryba1b*, *crybb1a*) to phenocopy lens apoptosis.

Our result of *cryaa* morpholino knock-down in zebrafish is in contradiction with a recent report, in which *cryaa* morpholino in *Astyanax* SF seems to induce an increase of TUNEL labeling in the lens (Ma et al., 2014). Several possibilities can explain the difference of results between the two studies: the most obvious is that cryaa downregulation in *Astyanax* SF may have a different effect than in zebrafish. It is also possible, although unlikely considering published results (Posner et al., 2013), that the translation-blocking morpholino we used in zebrafish is less efficient than the splice-blocking morpholino used in *Astyanax*. Of note in our experiments, concerning *crybgx* and *crybb1c*, morpholinos do not eliminate completely the correct form of mRNA, and there must be some active protein remaining in lens cells. This could also explain why no apoptosis occurs.

Contrarily to morpholinos knock-down experiments which are sometimes difficult to interpret, our results using transgenic rescue approach are clear-cut. They demonstrate that *cryaa* protects CF lens cells from apoptosis cell-autonomously. Thus, in the CF genetic background, *cryaa* reexpression is sufficient to prevent lens apoptosis, which is thought to be the trigger for eye degeneration in CF (Yamamoto and Jeffery, 2000). However, *cryaa* reexpression in the *Astyanax* CF lens by transgenesis does not rescue the adult eye in FO injected mosaic animals. It will be interesting to analyze the development of the eye in F1 transgenic larvae with homogeneous expression of *cryaa* in the lens.

CONCLUSION

In total, 7 out of the 14 crystallins that are identified and studied in this article are modified in CF in one way or another, expression-wise or sequence-wise. Our survey, therefore, points to a global defect in lens "crystallin biology" in CF, which occurs under the form of apparently little but multiple modifications. Among them, reduced *cryaa* expression in the CF lens is one of the major defects that could account for apoptosis and explain eye degeneration.

We thank Magalie Bouvet and Stéphane Père for taking care of our *Astyanax* colony, members of the AMAGEN platform for providing us with zebrafish eggs, Joanne Edouard for her guidance on molecular biology and for providing material to test GFP antibody, Pierre Affaticati for

the DAPI staining protocol, and all the members of the DECA group for discussions and suggestions. We also thank an anonymous reviewer for suggestions that improved the manuscript. This work has benefited from the facilities and expertise of the high throughput sequencing platform, the QPCR platform and the imaging platform of IMAGIF (Centre de Recherche de Gif - www.imagif.cnrs.fr). We thank Céline Noirot (GenoToul Bioinformatics Platform) for assembling the Illumina data to be published elsewhere. The authors declare no conflict of interest.

REFERENCES

Alunni A, Menuet A, Candal E, Pénigault J-B, Jeffery WR, Rétaux S. 2007. Developmental mechanisms for retinal degeneration in the blind cavefish Astyanax mexicanus. J Comp Neurol 505:221–233.

Andley UP. 2000. Differential protective activity of alpha A- and alpha B-crystallin in lens epithelial cells. J Biol Chem 275:36823–36831.

Behrens M. 1998. Cloning of the αA-crystallin genes of a blind cave form and the epigean form of Astyanax fasciatus: A comparative analysis of structure, expression and evolutionary conservation. Gene 216:319–326.

Chen J-Y, Chang B-E, Chen Y-H, Lin CJ-F, Wu J-L, Kuo C-M. 2001. Molecular cloning, developmental expression, and hormonal regulation of zebrafish (Danio rerio) β crystallin B1, a member of the superfamily of β crystallin proteins. Biochem Biophys Res Commun 285:105–110.

Dahm R, Schonthaler HB, Soehn AS, van Marle J, Vrensen GFJM. 2007. Development and adult morphology of the eye lens in the zebrafish. Exp Eye Res 85:74–89.

de Jong WW, Leunissen JA, Voorter CE. 1993. Evolution of the alpha-crystallin/small heat-shock protein family. Mol Biol Evol 10:103–126.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43:491–498.

Elipot Y, Legendre L, Père S, Sohm F, Rétaux S. 2014. *Astyanax* transgenesis and husbandry: How cavefish enters the laboratory. Zebrafish 11:291–299.

Fagerholm PP, Philipson BT, Lindström B. 1981. Normal human lens—the distribution of protein. Exp Eye Res 33: 615–620.

Goishi K. 2006. A-crystallin expression prevents γ-crystallin insolubility and cataract formation in the zebrafish cloche mutant lens. Development 133:2585–2593.

Greiling TM, Houck SA, Clark JI. 2009. The zebrafish lens proteome during development and aging. Mol Vis 15:2313.

Greiling TMS, Clark JI. 2012. New insights into the mechanism of lens development using zebra fish. In: International Review of Cell and Molecular Biology. Vol. 296. USA: Elsevier, pp 1–61.

Greiling TMS, Aose M, Clark JI. 2010. Cell fate and differentiation of the developing ocular lens. Invest. Ophthalmol Vis Sci 51:1540–1546.

Gross JB, Protas M, Conrad M, Scheid PE, Vidal O, Jeffery WR, Borowsky R, et al. 2008. Synteny and candidate gene prediction using an anchored linkage map of Astyanax mexicanus. Proc Natl Acad Sci USA 105:20106.

Hinaux H, Pottin K, Chalhoub H, Père S, Elipot Y, Legendre L, Rétaux S. 2011. A developmental staging table for Astyanax mexicanus surface fish and Pachón cavefish. Zebrafish 8:155–165.

Hinaux H, Poulain J, Da Silva C, Noirot C, Jeffery WR, Casane D, Rétaux S. 2013. De novo sequencing of astyanax mexicanus surface fish and Pachón cavefish transcriptomes reveals enrichment of mutations in cavefish putative eye genes. PLoS ONE 8:e53553.

Hooker L, Smoczer C, KhosrowShahian F, Wolanski M, Crawford MJ. 2012. Microarray-based identification of Pitx3 targets during Xenopus embryogenesis. Dev Dyn 241:1487–1505.

Ishibashi S, Yasuda K. 2001. Distinct roles of *maf* genes during *Xenopus* lens development. Mech Dev 101:155–166.

Javonillo R, Malabarba LR, Weitzman SH, Burns JR. 2010. Relationships among major lineages of characid fishes (Teleostei: Ostariophysi: Characiformes), based on molecular sequence data. Mol Phylogenet Evol 54:498–511.

Jeffery W, Strickler A, Guiney S, Heyser D, Tomarev S. 2000. Prox 1 in eye degeneration and sensory organ compensation during development and evolution of the cavefish Astyanax. Dev Genes Evol 210:223–230.

Jeffery WR. 2008. Emerging model systems in evo-devo: Cavefish and microevolution of development. Evol Dev 10:265–272.

Jeffery WR. 2009. Regressive evolution in *Astyanax* cavefish. Annu Rev Genet 43:25–47.

Jeffery WR, Martasian DP. 1998. Evolution of eye regression in the cavefish *Astyanax*: Apoptosis and the *Pax-6* gene. Integr Comp Biol 38:685–696.

Kim EB, Fang X, Fushan AA, Huang Z, Lobanov AV, Han L, Marino SM, et al. 2011. Genome sequencing reveals insights into physiology and longevity of the naked mole rat. Nature 479:223–227.

Litt M. 1998. Autosomal dominant congenital cataract associated with a missense mutation in the human alpha crystallin gene CRYAA. Hum Mol Genet 7:471–474.

Ma L, Parkhurst A, Jeffery WR. 2014. The role of a lens survival pathway including sox2 and αA-crystallin in the evolution of cavefish eye degeneration. EvoDevo 5:28.

Mackay DS, Andley UP, Shiels A. 2003. Cell death triggered by a novel mutation in the alphaA-crystallin gene underlies autosomal dominant cataract linked to chromosome 21q. Eur J Hum Genet 11:784–793.

Madej T, Addess KJ, Fong JH, Geer LY, Geer RC, Lanczycki CJ, Liu C, et al. 2012. MMDB: 3D structures and macromolecular interactions. Nucleic Acids Res 40: D461–D464.

McGaugh SE, Gross JB, Aken B, Blin M, Borowsky R, Chalopin D, Hinaux H, et al. 2014. The cavefish genome reveals candidate genes for eye loss. Nat Commun 5:5307.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20: 1297–1303.

Nikitina NV, Maughan-Brown B, O'Riain MJ, Kidson SH. 2004. Postnatal development of the eye in the naked mole rat (Heterocephalus glaber). Anat Rec A Discov Mol Cell Evol Biol 277:317–337.

Posner M, Skiba J, Brown M, Liang JO, Nussbaum J, Prior H. 2013. Loss of the small heat shock protein αA-crystallin does not lead to detectable defects in early zebrafish lens development. Exp Eye Res 116:227–233.

Rétaux S, Casane D. 2013. Evolution of eye development in the darkness of caves: Adaptation, drift, or both? EvoDevo 4:26.

Saitou N, Nei M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425.

Strickler AG, Yamamoto Y, Jeffery WR. 2007a. The lens controls cell survival in the retina: Evidence from the blind cavefish Astyanax. Dev Biol 311:512–523.

Strickler AG, Byerly MS, Jeffery WR. 2007b. Lens gene expression analysis reveals downregulation of the anti-apoptotic chaperone αA-crystallin during cavefish eye degeneration. Dev Genes Evol 217:771–782.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731–2739.

Templeton JP, Wang X, Freeman NE, Ma Z, Lu A, Hejtmancik F, Geisert EE. 2013. A crystallin gene network in the mouse retina. Exp Eye Res 116:129–140.

Van Montfort RLM, Bateman OA, Lubsen NH, Slingsby C. 2003. Crystal structure of truncated human betaB1-crystallin. Protein Sci. Publ. Protein Soc 12:2606–2612.

Wang H, Kesinger JW, Zhou Q, Wren JD, Martin G, Turner S, Tang Y, et al. 2008. Identification and characterization of zebrafish ocular formation genes. Genome Natl Res Counc Can Génome Cons Natl Rech Can 51: 222–235.

Wistow G, Wyatt K, David L, Gao C, Bateman O, Bernstein S, Tomarev S, et al. 2005. γN-crystallin and the evolution of the βγ-crystallin superfamily in vertebrates: γN-crystallin. FEBS J 272:2276–2291.

Yamamoto Y, Jeffery WR. 2000. Central role for the lens in cave fish eye degeneration. Science 289:631–633.

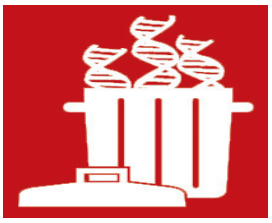
Zuckerlandl E, Pauling L. 1965. Evolutionary divergence and convergence in proteins. In: Bryson V, Vogel HJ, editors. Evolving Genes and Proteins, New York: Academic Press, pp 97–166.

B.2 L'apophénie d'ENCODE ou Pangloss examine le génome humain

médecine/sciences 2015 ; 31 : 680-6

L'apophénie d'ENCODE ou Pangloss examine le génome humain

Didier Casane[1,2], Julien Fumey[1], Patrick Laurenti[1,2]



[1] Laboratoire Évolution, génomes, comportement, écologie, CNRS université Paris-Sud UMR 9191, IRD UMR 247, Avenue de la Terrasse, bâtiment 13, boîte postale 1, 91198 Gif-sur-Yvette, France ;
[2] université Paris-Diderot, Sorbonne Paris-Cité, Paris, France.
patrick.laurenti@egce.cnrs-gif.fr
didier.casane@egce.cnrs-gif.fr
julien.fumey@egce.cnrs-gif.fr

> En septembre 2012, les principaux résultats du projet ENCODE (*Encyclopedia of DNA Elements*) furent publiés sous la forme de 30 articles, dont une grande partie dans les deux plus prestigieuses revues scientifiques généralistes *Nature* et *Science*. Ce projet, qui avait mobilisé des centaines de chercheurs et des centaines de millions de dollars, avait permis de renverser un des paradigmes les mieux établis portant sur l'évolution des génomes : notre génome ne compterait pas 80 % d'ADN « poubelle » mais il serait au contraire à 80 % utile. Il ne restait plus qu'à savoir à quoi ! Dans le microcosme de la biologie évolutive, il y eut quelques réactions amusées, voire agacées, face à cette revitalisation d'un adaptationnisme naïf. Des réfutations argumentées furent publiées, mais qui ne pouvaient rivaliser avec la publicité faite à ENCODE. En 2014, une nouvelle série d'articles présentaient les avancées récentes du projet. Étonnamment, on n'y trouvait plus aucune trace de l'extraordinaire découverte de 2012, sans que cette « disparition » d'un résultat majeur ne soit commentée. Une forme de rétractation par omission sans doute. Mais le mal était fait, l'affirmation « 80 % d'ADN utile » d'ENCODE est désormais intégrée par de nombreux biologistes comme le nouveau cadre d'étude des génomes, ou pour le moins comme une hypothèse alternative valable. Il nous semble donc indispensable de rappeler quelques concepts généraux qui expliquent l'architecture et le fonctionnement des génomes, concepts qui, à ce jour, n'ont pas été réfutés et qui permettent d'éviter l'écueil que représente l'approche panglossienne d'ENCODE. <

Le 7 septembre 2012, il était proclamé dans la revue *Science* : « *ENCODE project writes eulogy for junk DNA* » [1], ce qui peut être traduit par : « ENCODE écrit l'éloge funèbre de l'ADN poubelle ». L'ADN poubelle est mort, vive l'ADN fonctionnel ! L'invraisemblable nombre de fragments de transposons, de pseudogènes qui ne peuvent permettre la production d'ARN fonctionnels, tout ce fatras d'ADN qui ne ressemble à rien de nécessaire et qui apparaît et disparaît des génomes au fil du temps... Et bien, non, ce ne serait pas des déchets accumulés dans les génomes, mais de l'ADN fonctionnel ! Il ne restait plus alors qu'à identifier plus précisément ces fonctions [2]. Le consortium ENCODE annonçait avoir « renversé un paradigme », et il fallait se préparer à réécrire les manuels universitaires [1]. Mais, au fait, le projet ENCODE, c'est quoi au juste ? C'est l'*Encyclopedia of DNA Elements*. Il existe aussi modENCODE (*model organism ENCODE*). Le but de ces deux projets est d'identifier tous les éléments fonctionnels du génome humain et des génomes d'organismes modèles (drosophiles et nématodes). La stratégie retenue est de compiler, de la façon la plus exhaustive et précise possible, le niveau de transcription, l'association à des facteurs de transcription, l'organisation de la chromatine et les modifications des histones, à l'échelle de l'ensemble du génome, et ce pour différents types cellulaires et pour différentes espèces. C'est un travail colossal et une base de données de cette nature est potentiellement de première importance pour un très grand nombre de biologistes. Les investissements humains et financiers furent à l'échelle du projet : 442 chercheurs et 288 millions de dollars [1]. Mais peut-on aujourd'hui investir autant de moyens pour générer une base de données énorme sans obtenir un résultat biologique extraordinaire ? On peut facilement répondre oui en termes d'avancement des sciences,

m/s n° 6-7, vol. 31, juin-juillet 2015
DOI : 10.1051/medsci/20153106023

mais non en termes de publicité. Un fort investissement réclame un résultat fort. Ainsi, la NASA nous a accoutumés à des découvertes fracassantes, comme la démonstration de la présence de vie sur Mars ou de l'existence de molécules d'ADN contenant de l'arsenic… (➜). Ces découvertes furent réfutées, mais la NASA se doit de justifier de temps en temps son colossal budget. L'objectif d'une stratégie très coûteuse est en effet de découvrir ce qui n'est pas à la portée des chercheurs ordinaires, c'est-à-dire passablement isolés et disposant de peu de moyens. De fait, ENCODE a produit des milliers de jeux de données de grande taille et publié un grand nombre d'articles en 2012. Malheureusement, l'analyse de ces masses étourdissantes de données ne fut pas guidée par des questions scientifiques précises. Les graphiques colorés abondaient, mais il ne ressortait rien de vraiment nouveau de tout ça. Et c'était tout à fait normal puisque ce n'était pas le but du projet. Il fallait toutefois, et à toute force, obtenir un résultat percutant et un paradigme renversé. Ce fut donc « 80 % de l'ADN humain est fonctionnel ». Sachant qu'auparavant, il n'avait été identifié qu'environ 10 % à 15 % de génome utile [3, 4], il y aurait donc 70 % de matière noire en attente de fonction et à explorer [5]. Fallait-il être bête pour croire que l'ADN humain, la molécule la plus parfaite de la plus parfaite des créations de la nature n'abritait pas quelques fonctions cachées ? La publicité faite à cette découverte sortit alors largement du cadre des spécialistes de la génomique. Certains dès lors pouvaient aussi « légitimement » s'interroger sur la nécessité du maintien de la « SMALL » science, comme la biologie évolutive ou la génétique, reposant sur peu de chercheurs qui disposent de peu de moyens [6]. Ces affirmations semèrent la consternation chez les spécialistes du domaine. Mais, accoutumés à ce que la biologie évolutive soit l'objet de spéculations hasardeuses, la plupart des évolutionnistes haussèrent les épaules, et retournèrent à leurs chères études sous financées. Cependant, quelques uns, et non des moindres, firent l'effort de publier des analyses critiques très argumentées [7-10]. Dan Graur s'est fait le champion des pourfendeurs des inepties propagées à grande échelle autour de cette découverte. Adepte d'un humour grinçant, auteur de chroniques hilarantes sur la « BIG » science en général et ENCODE en particulier sur son blog dont nous conseillons vivement la lecture[1], il écrivit en 2013 un article très caustique [11] qui ramena la découverte faite par ENCODE à ce qu'elle est : un affichage publicitaire.

(➜) Voir la Brève de Sophia Häfner, *m/s* n° 2, février 2011, page 145

En 2014, nouvelle livraison d'articles par le consortium ENCODE dans la prestigieuse revue *Nature*. Les articles sont toujours aussi ennuyeux et toujours aussi bien illustrés par des graphiques toujours aussi colorés. L'information mise en avant est que la masse de données croît très régulièrement [12]. La base de données a plus que doublé de volume. Bravo ! En revanche, les conclusions des articles de 2014 semblent bien insipides. Qu'on en juge par la platitude de la conclusion de la comparaison des transcriptomes humains / drosophiles / nématodes : « *nos résultats soulignent l'importance de comparer des organismes modèles distants avec les humains afin de distinguer les principes biologiques* [*sic* !] *conservés des adaptations spécifiques aux lignées*[2] » [13]. Finalement, l'information principale n'apparaît qu'en creux : l'utilité de 80 % du génome n'est plus évoquée nulle part. Au contraire, comme l'a fait remarquer Dan Graur dans une chronique récente[3], le *News and Views* associé aux cinq publications de *Nature* prend soin de souligner qu'aucun de ces articles ne permet de conclure quant à la fonctionnalité des séquences étudiées [12]. En somme, les interprétations panglossiennes font une sortie de scène aussi discrète que leur entrée avait été tonitruante. Pourtant, ce même *News and Views* rappelle en introduction : « Les projets Encyclopédie des éléments d'ADN (ENCODE) et ENCODE des organismes modèles (modENCODE) furent lancés dans le but d'identifier tous les éléments fonctionnels des génomes de l'espèce humaine, de la mouche *Drosophila melanogaster* et du ver *Caenorhabditis elegans* »[4]. Si le but est toujours d'identifier l'ensemble des éléments fonctionnels des génomes, la disparition de la « fonction » de l'ADN poubelle et la critique des interprétations de 2012 auraient dû faire les grands titres des commentaires. Cette étonnante omission nous fait craindre que le ver ne soit toujours dans le fruit. Aussi, il nous semble important d'enfoncer encore quelques clous dans le cercueil du Dr Pangloss afin qu'il ne revienne hanter les interprétations des analyses génomiques à venir. De façon plus générale, nous nous interrogeons sur les dérives de la communication des grands projets scientifiques de type « BIG science », et plaidons pour une plus grande reconnaissance des bases théoriques qu'apportent les « SMALL sciences » comme la génétique ou la biologie évolutive.

L'apophénie d'ENCODE : ils voient des fonctions partout

Revenons tout d'abord à l'affirmation de 2012. Le génome serait à 80 % fonctionnel ! Bon, pourquoi pas, mais quelle est la définition de « fonction » pour ENCODE ? Dans son acception courante, une fonction implique une chose « utile » à quelque chose. Pour la plupart des biologistes, un fragment d'ADN possède une fonction au niveau de l'organisme qui le contient lorsqu'il est impliqué dans quelque processus utile à la vie de cet organisme et à

[1] http://judgestarling.tumblr.com/

[2] « *Overall, our results underscore the importance of comparing divergent model organisms to human to highlight conserved biological principles (and disentangle them from lineage-specific adaptations)* » [13].

[3] http://judgestarling.tumblr.com/post/95976986801/encode-2014-versus-encode-2012-with-translations

[4] « *In an effort to identify all functional elements in the genomes of humans,* Drosophila melanogaster *flies and* Caenorhabditis elegans *worms, the Encyclopedia of DNA Elements (ENCODE) and the model organism ENCODE (modENCODE) research projects were launched* ».

« Le melon a été divisé en tranches par la nature afin d'être mangé en famille ; la citrouille, étant plus grosse, peut être mangée avec les voisins. »
Jacques-Henri Bernardin de Saint-Pierre, *Études de la Nature,* 1784.

Pangloss enseignait la métaphysico-théologo-cosmolo-nigologie. Il prouvait admirablement qu'il n'y a point d'effet sans cause [...].
« Il est démontré, disait-il, que les choses ne peuvent être autrement : car, tout étant fait pour une fin, tout est nécessairement pour la meilleure fin. Remarquez bien que les nez ont été faits pour porter des lunettes ; aussi avons-nous des lunettes. Les jambes sont visiblement instituées pour être chaussées, et nous avons des chausses. Les pierres ont été formées pour être taillées et pour en faire des châteaux ; aussi monseigneur a un très beau château : le plus grand baron de la province doit être le mieux logé ; et les cochons étant faits pour être mangés, nous mangeons du porc toute l'année. Par conséquent, ceux qui ont avancé que tout est bien ont dit une sottise : il fallait dire que tout est au mieux. » [...]
Et Pangloss disait quelquefois à Candide : « Tous les événements sont enchaînés dans le meilleur des mondes possibles ; car enfin, si vous n'aviez pas été chassé d'un beau château à grands coups de pied dans le derrière pour l'amour de Mlle Cunégonde, si vous n'aviez pas été mis à l'Inquisition, si vous n'aviez pas couru l'Amérique à pied, si vous n'aviez pas donné un bon coup d'épée au baron, si vous n'aviez pas perdu tous vos moutons du bon pays d'Eldorado, vous ne mangeriez pas ici des cédrats confits et des pistaches. – Cela est bien dit, répondit Candide, mais il faut cultiver notre jardin. »

Voltaire, *Candide ou l'Optimisme*, 1759.

la transmission de son patrimoine génétique [14]. Mais ENCODE utilise une définition plus inattendue : le fragment d'ADN doit être transcrit et/ou impliqué dans des interactions avec des protéines connues pour leurs rôles dans la régulation de la transcription, et ce dans au moins un des types cellulaires testés. Donc, si un fragment d'ADN est transcrit ou s'il sert à transcrire un autre fragment d'ADN, c'est un fragment d'ADN fonctionnel. Il est depuis longtemps établi que l'essentiel du génome est transcrit [15] et que les séquences sur lesquelles se fixent les facteurs de transcription sont de très petite taille et donc présentes en très nombreuses copies dans les génomes [16]. La philosophie d'ENCODE, simpliste, est la suivante : si ça existe, c'est bien pour quelque chose. Pourquoi pas, mais encore faut-il le démontrer. Surtout dans le cas de l'ADN, pour lequel les preuves du contraire s'accumulent depuis trente ans [10]. L'erreur d'ENCODE fut de tenter de ressusciter des idées sur l'évolution qui furent progressivement abandonnées au cours de la fin du XX^{e} siècle. En effet, dans la première moitié du siècle dernier, la sélection a parfois fait figure de mécanisme tout puissant qui optimise toutes les caractéristiques des organismes, génome compris, et maximise ainsi la valeur sélective. Il n'y a pas de but prédéfini, comme dans une conception créationniste, mais un état optimal vers lequel tend la population sous l'effet de la sélection. En somme, tout va pour le mieux, ou pour le moins tend vers le mieux, dans le meilleur des mondes possibles. C'est ce qu'on appelle le paradigme panglossien : tout caractère observé ne peut s'expliquer que par son utilité (voir *Encadré*). Cette vision panadaptationniste de l'évolution fut énergiquement remise en cause au cours de la deuxième moitié du XX^{e} siècle, en particulier par S.J. Gould and R.C. Lewontin [17]. Une large part de la recherche en biologie évolutive depuis plus de quarante ans a donc consisté, et avec succès, à montrer les limites d'une vision « adaptationniste » du monde vivant. Tout d'abord, la plupart des mutations sont neutres : elles n'améliorent ni ne détériorent la valeur sélective des organismes. Ces modifications du génome disparaissent ou se fixent au gré du hasard. C'est donc essentiellement la dérive génétique, c'est-à-dire les fluctuations aléatoires des fréquences alléliques au cours des générations[5] qui fait évoluer l'architecture des génomes [18]. Par ailleurs, même pour les locus soumis à sélection, la sélection naturelle n'est pas toute puissante : la dérive génétique peut contrecarrer la sélection et permettre la fixation d'une mutation délétère et la disparition d'une mutation avantageuse [19]. Enfin, la sélection agit à différents niveaux d'organisation, entre gènes dans un génome, entre organismes dans une population, entre populations dans une espèce, voire entre espèces. C'est l'origine de nombreux conflits génétiques qui entraînent la mise en place de compromis qui ne sont pas des solutions optimales à chaque niveau d'organisation. Ainsi, à un locus, un allèle peut se fixer au détriment d'un autre allèle tout en abaissant la valeur sélective de l'organisme qui le porte [20]. L'architecture et le fonctionnement des génomes ne sont donc pas le résultat de la maximisation d'un bien commun.

Et pourquoi pas 100 % ?

L'ADN humain serait à 80 % fonctionnel, en utilisant la définition et les données d'ENCODE. Si c'est vrai, alors il l'est très vraisemblablement à 100 %, car tous les types cellulaires et tous les environnements cellulaires n'ont pas été testés. Par ailleurs, il fut souligné assez malicieusement que l'activité de réplication n'a pas été prise en compte, alors que 100 % du génome participe à sa propre réplication. Il n'y a donc rien à jeter dans notre génome. Encore plus malicieusement, Dan Graur

[5] http://fr.wikipedia.org/wiki/Dérive_génétique

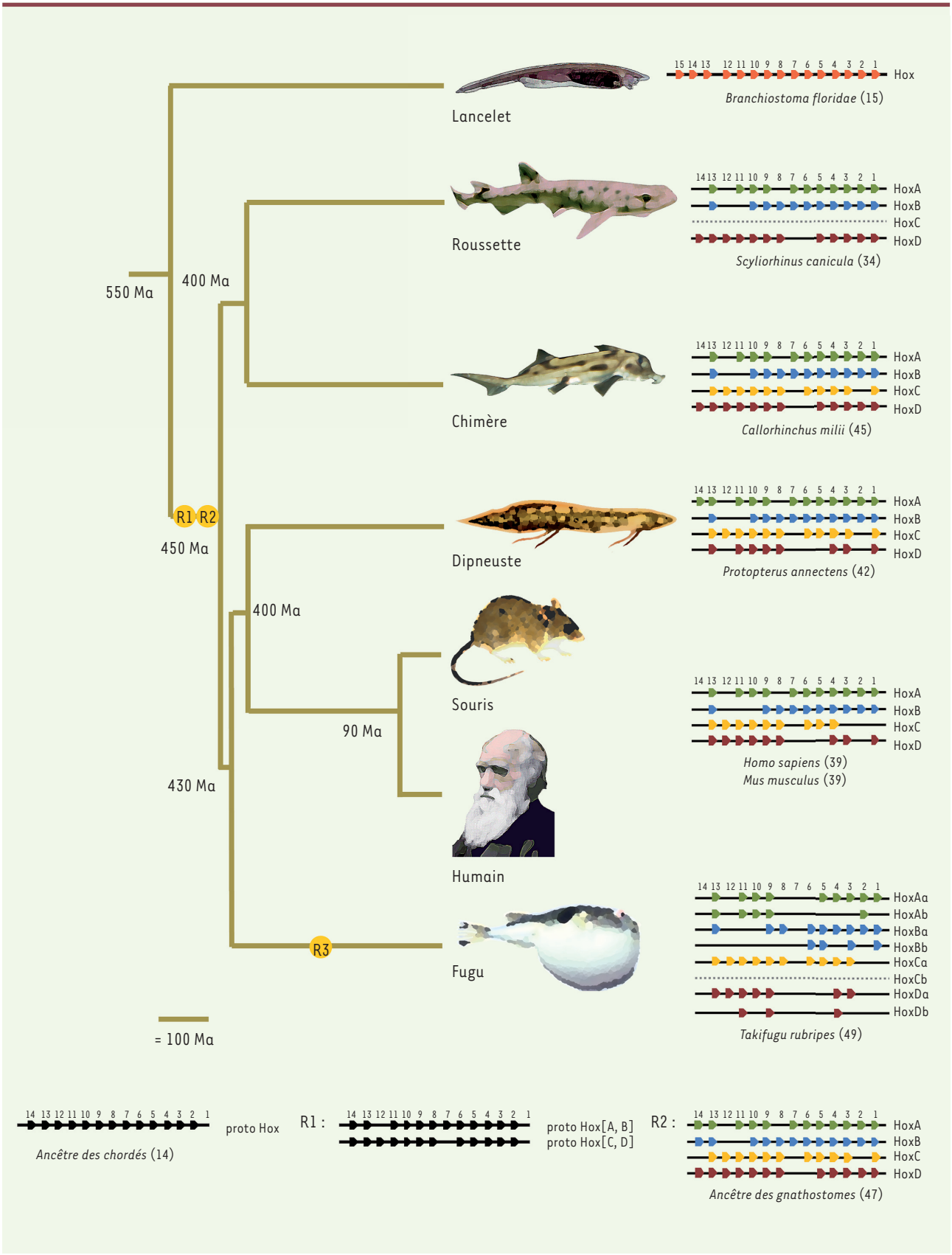


Figure 1. *Événements de tétraploïdisation et évolution des complexes Hox chez les vertébrés.*

s'interroge sur les raisons de cette retenue d'ENCODE. S'agit-il de donner un air plus scientifique à cette « découverte » ou de laisser un peu d'espace pour les concepts éculés de la biologie évolutive [11] ?

La matière noire du génome humain

Sachant que la biologie évolutive, en mesurant la pression de sélection sur les séquences, propose qu'il y a environ 10 % à 15 % d'ADN utile dans notre génome [3, 4], que contiennent les 85 % de matière noire aux supposées crypto-fonctions ? En fait, cet ADN n'est pas si noir que ça, car il est depuis longtemps établi qu'il est pour l'essentiel composé de fragments de transposons, plus ou moins anciens et donc parfois très difficiles à identifier [21], et de pseudogènes qui ne sont pas transcrits en ARN utiles au fonctionnement de l'organisme [22]. Il y a en outre une bonne dose de séquences de plus ou moins grandes tailles qui sont plus ou moins répétées, les microsatellites, minisatellites et autres satellites. L'origine de tous ces éléments est très bien identifiée.

• Les transposons ont la propriété de faire des copies d'eux-mêmes qui s'insèrent dans le génome, mais souvent seuls des fragments sont dupliqués et, pour la plupart, ils perdent ce faisant la capacité à se propager dans le génome. C'est un des exemples les mieux analysés de conflit génétique : les transposons maximisent leur transmission en proliférant dans les génomes tandis que le reste du génome lutte contre cette prolifération. La multiplication des copies de transposons se fait au détriment des autres gènes du fait de l'effet délétère de certaines de ces insertions. Les cycles de propagation et de contrôle des transposons sont très bien étudiés tant d'un point de vue théorique qu'expérimental, et, pour la plupart de ces transpositions, il n'a été détecté aucune utilité pour l'organisme [23, 24].

• La deuxième source massive d'ADN sans fonction est représentée par les pseudogènes qui se forment à la suite d'évènements de duplication de gènes. La formation de deux copies entraîne la présence de deux sources redondantes de la même fonction. Dans la plupart des très nombreux cas répertoriés, une des copies accumule des mutations qui la rendent non fonctionnelle, et ce sans perte de *fitness* au niveau de l'organisme [25].

• De plus, des séquences répétées naissent régulièrement, car la réplication présente une fâcheuse tendance : elle bégaie ! En présence d'une répétition de quelques nucléotides, l'ADN polymérase ajoute ou enlève fréquemment des répétitions. La recombinaison a également tendance à produire des séquences répétées en tandem.

Une grande quantité d'ADN sans fonction à l'échelle de l'organisme peut donc être produite et se maintenir longtemps, à la condition qu'il n'y ait pas de mécanisme efficace pour l'éliminer, c'est-à-dire des délétions de fragments d'ADN qui donnent un avantage sélectif suffisant aux individus porteurs pour qu'elles se fixent dans les populations. Sans gain de valeur sélective, la perte d'ADN inutile ne s'effectue donc que du seul fait du hasard, la dérive génétique. Donc, si cet ADN sans fonction est neutre ou presque neutre du point de vue de la sélection, il peut être stocké longtemps dans le génome. C'est l'origine de l'ADN poubelle (*junk DNA*), un ADN qui ne contient aucune information utile à la vie ni à la transmission du génome d'un individu. Il est vrai que des fragments de transposons, des pseudogènes et des séquences non codantes répétées

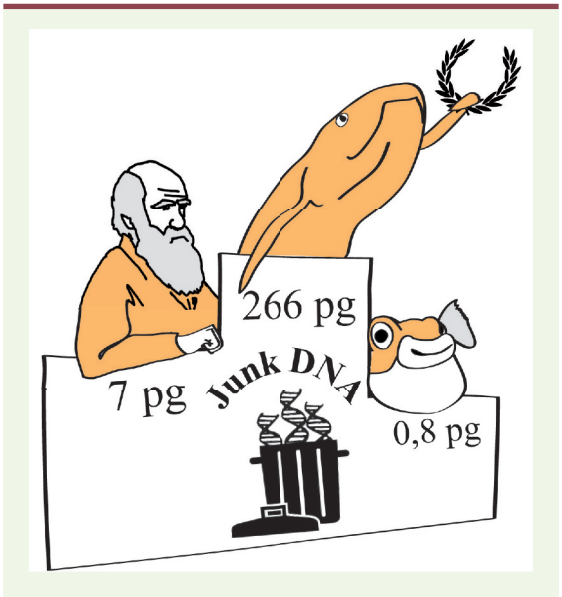


Figure 2. ***Classement de trois vertébrés en fonction de la taille de leur génome.***

en tandem acquièrent quelquefois des fonctions. On parle alors d'« exaptation » [26]. Mais ces observations confirment, et n'infirment en rien, l'existence de l'ADN poubelle. La grande quantité d'ADN poubelle repose sur la multitude de mécanismes qui le génèrent, le peu de puissance des mécanismes qui l'éliminent, et le fait que son recyclage fonctionnel est très rare. Cette conception de la dynamique de l'architecture des génomes n'est donc en rien invalidée par les résultats d'ENCODE. La seule nouveauté en la matière est la découverte récente de la possibilité de l'émergence au sein de l'ADN poubelle de nouveaux gènes codant des protéines [27].

Crypto-fonction de la matière noire : tout s'obscurcit !

La seule façon de maintenir l'hypothèse d'une fonction portant sur l'ensemble du génome est de proposer que cette fonction ne dépende pas de sa séquence, mais de la quantité totale d'ADN. Pour une raison absolument inconnue à ce jour, nos cellules devraient contenir environ 7 pg d'ADN. Une telle fonction de la « masse noire » s'oppose à l'ensemble des connaissances de la dynamique des génomes. Les deux mécanismes principaux qui permettent de faire varier fortement la quantité d'ADN d'un génome sont la polyploïdisation et les pics de transpositions [18]. Notre génome a ainsi vu son poids doubler deux fois à l'occasion de deux évènements de tétraploïdisation (R1, R2) qui ont eu lieu il y a environ 500 millions d'années. Il a ensuite perdu beaucoup de poids. On peut retracer cette histoire en étudiant la

structure des complexes de gènes Hox *(Figure 1)*. Ainsi, bien que chaque gène doive être présent sous la forme de quatre copies, une par complexe, il ne reste en général que deux ou trois copies, rarement les quatre copies de départ. Les requins ont perdu un complexe entier, le complexe HoxC [28, 29]. Chez les poissons actinoptérygiens, une tétraploïdisation supplémentaire (R3) est à l'origine de huit complexes de gènes. Après cette duplication des complexes de gènes, il y a eu de nombreuses pertes de gènes, voire de complexes complets. Finalement, chez tous les vertébrés à mâchoires, le nombre de gènes *Hox* varie dans une gamme assez étroite, à l'exception des espèces chez lesquelles une nouvelle duplication du génome s'est produite récemment, ne laissant pas le temps à la perte de gènes. Ainsi, on compte 105 gènes *Hox* chez le saumon [30]. Les génomes ont donc tendance à revenir à des tailles intermédiaires après les évènements d'amplification massive, mais on observe chez certaines espèces des génomes extraordinairement grands et, chez d'autres, des génomes extraordinairement petits. Parmi les vertébrés, le génome humain a une taille intermédiaire (7 pg/cellule diploïde). Le fugu possède un des plus petits génomes (0,8 pg) et un dipneuste africain possède un des plus grands génomes (266 pg) *(Figure 2)*. La découverte que la taille du génome du fugu est huit fois inférieure à celle du génome humain pour un nombre de gènes équivalent fut d'ailleurs saluée, il y a plus de vingt ans par Bertrand Jordan dans ces mêmes colonnes, comme un argument définitif en faveur de l'absence de fonction de l'ADN poubelle [31]. Aucune explication fonctionnelle n'a pu être proposée, à ce jour, pour expliquer ces différences de taille de génomes. Il est probable que la taille de ces génomes dépende essentiellement du *ratio* prolifération/élimination de l'ADN poubelle [32, 33].

Le dessin inintelligent de notre génome

Reconsidérons maintenant les observations d'ENCODE en prenant en compte que notre génome contient une multitude de séquences sur lesquelles des facteurs de transcription peuvent se fixer. Il n'est pas étonnant qu'il soit entièrement transcrit, au moins à faible niveau. Le très fort taux de *turnover* de ces sites de fixation et de ces transcrits au cours de l'évolution, ainsi que la difficulté à leur trouver quelque fonction biologique, montrent qu'ils ne sont très vraisemblablement pas indispensables [34]. La petite taille des motifs reconnus par les facteurs de transcription explique très simplement la forte probabilité d'apparition d'un tel site dans un quelconque fragment d'ADN [16]. La transcription semble être un processus biochimique qui peut démarrer assez facilement un peu n'importe où dans un génome [15]. Clairement, notre génome n'est pas optimisé, ni au niveau de son organisation, ni au niveau de son fonctionnement. Mais, c'est aussi vrai au niveau d'autres processus biochimiques et jusqu'à celui de la morpho-anatomie. Le mieux n'est pas nécessairement accessible, parce que tout simplement il n'apparaît pas (toutes les mutations possibles ne se réalisent pas, et loin de là), ou parce que lorsque la (ou les) mutation(s) nécessaire(s) se produise(nt) chez un individu, elle(s) ne peut (peuvent) pas se fixer dans la population, du simple fait du hasard ou à cause de forces sélectives antagonistes. L'ADN poubelle n'est donc probablement pas maintenu pour son utilité aujourd'hui, et encore moins, bien sûr, pour une éventuelle utilité dans un avenir indéterminé. Les organismes ne peuvent prévoir de quoi sera fait leur futur. En revanche, cet ADN constituerait une source d'adaptabilité, ou de façon plus générale une source d'« évolvabilité », qui se serait formée fortuitement, comme un produit secondaire de l'impuissance des mécanismes moléculaires à évacuer l'ADN sans fonction. Par bien des égards, il y a une forte similitude avec le maintien de la reproduction sexuée. Alors qu'il y a un avantage pour une femelle à éliminer la production de mâle et à ne produire que des femelles, chez les eucaryotes la reproduction sexuée est la norme et la reproduction asexuée l'exception [35]. La distribution phylogénétique des organismes asexués indique que ceux-ci apparaissent régulièrement, mais qu'ils ne se maintiennent pas sur le long terme, même si quelques rares contre-exemples existent [36, 37]. Des arguments théoriques et expérimentaux suggèrent que les organismes asexués ont une adaptabilité réduite du fait de la perte de la recombinaison qui permet la combinaison d'allèles favorables et l'élimination d'allèles délétères. La sélection entre lignées éliminerait les lignées asexuées, avantagées sur le court terme, mais désavantagées sur le long terme [35]. De façon analogue, un génome réduit à sa partie utile serait un avantage mineur à court terme et un désavantage à long terme, et la sélection de groupe entre lignées évolutives pourrait alors expliquer le maintien de l'ADN poubelle dans la plupart des génomes. Cette hypothèse peut être testée en étudiant l'âge des génomes exceptionnellement petits chez les plantes et les animaux. Ils devraient être pour la plupart relativement récents, à l'échelle de l'évolution de ces taxons.

Plaidoyer en faveur de la « petite » science

Rien en biologie n'a de sens, si ce n'est à la lumière de l'évolution. Cet aphorisme de Theodosius Dobzhansky est plus que jamais d'actualité. À l'heure des grands projets et lorsque les exploits technologiques et la quantité de données produites servent de mesure de la qualité, il est devenu difficile de défendre les approches reposant sur l'élaboration d'hypothèses, leurs développements formels et leurs tests expérimentaux. Force est de constater que des énormités sont régulièrement proférées pour justifier l'argent investi dans les projets « BIG » science. Seule une bonne connaissance des acquis théoriques et expérimentaux en biologie évolutive pourrait limiter ces errements répétés d'une techno-science irréfléchie, notamment les raisonnements panglossiens. Malheureusement, les universités françaises ne recrutent que peu (voire pas du tout !) d'enseignants-chercheurs ayant une solide formation en biologie évolutive. De plus, la

principale agence de financement de la recherche en France n'a plus de programme spécifique pour soutenir les chercheurs travaillant dans ce domaine. Pourrons-nous encore longtemps contenir les errements des Dr Pangloss et, en bon Candide, cultiver notre jardin ? ◊

SUMMARY

ENCODE apophenia or a panglossian analysis of the human genome

In September 2012, a batch of more than 30 articles presenting the results of the ENCODE (Encyclopaedia of DNA Elements) project was released. Many of these articles appeared in *Nature* and *Science*, the two most prestigious interdisciplinary scientific journals. Since that time, hundreds of other articles dedicated to the further analyses of the Encode data have been published. The time of hundreds of scientists and hundreds of millions of dollars were not invested in vain since this project had led to an apparent paradigm shift: contrary to the classical view, 80% of the human genome is not junk DNA, but is functional. This hypothesis has been criticized by evolutionary biologists, sometimes eagerly, and detailed refutations have been published in specialized journals with impact factors far below those that published the main contribution of the Encode project to our understanding of genome architecture. In 2014, the Encode consortium released a new batch of articles that neither suggested that 80% of the genome is functional nor commented on the disappearance of their 2012 scientific breakthrough. Unfortunately, by that time many biologists had accepted the idea that 80% of the genome is functional, or at least, that this idea is a valid alternative to the long held evolutionary genetic view that it is not. In order to understand the dynamics of the genome, it is necessary to re-examine the basics of evolutionary genetics because, not only are they well established, they also will allow us to avoid the pitfall of a panglossian interpretation of Encode. Actually, the architecture of the genome and its dynamics are the product of trade-offs between various evolutionary forces, and many structural features are not related to functional properties. In other words, evolution does not produce the best of all worlds, not even the best of all possible worlds, but only *one* possible world. ◊

REMERCIEMENTS

Nous tenons à exprimer ici toute notre gratitude à nos collègues Mélanie Debiais-Thibaud et Alice Michel-Salzat pour leur relecture attentive et critique de notre manuscrit, ainsi qu'à notre collègue Cushla Metcalfe, pour l'amélioration des titres et du résumé en anglais.

LIENS D'INTÉRÊT

Les auteurs déclarent n'avoir aucun lien d'intérêt concernant les données publiées dans cet article.

RÉFÉRENCES

1. Pennisi E. ENCODE project writes eulogy for junk DNA. *Science* 2012 ; 337 : 1159-61.
2. Dunham I, Kundaje A, Aldred SF, *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012 ; 489 : 57-74.
3. Rands CM, Meader S, Ponting CP, Lunter G. 8.2% of the Human genome is constrained: variation in rates of turnover across functional element classes in the Human lineage. *PLoS Genet* 2014 ; 10 : e1004525.
4. Ponting CP, Hardison RC. What fraction of the human genome is functional? *Genome Res* 2011 ; 21 : 1769-76.
5. Ecker JR. Forum: Genomics ENCODE explained. *Nature* 2012 ; 489 : 52-3.
6. Alberts B. The End of small science? *Science* 2012 ; 337 : 1583.
7. Doolittle WF. Is junk DNA bunk? A critique of ENCODE. *Proc Natl Acad Sci USA* 2013 ; 110 : 5294-300.
8. Eddy SR. The ENCODE project: missteps overshadowing a success. *Curr Biol* 2013 ; 23 : R259-61.
9. Niu DK, Jiang L. Can ENCODE tell us how much junk DNA we carry in our genome? *Biochem Biophys Res Commun* 2013 ; 430 : 1340-3.
10. Palazzo AF, Gregory TR. The case for junk DNA. *PLoS Genet* 2014 ; 10 : e1004351.
11. Graur D, Zheng YC, Price N, *et al.* On the immortality of television sets: function in the Human genome according to the evolution-free gospel of ENCODE. *Genome Biol Evol* 2013 ; 5 : 578-90.
12. Muerdter F, Stark A. Genomics: hiding in plain sight. *Nature* 2014 ; 512 : 374-5.
13. Gerstein MB, Rozowsky J, Yan KK, *et al.* Comparative analysis of the transcriptome across distant species. *Nature* 2014 ; 512 : 445-8.
14. Doolittle WF, Brunet TDP, Linquist S, Gregory TR. Distinguishing between function and effect in genome biology. *Genome Biol Evol* 2014 ; 6 : 1234-7.
15. Struhl K. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat Struct Mol Biol* 2007 ; 14 : 103-5.
16. Ruths T, Nakhleh L. ncDNA and drift drive binding site accumulation. *BMC Evol Biol* 2012 ; 12 : 159.
17. Gould SJ, Lewontin RC. The spandrels of San Marco and the panglossian paradigm: a critique of the adaptationist programme. *Proc R Soc Lond B* 1979 ; 205 : 581-98.
18. Lynch M. *The origins of genome architecture*. Sunderland, Massachusetts : Sinauer, 2007.
19. Kimura M. *The neutral theory of molecular evolution.* New York : Cambridge University Press, 1983.
20. Rice WR. Nothing in genetics makes sense except in light of genomic conflict. *Annu Rev Ecol Evol Syst* 2013 ; 44 : 217-37.
21. De Koning APJ, Gu WJ, Castoe TA, *et al.* Repetitive elements may comprise over two-thirds of the Human genome. *Plos Genet* 2011 ; 7 : e1002384.
22. Zhang ZL, Harrison PM, Liu Y, Gerstein M. Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. *Genome Res* 2003 ; 13 : 2541-58.
23. Jacobs FMJ, Greenberg D, Nguyen N, *et al.* An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. *Nature* 2014 ; 516 : 242-5.
24. Hua-Van A, Le Rouzic A, Boutin TS, *et al.* The struggle for life of the genome's selfish architects. *Biol Direct* 2011 ; 6 : 19.
25. Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. *Science* 2000 ; 290 : 1151-5.
26. De Souza FSJ, Franchini LF, Rubinstein M. Exaptation of transposable elements into novel cis-regulatory elements: is the evidence always strong? *Mol Biol Evol* 2013 ; 30 : 1239-51.
27. Casane D, Laurenti P. Syllogomanie moléculaire : l'ADN non codant enrichit le jeu des possibles. *Med Sci (Paris)* 2014 ; 30 : 1177-83.
28. Oulion S, Debiais-Thibaud M, d'Aubenton-Carafa Y, *et al.* Evolution of Hox gene clusters in gnathostomes: insights from a survey of a shark (*Scyliorhinus canicula*) transcriptome. *Mol Biol Evol* 2010 ; 27 : 2829-38.
29. Oulion S, Laurenti P, Casane D. Organisation des gènes *Hox* : l'étude de vertébrés non-modèles mène à un nouveau paradigme. *Med Sci (Paris)* 2012 ; 28 : 350-3.
30. Pascual-Anaya J, D'Aniello S, Kuratani S, Garcia-Fernandez J. Evolution of *Hox* gene clusters in deuterostomes. *BMC Dev Biol* 2013 ; 13 : 26.
31. Jordan B. Fugu story. *Med Sci (Paris)* 1994 ; 10 : 1154-6.
32. Metcalfe CJ, Casane D. Accommodating the load: the transposable element content of very large genomes. *Mob Genet Elements* 2013 ; 3 : e24775.
33. Metcalfe CJ, Filee J, Germon I, *et al.* Evolution of the Australian lungfish (*Neoceratodus forsteri*) genome: a major role for CR1 and L2 LINE elements. *Mol Biol Evol* 2012 ; 29 : 3529-39.
34. Kapusta A, Feschotte C. Volatile evolution of long noncoding RNA repertoires: mechanisms and biological implications. *Trends Genet* 2014 ; 30 : 439-52.
35. De Vienne DM, Giraud T, Gouyon PH. Lineage selection and the maintenance of sex. *PLoS One* 2013 ; 8 : e66906.
36. Boschetti C, Carr A, Crisp A, *et al.* Biochemical diversification through foreign gene expression in bdelloid rotifers. *PLoS Genet* 2012 ; 8 : e1003035.
37. Flot JF, Hespeels B, Li X, *et al.* Genomic evidence for ameiotic evolution in the bdelloid rotifer *Adineta vaga*. *Nature* 2013 ; 500 : 453-7.

TIRÉS À PART

P. Laurenti

B.3 Evidence of Late Pleistocene origin of *Astyanax mexicanus* cavefish

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 x 10,000 x 2 x 10^{-2} = 400 new SNPs at each generation, and that these 400 new SNPs appear with an initial frequency of 1 / (2 x 10,000) = 5 x 10^{-5}. In parallel, 25 new SNPs appear with initial frequency of 8 x 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 5×10^{-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 $t = (\delta\mu)^2 / \mu \times [(g_{CF} \times g_{SF})/(g_{CF} + g_{SF})]$, 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.

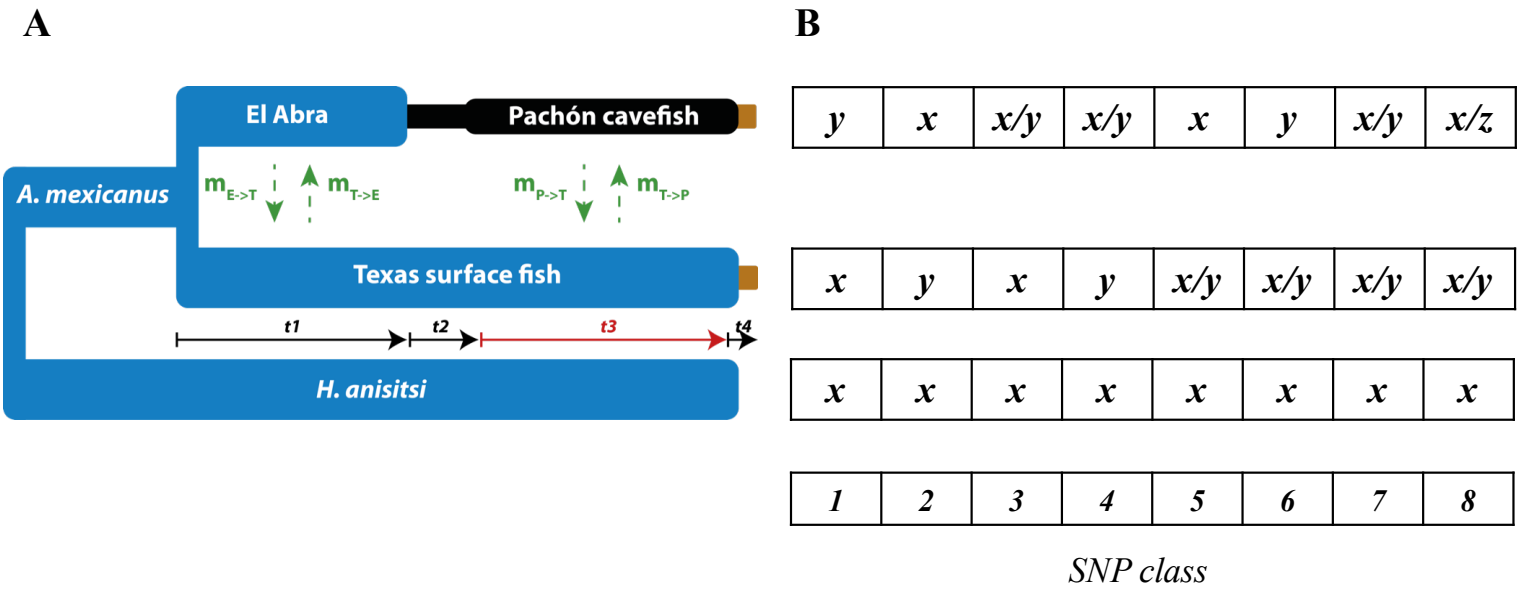

A
B
El Abra
Pachón cavefish
A. mexicanus
m E->T
m T->E
m P->T
m T->P
Texas surface fish
t1
t2
t3
t4
H. anisitsi
y x x/y x/y x y x/y x/z
x y x y x/y x/y x/y x/y
x x x x x x x x
1 2 3 4 5 6 7 8
SNP class

Figure 1

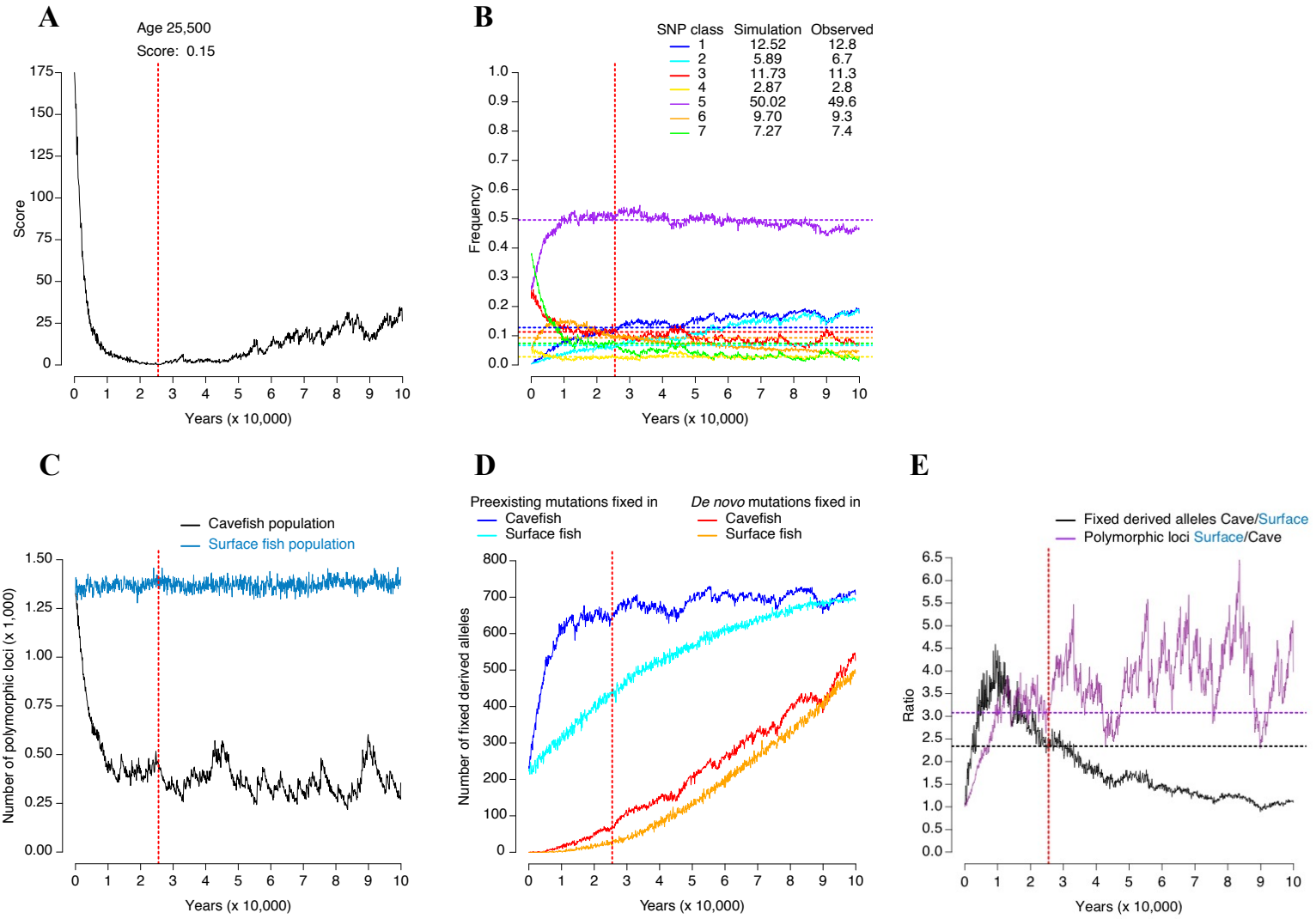

A
Age 25,500
Score: 0.15
Score
Years (x 10,000)
B
SNP class Simulation Observed
1 12.52 12.8
2 5.89 6.7
3 11.73 11.3
4 2.87 2.8
5 50.02 49.6
6 9.70 9.3
7 7.27 7.4
Frequency
Years (x 10,000)
C
Cavefish population
Surface fish population
Number of polymorphic loci (x 1,000)
Years (x 10,000)
D
Preexisting mutations fixed in
Cavefish
Surface fish
De novo mutations fixed in
Cavefish
Surface fish
Number of fixed derived alleles
Years (x 10,000)
E
Fixed derived alleles Cave/Surface
Polymorphic loci Surface/Cave
Ratio
Years (x 10,000)

Figure 2

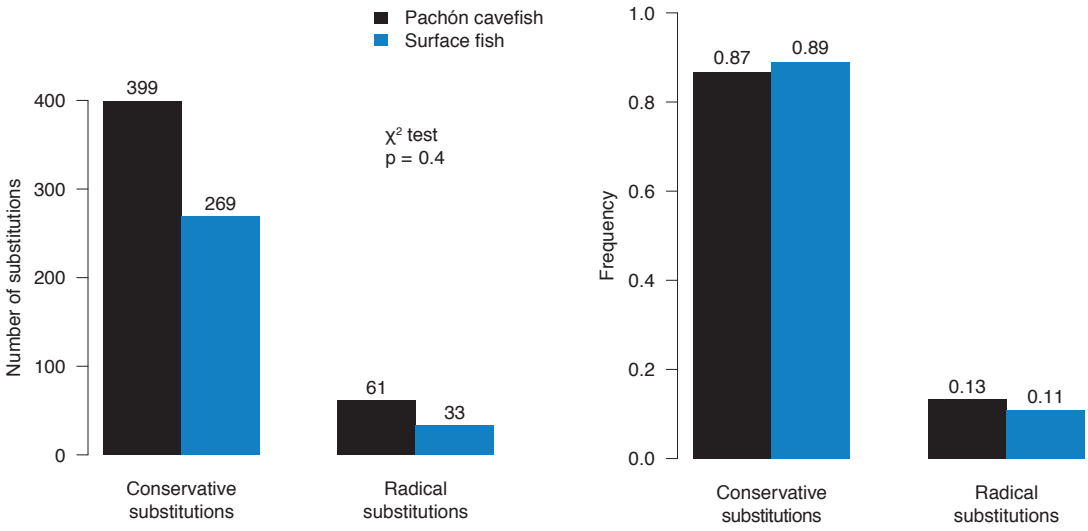

Pachón cavefish
Surface fish
Number of substitutions
400
300
200
100
0
399
269
61
33
χ² test
p = 0.4
Conservative substitutions
Radical substitutions
Frequency
1.0
0.8
0.6
0.4
0.2
0.0
0.87
0.89
0.13
0.11
Conservative substitutions
Radical substitutions

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.

Tempo et mode de l'évolution des populations cavernicoles de l'espèce *Astyanax mexicanus*

Le poisson *Astyanax mexicanus* est un modèle particulièrement intéressant pour l'étude de l'évolution. En effet, dans cette espèce de poissons d'eau douce, il existe des populations vivant de façon pérenne dans des grottes. Dans cet environnement, l'obscurité est totale et permanente et les ressources en nourriture souvent faibles. Les poissons cavernicoles se sont adaptés à la vie souterraine et ils présentent de nombreuses modifications phénotypiques comme la dépigmentation, la perte des yeux, l'augmentation du nombre et de la taille d'organes sensoriels non-visuels et plusieurs changements du comportement. Un des problèmes majeurs est de savoir si ces modifications phénotypiques sont dues à des mutations préexistantes à la colonisation de l'environnement cavernicole ou si elles sont apparues après. Pour répondre à cette question, connaître l'âge des populations est un facteur important car dans une population récente, il n'y aura probablement pas eu suffisamment de temps pour l'apparition de beaucoup de mutations et leur fixation. L'objet de cette thèse est donc l'estimation de l'âge d'une population, celle de la grotte Pachón qui est souvent considérée comme étant une des plus anciennes et une des plus isolées. Au cours de ces travaux de thèse, nous avons développé une nouvelle méthode de datation qui repose d'une part sur la caractérisation du polymorphisme nucléotidique à l'intérieur de chaque population et entre populations, et d'autre part la comparaison de ces données avec des simulations de l'évolution du polymorphisme. Les résultats obtenus, ainsi que la réanalyse de données sur le polymorphisme d'haplotypes mitochondriaux et de loci microsatellites précédemment publiées, suggèrent que les populations cavernicoles seraient bien plus récentes qu'habituellement indiqué dans la littérature (quelques milliers d'années, et non plusieurs centaines de milliers d'années). Les conséquences d'un tempo rapide d'évolution sur le mode d'évolution de ces poissons cavernicoles ont aussi été présentées.

Mots- clés : poissons cavernicoles, séquençage à haut-débit, transcriptomique comparé, modélisation, datation moléculaire, adaptation

Tempo and mode of the *Astyanax mexicanus* cavefish evolution

The fish *Astyanax mexicanus* is a particularly suitable model for evolutionary biology studies. Indeed, in this species there are several subterranean populations which live in the total and permanent darkness of cave. These cavefish are well adapted to the life in this inhospitable environment and they show several differences with their surface conspecific such as depigmentation, eye loss and behavioral changes. A major unresolved issue is about the relative role of surface fish standing genetic variation and de novo mutations appeared in cavefish populations after their settlement in caves in their phenotypic evolution. In order to examine this issue, accurate estimations of population ages are very important because many new mutations cannot appear and fix in a recent population. In this thesis we aimed to estimate the age of the Pachón cave population which is considered as one of the oldest and most isolated populations. We developed a new method which is based on measures of the distribution of single nucleotide polymorphism within each population and between populations. Our results, as well as reanalyses of published data about mitochondrial haplotypes and microsatellite loci polymorphism suggest that cavefish populations are much more recent than previously thought (several thousand years and not several hundred thousand years). The consequences of a fast tempo of evolution on the mode of evolution of cavefish is also discussed.

Keyword : cavefish, high-throughput sequencing, comparative transcriptomic, modelisation, molecular datation, adaptation.

Université Paris Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France