Évolution de gènes de la vision et de l'olfaction chez les poissons à nageoires rayonnées

Maxime Policarpo

▶ **To cite this version:**

Maxime Policarpo. Évolution de gènes de la vision et de l'olfaction chez les poissons à nageoires rayonnées. Génétique animale. Université Paris-Saclay, 2021. Français. NNT : 2021UPASL117 . tel-03563536

HAL Id: tel-03563536

https://theses.hal.science/tel-03563536v1

Submitted on 9 Feb 2022

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

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

THESE DE DOCTORAT

NNT : 2021UPASL117

université PARIS-SACLAY

Évolution de gènes de la vision et de l'olfaction chez les poissons à nageoires rayonnées

Evolution of vision and olfaction genes in ray-finned fishes

Thèse de doctorat de l'université Paris-Saclay

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

Spécialité de doctorat : Sciences de la vie et de la santé

Graduate School : Sciences de la vie et santé. Référent : Faculté des sciences d'Orsay

Thèse préparée dans les unités de recherche **EGCE** (Université Paris-Saclay, CNRS, IRD) et **Institut des Neurosciences Paris-Saclay** (Université Paris-Saclay, CNRS) sous la direction de **Didier CASANE**, professeur, et la co-direction de **Sylvie RETAUX**, directrice de recherche

Thèse soutenue à Paris-Saclay, le 20 Décembre 2021, par

Maxime POLICARPO

Composition du Jury

Hélène MORLON Directrice de recherche, IBENS	Présidente & Rapporteur
Christophe DOUADY Professeur, Université Claude Bernard-Lyon-1	Rapporteur & Examinateur
Aurélie HUA-VAN Maitresse de conférences, Université Paris-Saclay	Examinatrice
Filippo DEL BENE Directeur de recherche, INSERM	Examinateur
Didier CASANE Professeur, Université de Paris	Directeur de thèse

Titre : Évolution de gènes de la vision et de l'olfaction chez les poissons à nageoires rayonnées

Mots clés : Vision, olfaction, génomique, évolution, adaptation, poissons

Résumé : Avec environ 35 000 espèces décrites, les poissons à nageoires rayonnées, ou actinoptérygiens, représentent la moitié des espèces de vertébrés et plus de 95% des vertébrés aquatiques. Ces animaux ont colonisé un grand nombre de milieux différents, que ce soit en eaux douces ou salées et à des profondeurs allant jusqu'à 8000 mètres. Cette grande hétérogénéité écologique au sein d'un groupe d'espèces rend particulièrement intéressant l'étude de l'évolution de leurs systèmes sensoriels qui sont impliqués dans un grand nombre de fonctions essentielles comme le déplacement, l'alimentation et la reproduction. Dans cette thèse, des études comparatives sont menées afin de mieux comprendre l'évolution de l'olfaction chez les actinoptérygiens et de la vision chez des actinoptérygiens cavernicoles.La détection de molécules odorantes est assurée par les récepteurs olfactifs, protéines présentes à la surface de la membrane des neurones sensoriels insérés dans l'épithélium olfactif. Cet épithélium peut avoir une organisation plus ou moins complexe en fonction de l'espèce, notamment par la présence d'un nombre différent de lamelles, impliqués dans la dimension de la surface d'échange avec le milieu extérieur. En analysant le génome de 185 actinoptérygiens, nous avons observé une grande variabilité du nombre de gènes codant les récepteurs olfactifs et une forte corrélation entre la taille de ce répertoire de gènes et le nombre de lamelles, probablement parce qu'une grande surface d'épithélium olfactif est nécessaire pour qu'un grand nombre de récepteurs puisse être présent. Ainsi, alors que les polyptères (*Erpetoichthys calabaricus* et *Polypterus senegalus*) peuvent produire plus d'un millier de récepteurs différents et que leur organe olfactif est constitué de centaines de lamelles, le poisson lune (*Mola mola*) ne présente qu'un repli médian au niveau d'un épithélium olfactif réduit et ne peut produire qu'une trentaine de récepteurs différents. Les données au niveau morphologique et moléculaire montrent que les capacités olfactives sont très différentes entre actinoptérygiens. Par ailleurs, nous avons étudié l'évolution de gènes impliqués dans la photoréception chez des poissons cavernicoles qui ont des yeux plus ou moins dégénérés. En recherchant la présence de mutations perte de fonction (mutations qui rendent un gène non fonctionnel, ou pseudogène), nous avons montré la présence de pseudogènes plus ou moins nombreux en fonction de l'espèce et nous en avons déduit qu'ils sont dispensables chez ces poissons. Les gènes impliqués dans le cycle circadien et la pigmentation sont pour la plupart sous sélection purificatrice, probablement à cause de leur pléiotropie (gènes impliqués dans de multiples processus biologiques). Seulement quelques gènes sont pseudogénisés de façon récurrente chez des poissons cavernicoles, suggérant qu'ils sont dispensables dans l'obscurité des grottes et seraient souvent impliqués dans la dépigmentation et la perte du cycle circadien dans les grottes.En utilisant différentes méthodes qui reposent sur l'analyse du changement de pression de sélection sur les gènes de la photoréception au moment du passage de la vie en surface à la vie dans les grottes, nous avons estimé que les poissons cavernicoles étudiés sont relativement récents, c'est-à-dire qu'ils ont colonisé des grottes entre le début du Pliocène (~5 millions d'années) et la fin du Pléistocène (~10 000 ans).

Title : Evolution of vision and olfaction genes in ray-finned fishes

Keywords : Vision, olfaction, genomic, evolution, adaptation, fishes

Abstract : With about 35,000 described species, ray-finned fishes, also called actinopterygians, represent half of all vertebrate species and more than 95% of all aquatic vertebrates. These animals have colonized a large number of different environments, both in fresh and salt water and at depths of up to 8000 meters. These highly heterogeneous ecological niches within a species group makes it particularly interesting to study the evolution of their sensory systems, which are involved in a number of essential functions such as movement, feeding and reproduction. In this thesis, comparative studies are conducted to better understand the evolution of olfaction in actinopterygians and vision in cave-dwelling actinopterygians.

The detection of odorant molecules is ensured by olfactory receptors, proteins present on the surface of the membrane of sensory neurons inserted into the olfactory epithelium. This epithelium can have a more or less complex organization depending on the species, in particular by the presence of a different number of lamellae, regulating the size of the exchange surface with the external environment. By analyzing the genome of 185 actinopterygians, we observed a great variability in the number of genes encoding olfactory receptors and a strong correlation between the size of this gene repertoire and the number of lamellae, probably because a large surface area of olfactory epithelium is necessary for a large number of receptors to be present.

Thus, whereas polypterids (*Erpetoichthys calabaricus* and *Polypterus senegalus*) can produce more than a thousand different receptors with an olfactory organ that consists of hundreds of lamellae, the sunfish (*Mola mola*) has only a median fold in a reduced olfactory epithelium and can produce only about thirty different receptors. Morphological and molecular data show that the olfactory capacities are very different between actinopterygians. In addition, we have studied the evolution of genes involved in photoreception in cave-dwelling fish with more or less degenerate eyes. By looking for the presence of loss-of-function mutations (mutations that makes a gene non-functional, a pseudogene), we have shown the presence of more or less numerous pseudogenes depending on the species, suggesting that photoreception genes are dispensable in these fishes. Most genes involved in the circadian cycle and pigmentation are under purifying selection, probably because of their pleiotropy (genes involved in multiple biological processes). Only a few genes are recurrently pseudogenized in cavefishes, suggesting that they are dispensable in the darkness of caves and could often be involved in depigmentation and loss of circadian clock in caves.Using different methods relying on the analysis of selective shift on photoreception genes at the time of transition from surface to cave, we estimated that the cavefishes we studied are relatively recent, i.e. they colonized caves between the early Pliocene (~5 million years ago) and the late Pleistocene (~10,000 years ago).

« *Si les animaux n'existaient pas, ne serions-nous pas encore plus incompréhensibles à nous-mêmes ?* »

Georges-Louis Leclerc de Buffon (1707 – 1788)
(Discours sur la nature des animaux, 1753)

Remerciements

Je tiens tout d'abord à remercier les rapporteurs et examinateurs de ma thèse d'avoir accepté d'évaluer mon travail.

Merci à Didier de m'avoir accueilli en stage de M2 et de m'avoir encouragé à passer le concours de l'école doctorale pour faire cette thèse. Merci pour tous les conseils que tu m'as donnés pendant ma thèse, que je m'efforcerai de retenir pendant le reste de mon parcours. Partager ce bureau au fond du couloir avec toi a été un grand plaisir. Je me souviendrai de toutes ces discussions autour du tableau blanc et de tes graphiques à main levée, de toutes ces séances Zoom, de ton engouement lorsque je te montrais quelques résultats, de ta gentillesse et de ton humour. Si je devais refaire une thèse, je la referais avec toi, sans hésitation.

Merci à Sylvie d'avoir accepté de co-encadrer ma thèse et d'avoir été un soutien majeur pendant ces trois dernières années. J'ai été et continue d'être impressionné par l'éventail de connaissances et de compétences que tu possèdes. J'aimerais pouvoir un jour diriger une équipe de recherche comme tu le fais, alliant bienveillance et rigueur scientifique.

Merci à Didier et Sylvie de m'avoir donné la chance de faire partie d'une expédition au Mexique. Merci à toutes celles et tous ceux qui ont participé au voyage. Un grand merci à Maryline pour toutes ces discussions, et de ta perpétuelle envie d'aider les autres. Merci pour ton humour et ton rire que je n'oublierai pas de sitôt. Merci à Boudjema. J'ai eu la chance de te découvrir et de forger une amitié avec toi pendant ce voyage. Merci à Constance, à Julien et Julien pour votre gentillesse. Merci à Laurent et Stéphane, sans qui j'aurais sûrement été beaucoup moins rassuré dans les grottes. Merci à Jorge qui est un vrai bosseur et un modèle. Merci à François pour ton humour décalé et ta capacité d'autodérision. Merci à Carole avec qui j'ai beaucoup apprécié discuter pendant le peu de temps qui nous était accordé. Un grand merci à toute l'équipe DECA pour vos magnifiques présentations du vendredi et les discussions qui ont suivi.

Merci à tous les membres de l'EGCE avec qui j'ai eu la chance de partager mon quotidien et qui font de ce laboratoire un lieu unique.
Merci pour les nombreuses présentations à la bibliothèque, qui m'ont fait découvrir des sujets de recherche divers et qui m'ont permis d'avoir des regards extérieurs sur mes travaux. Merci à tous les doctorants et stagiaires avec qui j'ai partagé de bons moments autour d'une bière entre autres. Merci à Véronique et Alice qui m'ont fait confiance et qui m'ont permis d'être chargé de TD et TP. Merci à toute l'équipe enseignante de Paris Université pour leur accueil. Merci à Isabelle de m'avoir accompagné pendant mes premières PCR et de s'être toujours intéressé à ce que je faisais et à ce que je vivais. Merci à Émilie, David et Arnaud avec qui j'ai partagé de nombreuses pauses café/clope. Merci à Jean-Luc pour son humour, sa sincérité et son engagement écologique. Merci à Pierre de m'avoir soutenu dans ma candidature auprès de l'école doctorale. Merci à Laure pour sa gentillesse et sa bienveillance envers tous les membres du laboratoire. Merci à Hélène, Sylvie, Magalie et Valérie, anciennes et nouvelles secrétaires du laboratoire qui m'ont toujours aidé dans les démarches administratives.

Merci à Éric, Marc et Michel d'avoir accepté de faire partie de mon comité de thèse pendant ces 3 ans. Merci à nos collaborateurs pour leur aide dans nos différentes études, notamment l'INRAE de Toulouse, l'I2BC, l'IGLF et la Smithsonian Institution. Merci à Patrick pour tous les poissons que tu nous as dessiné.

Merci aux femmes et hommes de ménage pour leur travail et sans lesquels mon bureau serait dans un piteux état.

Merci à Sarah que j'ai connue pendant son stage et avec qui j'ai maintenant la chance de partager ma vie. Merci de m'avoir toujours soutenu et de m'avoir donné confiance en moi quand je ne me sentais pas au niveau.
Un grand merci à ma famille qui m'a toujours encouragé dans mes études. Merci à tous mes amis, qui égayent ma vie et qui ont toujours été à l'écoute. Merci pour toutes ces soirées passées à discuter, à jouer, et parfois même à tuer des zombies.

Merci enfin à tous mes proches, à Didier et à Sylvie pour avoir pris le temps de relire et corriger ma thèse.

Table des matières

I) Introduction générale 13

A- Les actinoptérygiens : un clade majeur de vertébrés 14

1) Diversité et abondance d'éléments transposables 18

2) Duplications entières du génome 19

B- Évolution moléculaire par gains et pertes de gènes. 20

1) Gains de gènes 20

a) Créations de-novo 21

b) Transferts horizontaux de gènes 22

c) Duplications de gènes 22

2) Pertes de gènes 23

C- Modalités sensorielles chez les actinoptérygiens 30

1) L'électroperception 30

2) La magnétoréception 32

3) La mécanoréception 34

4) La photoréception et la chémoréception 36

II) La chémoréception chez les vertébrés 36

A- La gustation 36

B- L'olfaction 38

1) Rôle de l'olfaction 38

2) Organisation du système olfactif chez les actinoptérygiens 38

3) Reconnaissance des molécules odorantes par les récepteurs olfactifs 40

a) Spécificité des récepteurs olfactifs 40

b) Variabilité du nombre de récepteurs olfactifs 41

c) Mécanismes d'expansion et de contraction du répertoire de récepteurs olfactifs 44

III) Résultats - Évolution des quatre familles de récepteurs olfactifs chez les actinoptérygiens 47

A- Génomes utilisés et méthode de détection des gènes 47

1) Récupération et sélection de génomes depuis les bases de données publiques 47

2) Méthodes de récupération des séquences de récepteurs olfactifs 51

3) Évaluation des méthodes utilisées 53

B- Nombre de récepteurs olfactifs chez les actinoptérygiens 56

1) *OR* 61

2) *TAAR* et *OlfC* 64

3) *ORA* 69

4) Les récepteurs olfactifs : une coévolution de trois familles de gènes 72

C- Histoire évolutive des récepteurs olfactifs 74

1) Méthode d'inférence de gains et pertes de gènes 74

2) Taux de gains et de pertes de gènes dans l'arbre des espèces 78

D- Influence de facteurs environnementaux et morphologiques sur le répertoire de gènes olfactifs 81

E- Évolution de l'épithélium olfactif des actinoptérygiens 83

1) Récupération de données dans la littérature 83

2) Reconstruction de la forme ancestrale de l'épithélium olfactif 85

3) Corrélation de la richesse du répertoire olfactif avec la complexité morphologique de l'épithélium olfactif 88

F- Discussion 91

1) Dynamique évolutive des récepteurs olfactifs 91

2) Coévolution moléculaire et morphologique du système olfactif 93

3) Caractéristiques écologiques liées à la taille du répertoire olfactif 96

IV) La photoréception : modalité sensorielle obsolète dans les grottes 97

A- La photoréception visuelle 97

1) L'œil des vertébrés 97

2) La pigmentation : déjouer et attirer l'attention 99

3) Gènes impliqués dans la photoréception visuelle 102

B- La photoréception non-visuelle 104

C- Les organismes cavernicoles 105

1) Définition d'une grotte et de sa biodiversité 106

2) Adaptations morphologiques à la vie dans les grottes 107

3) Datation des organismes cavernicoles 110

a) Les méthodes de datation moléculaire 110

b) La datation moléculaire de la colonisation des grottes 112

4) Analyses moléculaires chez les organismes cavernicoles 113

V) Résultats : Étude des gènes de la photoréception, de la pigmentation et du cycle circadien chez les poissons cavernicoles 116

A- Hypothèses de travail 116

B- Génomes de poissons cavernicoles et de surface proches disponibles 117

1) *Astyanax mexicanus* 117
2) *Sinocyclocheilus* spp. 119
3) *Typhlichthys subterraneus* 119
4) *Lamprologus lethops* : un poisson aveugle non cavernicole 121
5) Séquençage des génomes de *Lucifuga dentata* et *Lucifuga gibarensis* 121

C- Construction de trois groupes de gènes 123

1) Gènes de la photoréception 123
2) Gènes du cycle circadien et gènes de la pigmentation 124

D- Méthodes et résultats de la recherche de pseudogènes 124

1) Nombre de mutations perte de fonction et nombre de pseudogènes 124
   a) Gènes incomplets 126
   b) Photoréception 129
   c) Cycle circadien 135
   d) Pigmentation 136
2) Nombre de mutations perte de fonction et leur distribution 139
   a) Distribution de la taille des indels 139
   b) Fréquence théorique de chaque type de mutation perte de fonction 140
   c) Position relative des mutations perte de fonction le long des séquences 143

E- Gènes mutés dans les trois groupes 145

F- Estimation du nombre de gènes neutres à partir du nombre de mutations perte de fonction et du nombre de pseudogènes 150

1) Répartition des mutations perte de fonction 150
2) Comparaison des pseudogènes de *L. dentata* et *T. subterraneus*. 154

G- Estimation de la neutralité à l'aide du dN/dS 156

H- Estimation du régime d'évolution des groupes de gènes à partir de l'effet délétère des mutations non synonymes 163

1) Gènes de la photoréception 165
2) Gènes du cycle circadien et de la pigmentation 169

I- Datations de la colonisation des grottes par différentes méthodes 170

1) Datation à partir du nombre de pseudogènes 170

2) Datation à partir du nombre de substitutions 174

3) Datation à partir des estimations de dN/dS 177

4) Datation à partir des effets délétères des mutations non-synonymes 178

5) Comparaison des différentes méthodes de datation 180

6) Comparaison avec les mammifères souterrains 181

J- Discussion 183

VI) Étude et échantillonnage de populations sauvages d'Astyanax mexicanus 186

A- Étude des capacités olfactives 186

B- Échantillonnage d'ADN 188

VII) Conclusion générale 192

Table des figures 193

Crédits photographiques et graphiques 201

Boites 203

Références 204

A- Articles publiés en premier auteur 237

1) Contrasting Gene Decay in Subterranean Vertebrates: Insights from Cavefishes and Fossorial Mammals 237

2) Evolutionary Dynamics of the OR Gene Repertoire in Teleost Fishes: Evidence of an Association with Changes in Olfactory Epithelium Shape 255

B- Articles publiés en coauteur 268

1) Pourquoi le taux de mutation n'est-il jamais égal à zéro ? 268

2) Diversity of Olfactory Responses and Skills in Astyanax Mexicanus Cavefish Populations Inhabiting different Caves 276

C- Article en premier auteur uniquement disponible en version « pre-print » 298

I) Introduction générale

La biologie évolutive cherche à comprendre les mécanismes ayant mené à l'évolution et la diversification de tous les êtres vivants sur Terre, depuis un unique ancêtre commun cellulaire qui serait apparu il y a 3,9 milliards d'années (Betts et al. 2018). Je porte un grand intérêt pour cette science et rejoins l'idée de Theodosius Dobzhansky, qui dans son essai de 1973 écrit « *Seen in the light of evolution, biology is, perhaps, intellectually the most satisfying and inspiring science. Without that light it becomes a pile of sundry facts-some of them interesting or curious but making no meaningful picture as a whole* »[1].

La biologie évolutive peut prendre une dimension philosophique. Longtemps considérée comme la plus évoluée, nous savons maintenant que l'espèce humaine est soumise aux mêmes lois que celles qui régissent le reste du monde vivant. En ce sens, nous ne sommes que des transporteurs de gènes en compétition pour assurer leur propre transmission.

Outre les aspects fondamentaux de cette science, son importance dans la médecine ne fait plus de doute. Elle joue un rôle majeur dans la compréhension des cancers. Ces derniers apparaissent suite à la mutation d'un ou plusieurs gènes dans une cellule, qui va lui conférer un avantage sélectif majeur et mener à son amplification. Cette science permet également de mieux appréhender nos combats face aux microbes. On s'est notamment rendu compte que la pression de sélection induite par les antibiotiques sur les communautés bactériennes favorisait l'émergence de gènes de résistance, rendant d'autant plus compliqué le traitement des personnes malades. Ainsi, les pouvoirs publics, depuis les années 2000, ont lancé la campagne « Les antibiotiques, c'est pas automatique ». D'actualité au moment où je rédige cette thèse, on peut également noter à quel point il est important d'intégrer nos connaissances sur les mécanismes évolutifs du vivant dans la lutte contre les pandémies virales, par exemple dans la gestion de l'apparition de variants.

[1] Considérée à la lumière de l'évolution, la biologie est peut-être intellectuellement la science la plus satisfaisante et la plus inspirante. Sans cette lumière, cela devient un tas de faits divers - certains d'entre eux intéressants ou curieux mais ne faisant aucune image significative dans son ensemble.

Enfin, la biologie évolutive a une place centrale dans la protection de la biodiversité, menacée par les activités humaines. Une des grandes questions est de savoir à quel point et à quelle vitesse les espèces peuvent s'adapter aux modifications environnementales, comme la hausse de température ou la dégradation de leur milieu de vie. Dans ce cadre, il est important de mieux comprendre le fonctionnement et l'évolution des systèmes sensoriels qui permettant aux êtres vivants, que ce soient les procaryotes ou les eucaryotes, de percevoir l'environnement qui les entoure.

Pendant ma thèse, je me suis intéressé à l'évolution des systèmes sensoriels chez les actinoptérygiens, également appelés poissons à nageoires rayonnées. Ces derniers sont des modèles d'étude idéaux, car ils ont colonisé de nombreuses niches écologiques et présentent des caractéristiques morphologiques très diverses. J'ai notamment examiné les relations qui pourraient exister entre les changements de milieu de vies, de la morphologie, des yeux et de l'appareil olfactif, et l'évolution moléculaire des gènes liés à la vision et l'olfaction. Dans les sections suivantes, je présenterai le cadre de ma thèse, en décrivant les actinoptérygiens et plus particulièrement les téléostéens qui constituent la majeure partie de ce clade. Des concepts d'évolution moléculaire seront ensuite définis, notamment les gains et pertes de gènes en relation ou non avec des transitions écologiques. Enfin, les différentes modalités sensorielles qui existent chez les actinoptérygiens, et leur évolution depuis l'ancêtre commun le plus récent des vertébrés, seront introduites.

A- Les actinoptérygiens : un clade majeur de vertébrés

Les vertébrés, clade notamment caractérisé par la présence de vertèbres dont il tire son nom, semblent être apparus durant le Cambrien[2], leurs premières traces de vie ayant été retrouvées dans la province de Yunnan en Chine (Shu et al. 2003). Plus de 73 000 espèces sont maintenant décrites dans ce taxon (IUCN 2021), avec approximativement 6 500 mammifères, 11 000 oiseaux, 11 000 reptiles, 8 300 amphibiens, 35 000 actinoptérygiens (Eschmeyer, Fricke, et Van

[2] Période qui s'étend de −541 ± 1 à −485,4 ± 1,9 millions d'années.

Der Laan 2021), 1 200 chondrichtyens (Cavanagh et Gibson 2007) et une centaine d'agnathes (Janvier 2010) (Figure 1).

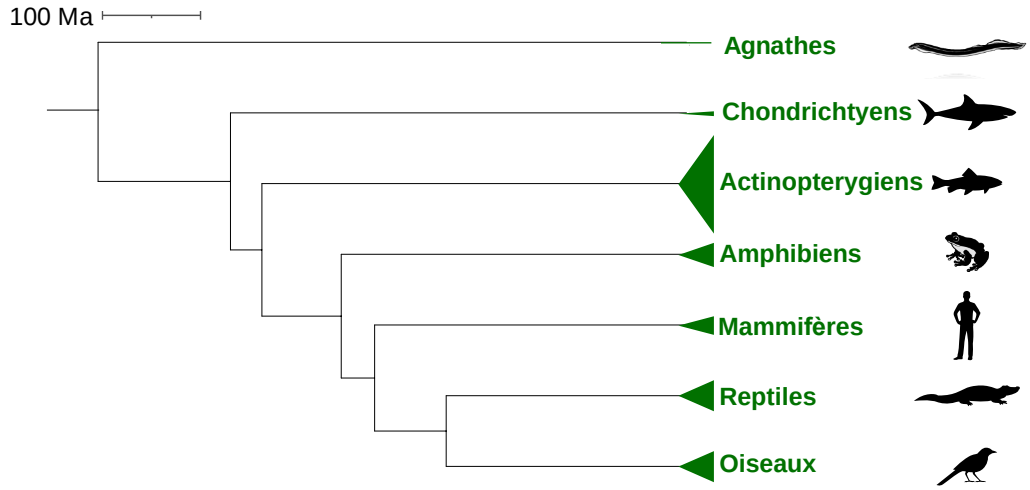


Figure 1 : Phylogénie des vertébrés.
Figure réalisée avec iTOL et temps de divergences obtenus sur timetree.org. Les tailles des triangles sont proportionnelles au nombre d'espèces dans chaque clade.

Les actinoptérygiens[3], également appelés poissons à nageoires rayonnées, partagent un ancêtre commun daté à plus de 400 millions d'années (Near et al. 2012). Ce groupe représente à lui seul la moitié des espèces de vertébrés et 95% des vertébrés aquatiques (Hughes et al. 2018). La majorité des actinoptérygiens fait partie des téléostéens, groupe d'espèces partageant une duplication entière du génome (appelé TGD pour « Teleost-specific Genome Duplication ») et dont une des caractéristiques est la présence d'un prémaxillaire (os de la mâchoire supérieure) mobile.

Une diversité fascinante existe au sein des téléostéens, aussi bien en ce qui concerne la forme (fusiforme, allongé, plat …) que la taille : les individus de certaines espèces mesurent quelques millimètres tels que le gobie pygmée (*Pandaka pygmaea*, *Gobiiformes*) ou le cyprinidé *Paedocypris progenetica*. D'autres mesurent plusieurs mètres comme le régalec (*Regalecus*

[3] Dans cette thèse, le mot poisson sera utilisé pour parler des actinoptérygiens.

glesne, *Lampridiformes*) et peuvent peser plusieurs tonnes comme le poisson lune (*Mola mola*, *Tetraodontiformes*) (Froese et Pauly 2021). Cette diversité se manifeste également dans les milieux de vie de ces poissons, avec des espèces vivant en eaux salée et/ou en eau douce, dans des espaces de la dimension d'une flaque d'eau à celle d'un océan, proches de la surface ou à des profondeurs allant jusqu'à 8000 mètres (Yancey et al. 2014) et plus ou moins près des côtes et des récifs coralliens. Certains poissons sont également amphibies, par exemple dans le genre *Boleophthalmus*.

Seulement une cinquantaine d'espèces d'Actinoptérygiens décrites ne sont pas des téléostéens et elles sont réparties en 4 ordres : les Polypteriformes, les Acipenseriformes, les Amiiformes et les Lepisosteiformes (Eschmeyer, Fricke, et Van Der Laan 2021, Figure 2). Ces espèces présentent des caractéristiques très différentes, que ce soit au niveau de leur régime alimentaire, de leur taille ou de leur milieu de vie. Par exemple, le polypteriforme *Erpetoichthys calabaricus* ne mesure que quelques dizaines de centimètres, se nourri de petits vertébrés, d'insectes et de crustacés tandis que l'acipenseriforme *Polyodon spathula* mesure entre 1 et 2 mètres et se nourri essentiellement de plancton (Froese et Pauly 2021).

Cette hétérogénéité morphologique, comportementale et écologique fait des actinoptérygiens un groupe idéal d'étude en biologie évolutive. Depuis quelques années, un nombre croissant de données génomiques est disponible pour ces espèces, avec 685 génomes d'actinoptérygiens, dont 678 téléostéens, assemblés sur la base de données NCBI[4], ce qui a permis d'améliorer considérablement la phylogénie de ce taxon pour laquelle il reste néanmoins quelques incertitudes (Figure 2, Hughes et al. 2018).

[4] A la date du 10 juin 2021.

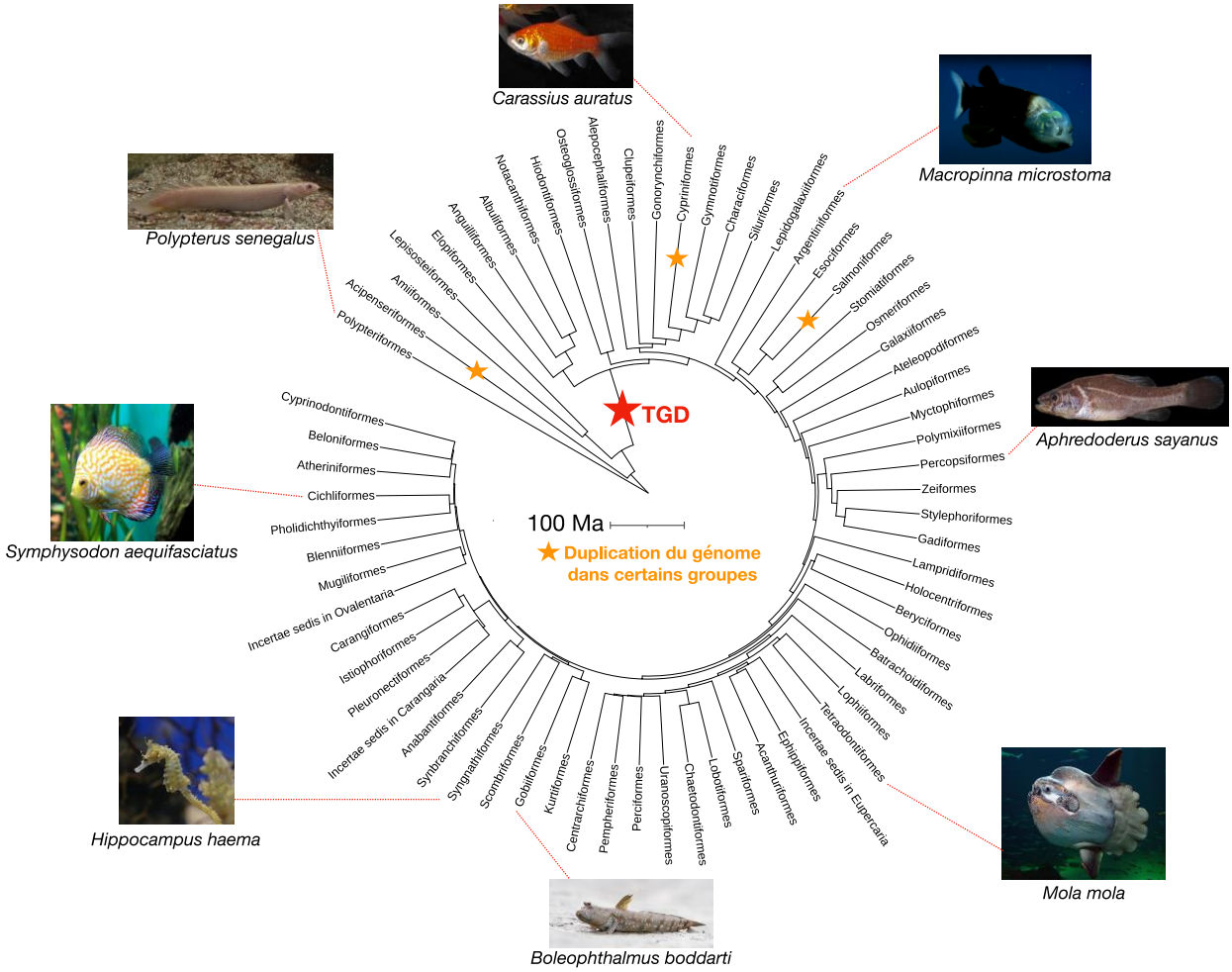


Figure 2 : Phylogénie des ordres d'actinoptérygiens avec quelques espèces représentatives de la diversité existante.
Figure réalisée avec iTOL. Les datations et relations phylogénétiques entre les différents ordres sont issus de https://fishtreeoflife.org/. Les ordres nommés « Incertae » ne sont plus monophylétiques d'après les données moléculaires récentes. L'évènement de duplication entière du génome commun aux téléostéens est indiqué par une étoile rouge, tandis que les étoiles oranges représentent des duplications entières de génomes plus récentes.

La diversification des téléostéens a impliqué plusieurs radiations évolutives[5], au niveau de l'ancêtre commun des téléostéens mais aussi plus récemment dans deux clades majeurs qui représentent 90% des espèces actuelles, que sont les Ostariophysi [6]et les Perciformes (Santini et al. 2009).
Deux principales hypothèses reposant sur l'architecture des génomes, non mutuellement exclusives, existent pour expliquer ces radiations évolutives : la grande diversité et abondance d'éléments transposables chez ces espèces et les duplications entières de génomes.

1) Diversité et abondance d'éléments transposables

Les éléments transposables, également appelés transposons, sont des séquences d'ADN capables de se déplacer et de s'amplifier dans les génomes. Les téléostéens possèdent la plus grande diversité d'éléments transposables des vertébrés, avec en moyenne 24 superfamilles de transposons par espèce contre uniquement 14 chez l'humain ou encore 13 chez la souris (*Mus musculus*) (Chalopin et al. 2015). Cette grande diversité d'éléments transposables va engendrer de nombreux réarrangements génomiques lors de leurs transpositions, mais également activer, désactiver, modifier, ou créer des gènes et éléments régulateurs en fonction de leur site d'intégration.

Cette activité va renforcer l'apparition de barrières reproductive et engendrer des évènements de spéciation. Ainsi, les génomes de téléostéens ont des tailles très hétérogènes, par exemple 358 Mb chez le tetraodon (*Dichotomyctere nigroviridis*) et 1317 Mb chez le poisson-zèbre (*Danio rerio*), principalement dues à l'abondance relative d'éléments transposables,

[5] Radiation évolutive : accélération du taux de spéciation, menant à l'apparition de nombreuses espèces dans une période de temps courte.
[6] Super-odre regroupant les Characiformes, les Cypriniformes, les Gonorynchiformes, les Gymnotiformes et les Siluriformes.

représentant 6% et 55% du génome respectivement chez ces deux espèces (Chalopin et al. 2015).

L'implication des transposons dans le remodelage des génomes et la spéciation est notamment visible dans le groupe des Nototheniidae (ordre des Perciformes), chez qui le nombre de chromosomes varie de 22 à 58, ce qui serait dû à la forte activité du transposon *DIRS1* (Auvinet et al. 2018) (Figure ci-dessous).

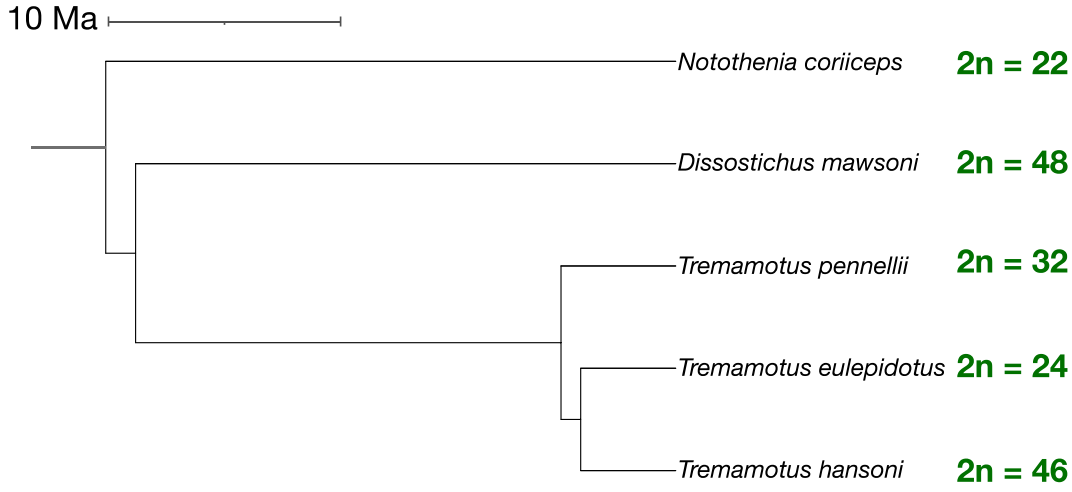


Figure 3 : Phylogénie avec cinq espèces représentatives des Nototheniidae. Pour une même taille de génome, ces espèces ont des nombres de chromosomes très variables (2n = 22 à 2n = 48). Cela serait dû à de nombreux réarrangements génomiques provoqués par la forte activité de certains transposons.

2) Duplications entières du génome

Une duplication entière du génome (=WGD pour « whole genome duplication ») double le matériel génétique d'une espèce, qui doublera ainsi son nombre de gènes. La divergence de ces copies de gènes, et l'éventuelle apparition de nouvelles fonctions pourront ainsi être facteurs de divergence et de spéciation (Volff 2005).

Il semblerait néanmoins que les duplications de génomes ne soient pas toujours corrélées avec des évènements de diversification. Par exemple, le taux de spéciation des plantes polyploïdes

n'est pas supérieur à celui des plantes diploïdes, et l'inverse est même observé, notamment dû à un taux d'extinction plus grand des espèces ayant subi une duplication récente de génome (Mayrose et al. 2011). De la même façon, une WGD chez les salmonidés, arrivée il y a environ 90 millions d'années, ne semble pas être liée à une augmentation du taux de spéciation, la plupart des lignées s'étant formées dans les 10 dernières millions d'années (Macqueen et Johnston 2014).
Il a été longtemps considéré que le TGD était à l'origine du succès évolutif des téléostéens. Cependant, des études récentes montrent que cette duplication a eu lieu il y a 235 millions d'années, soit bien avant la radiation et les innovations morphologiques observées chez ces espèces (Davesne et al. 2021).

Inversement, les cichlidés (ordre des Perciformes), qui ont connu une radiation évolutive dans plusieurs grands lacs d'Afrique, ne montrent pas d'activité particulièrement élevée d'éléments transposables et n'ont pas connu de duplications de génome récentes (Ronco et al. 2021).
Il apparaît donc que les bases moléculaires des rapides diversifications observées au sein des téléostéens restent encore mal comprises.

B- Évolution moléculaire par gains et pertes de gènes.

1) Gains de gènes

L'évolution d'une espèce peut impliquer l'apparition de nouveaux gènes, liés ou non à l'apparition de nouvelles fonctions. Par exemple, l'ancêtre commun des primates possédait une vision dichromatique, c'est-à-dire reposant uniquement sur deux opsines visuelles (protéines permettant de détecter la lumière). L'acquisition d'une troisième opsine, par le mécanisme de duplication de gène qui sera détaillé dans les chapitres suivants, chez l'ancêtre commun des Catarhiniens[7], a permis une vision trichromatique chez les espèces appartenant à ce groupe.

[7] Clade comprenant les humains et les macaques.

Trois mécanismes majeurs, décrits dans les sections suivantes, permettent l'apparition de nouveaux gènes au sein d'un génome (Andersson, Jerlström-Hultqvist, et Näsvall 2015).

a) Créations de-novo

L'apparition de gènes de-novo correspond à la naissance de gènes à partir de régions non-codantes de l'ADN. Pour que cela soit possible, ces régions doivent acquérir un promoteur, permettant l'initiation de la transcription, et une phase ouverte de lecture, c'est-à-dire une séquence débutant par un codon d'initiation de la traduction « ATG » et terminant par un codon de terminaison (« TGA », « TAA » ou « TGA »).

Les nouveaux gènes ainsi créés peuvent être impliqués dans l'apparition d'innovations évolutives. Chez les Gadidae (regroupant, entre autres, le tacaud norvégien *Trisopterus esmarkii* et la morue de l'atlantique *Gadus morhua*), l'apparition de-novo d'une protéine antigel a permis à certaines espèces de survivre dans les eaux glaciales de l'Arctique (Zhuang et al. 2019) (Figure ci-dessous).

Arctogadus glacialis

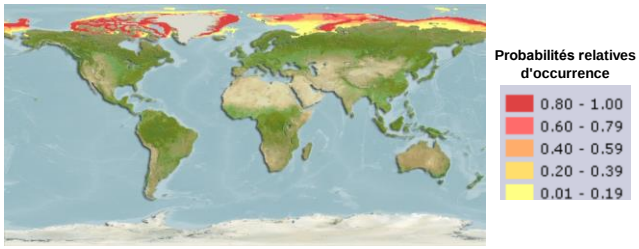


Figure 4 : Photographique du cabillaud arctique Arctogadus glacialis et carte de sa répartition.
Données issues d'AquaMaps 2019.

b) Transferts horizontaux de gènes

Les transferts horizontaux de gènes correspondent au passage de matériel génétique entre deux individus qui n'implique pas une relation de parent-descendant (qui seraient alors des transferts verticaux). Ces transferts sont fréquents chez les bactéries mais semblent plus rares chez les eucaryotes chez qui les mécanismes sous-jacents restent encore mal connus (Keeling et Palmer 2008). Ils peuvent toutefois avoir un rôle majeur dans l'apparition d'innovations chez une espèce. Le hareng (*Clupea harengus*) et l'éperlan arc-en-ciel (*Osmerus mordax*) possèdent le gène AFP, qui assure la même fonction antigel que le gène AFGP présent chez les Gadidae. Or, des analyses génomiques récentes ont permis de montrer que le gène AFP n'est pas apparu indépendamment chez ces deux espèces. Il n'a pas non plus été transmis de façon verticale depuis un ancêtre commun mais transmis de façon horizontale du hareng à l'éperlan (Graham et Davies 2021).

c) Duplications de gènes

La duplication est le mécanisme principal de création de nouveaux gènes. Un gène peut subir une duplication par différents mécanismes, impliquant soit une petite région du génome contenant ce gène soit le génome entier. Les duplications vont être à l'origine de la présence de deux copies de ce gène qui vont initialement assurer des fonctions redondantes. Les deux copies du gène vont ensuite accumuler des mutations pouvant mener à trois destins évolutifs différents (Lynch et Conery 2000).

Le destin le plus fréquent est la non-fonctionnalisation, c'est-à-dire l'extinction d'une des deux copies du gène par accumulation de mutations perte de fonction ou par délétion.

Une autre possibilité est l'acquisition d'une nouvelle fonction par une des deux copies du gène, que l'on appelle néo-fonctionnalisation tandis que l'autre copie garde la fonction ancestrale.

Enfin, il est possible, notamment pour les gènes codant des protéines pléiotropes (avec plusieurs fonctions) qu'apparaisse une sous-fonctionnalisation : chaque copie ne retiendra qu'une partie des fonctions ancestrales. Le maintien des deux copies du gène sera indispensable pour assurer l'ensemble des fonctions ancestrales.

Un des exemples les plus connus de néo- et sous-fonctionnalisation chez les téléostéens concerne les opsines des cichlidés. Chez l'espèce *Oreochromis niloticus,* la vision est assurée par sept opsines exprimées dans les cônes (cellules photoréceptrices de la rétine). Ce relativement grand répertoire d'opsine s'est mis en place après plusieurs évènements de duplication, à la suite desquels ces gènes ont évolué et acquis des spectres différents d'absorption de lumière (= néo-fonctionnalisation). De plus, ces gènes sont exprimés de manière différente pendant le développement de l'espèce, avec certaines opsines uniquement exprimées pendant le stade larvaire et d'autres plutôt exprimées chez l'adulte. La vision est donc assurée dans le temps par ces différentes opsines (= sous-fonctionnalisation temporelle) (Spady et al. 2006).

2) Pertes de gènes

L'évolution d'un génome peut également impliquer des pertes de gènes et des fonctions associées. Ils vont disparaitre par sélection naturelle si leur présence réduit la valeur sélective de l'organisme, par exemple dans le cas d'un gène de résistance à un antibiotique quand ce dernier n'est plus présent dans le milieu d'une bactérie (Koskiniemi et al. 2012). Les pertes vont également survenir par dérive génétique : lorsqu'un gène n'est plus utile pour un organisme et bien que sa présence ait un coût énergétique insignifiant, la pression de sélection pour son maintien sera relâchée et la séquence évoluera de façon neutre. Cette dernière accumulera des mutations jusqu'à ne plus être identifiable comme une ancienne séquence codante, ou sera éliminée du génome par des grandes délétions.

Sélection naturelle

La sélection naturelle est le processus par lequel un allèle qui confère un avantage reproductif aura plus de chance d'être transmis aux générations suivantes. Cet allèle augmentera en fréquence dans la population (On parle alors de sélection positive) pour finalement atteindre la fixation (fréquence de 1). Au contraire, un allèle désavantageux a moins de chance d'être transmis et sa fréquence dans la population décroîtra, jusqu'à potentiellement disparaître (sélection purificatrice). L'amplitude d'un avantage ou désavantage conféré par un allèle à un individu dépend de l'espèce et de l'environnement. Si les hétérozygotes ont une meilleure valeur sélective que les homozygotes, la diversité allélique peut se maintenir, et on parle dans ce cas de sélection balancée.

Dérive génétique

En 1968, Motoo Kimura dans son célèbre article « Evolutionary rate at the molecular level » (Kimura 1968) propose que la grande majorité des mutations qui apparaissent est neutre, ne conférant ni avantage ni désavantage. La fluctuation en fréquence d'une mutation neutre au sein d'une population serait alors uniquement due au hasard, mécanisme appelé dérive génétique, jusqu'à atteindre la fixation ou disparaître. La dérive génétique est d'autant plus forte que la taille efficace[1] (= Ne) de la population est faible (simplement dû au processus d'échantillonnage ; il est plus probable d'obtenir uniquement des 6 lors de trois lancés de dés que lors de dix lancés de dés). Dans le cas d'un allèle neutre venant d'apparaître dans une population diploïde de taille efficace Ne, sa probabilité de fixation est égale à $\frac{1}{2Ne}$, et elle prendra en moyenne un temps égal à $4Ne$ (Kimura et Ohta 1969). Il est également important de noter que dans une population, une mutation, même avantageuse, peut disparaître à cause de la dérive génétique, et de même un allèle désavantageux peut se fixer.

[1] : Taille d'une population idéale avec une même proportion de mâles et de femelles qui peuvent se reproduire de façon aléatoire et avec la même probabilité.

Les disparitions de séquences par grandes délétions restent difficiles à étudier car l'absence d'une séquence dans un génome peut être une réalité biologique, mais elle peut aussi être la conséquence d'un assemblage de génome incomplet. Par ailleurs, leur détection nécessite souvent une comparaison avec des espèces proches pour lesquelles un assemblage du génome est également disponible.

Les gènes fonctionnels peuvent également être perdus par pseudogénisation, processus directement observable à partie de la séquence d'un génome.
Un pseudogène est un gène devenu non fonctionnel suite à l'apparition de mutations « perte de fonction ». Ces dernières peuvent être : 1) des codons de terminaison de la traduction (= codon stop) prématurés ; 2) des petites délétions ou insertions (appelés « indels » en anglais) impliquant un nombre de nucléotides qui n'est pas un multiple de 3, et à l'origine d'un décalage du cadre de lecture du gène ; 3) des mutations aux sites d'épissage créant une séquence d'ARN messager contenant des introns ; 4) des mutations faisant disparaître le codon d'initiation (ATG) ou le codon de terminaison de la traduction (TGA, TAA ou TAG) du gène (Figure 5). Un gène peut également devenir non fonctionnel suite à l'accumulation de mutations non-synonymes délétères, mais l'impact de ces dernières reste difficile à évaluer sans études fonctionnelles.

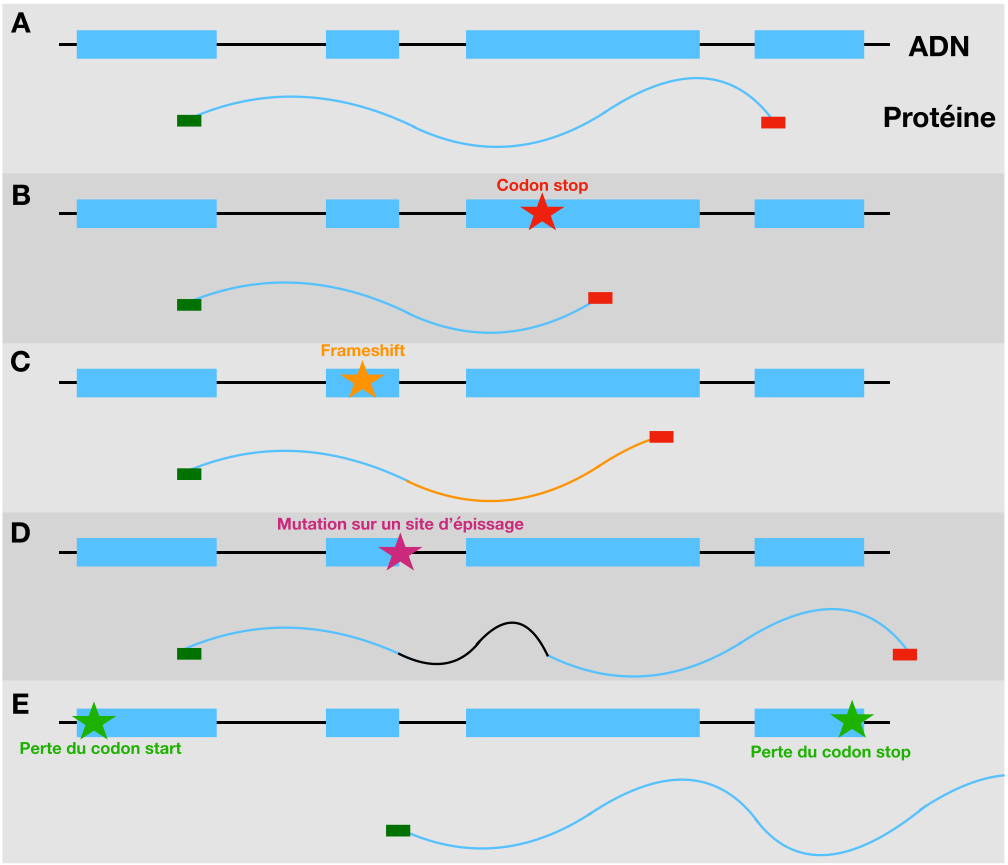


Figure 5 : Exemples de mutations perte de fonction et leur conséquence sur la séquence protéique.

A- Gène et protéine fonctionnels B- Codon stop prématuré C- Frameshift = décalage du cadre de lecture suite à une petite délétion ou insertion D- Mutation sur un site d'épissage E- Perte du codon d'initiation de la traduction (= codon start) et du codon de terminaison de la traduction (= codon stop). Souvent, les ARNm contenant des codons stop prématurés vont être dégradés avant même la traduction, par un mécanisme appelée « Nonsense-mediated mRNA decay ». Ainsi, les impacts sur les séquences protéiques indiquées dans la figure sont des impacts théoriques si l'ARNm n'est pas dégradé et qu'il est traduit.

De nombreux cas de pseudogénisation suite à des transitions écologiques ou comportementales sont connus au sein des vertébrés.

Le panda géant (*Ailuropoda melanoleuca*) est un mammifère de la famille des Ursidés (ours) et il en est le seul représentant herbivore. Le génome de cette espèce qui passe 55% de la journée à se nourrir principalement de bambou (99% de son alimentation) (IUCN 2016) a été séquencé en 2010 (Li et al. 2010). Ce génome a permis de mettre en évidence la pseudogénisation d'un gène (*Tas1r1*) responsable de la perception de l'umami, saveur que l'on retrouve dans les aliments riches en protéines tels que la viande et le fromage. Ainsi, la transition alimentaire du panda géant a rendu cette perception de l'umami dispensable, entrainant une évolution neutre du gène *Tas1r1* (Zhao et al. 2010).

Il est important de noter ici qu'il est souvent difficile de savoir si la perte d'un gène est la cause ou la conséquence d'une transition écologique ou d'une perte de structure. Dans le cas du panda, une première publication proposa que la perte de *Tas1r1* avait causé sa transition vers un régime herbivore (Li et al. 2010). Une seconde publication, utilisant une autre méthode de datation moléculaire et à l'aide de fossiles, propose que la perte de *Tas1r1* est une conséquence du régime herbivore, renforçant de fait ce mode d'alimentation (Zhao et al. 2010). Les datations moléculaires de cette seconde étude ayant un intervalle de confiance assez grand, le choix entre ces deux scenarios est difficile (Figure 6).

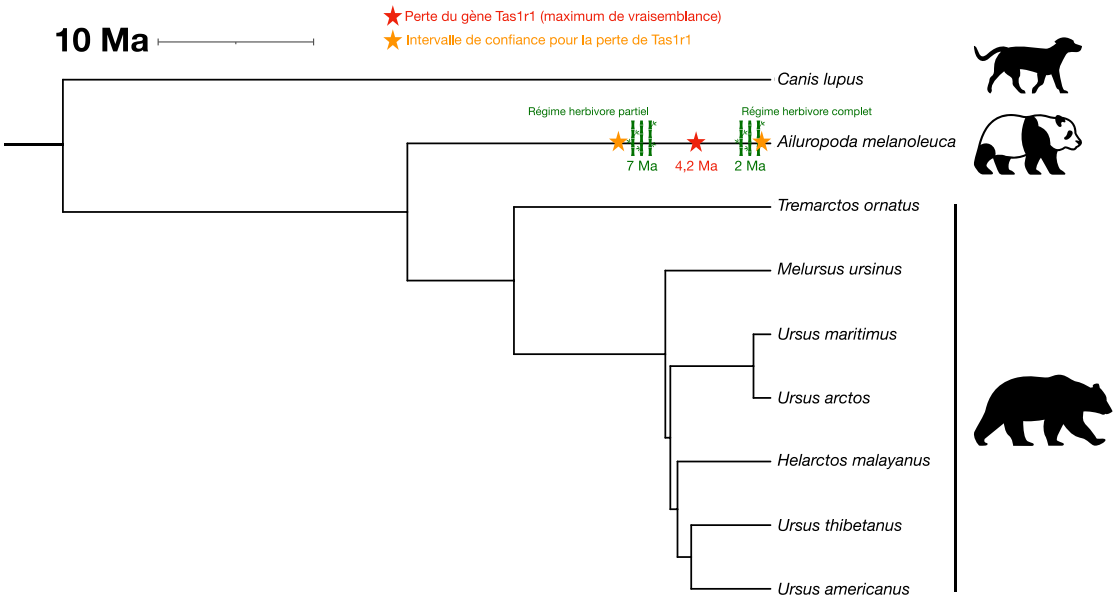


Figure 6 : Phylogénie des Ursidés raciné avec Canis lupus comme groupe externe. Temps de divergences obtenus sur TimeTree.org et visualisation sur iTOL. Les datations de la transition de régime alimentaire et de la perte du gène Tas1r1 sont issues de Zhao et al. (2010).

Un deuxième exemple concerne les oiseaux, qui possèdent un bec et un gésier et qui sont dépourvus de dents. Une recherche de gènes impliqués dans la formation des dents dans les génomes de 48 oiseaux a montré que ceux-ci étaient sous forme de pseudogènes et avec les mêmes mutations perte de fonction chez les différentes espèces. La probabilité qu'une même mutation se produise de façon indépendante étant pratiquement nulle, la perte des dents et des gènes s'est très certainement produite chez l'ancêtre commun de tous les oiseaux actuels, il y a plus de 116 millions d'années (Meredith et al. 2014).

Dans le cas d'une duplication entière et récente du génome, tous les gènes vont se trouver en deux copies, et pour la plupart de ces paires, il y aura non-fonctionnalisation d'une des deux versions du gène. Ainsi, l'analyse de génomes de cyprinidés et de salmonidés révèle la présence de milliers de pseudogènes sans perte de structures ou de fonctions au niveau de l'organisme. Comme indiqué précédemment, il faut noter que suite à l'accumulation de mutations, un gène puis un pseudogène évoluant de façon neutre depuis une longue période ne sera plus du tout détectable dans le génome.

Les gains et pertes de gènes sont essentiels dans l'évolution des familles multigéniques, groupes de gènes homologues qui assurent des fonctions similaires. Ces familles évoluent selon un modèle de naissance-et-mort[8], c'est-à-dire par des gains et des pertes de gènes récurrents qui font varier le nombre de gènes. Chez les vertébrés, de nombreuses familles, plus ou moins riches en gènes, suivent ce modèle d'évolution, comme les récepteurs des cellules T impliqués dans l'immunité, les gènes codant les histones ou encore les gènes impliqués dans la chémoréception, système sensoriel important pour de nombreux vertébrés (Nei et Rooney 2005).

C- Modalités sensorielles chez les actinoptérygiens

Il est nécessaire pour tout organisme vivant d'avoir une perception de son environnement afin de pouvoir survivre et se reproduire. Cinq classes de sensorialités existent au sein des actinoptérygiens : l'électroperception, la magnétoréception, la mécanoréception, la chémoréception et la photoréception (Atema et al. 1988; Naisbett-Jones et al. 2020). L'importance relative de ces différentes modalités sensorielles varie entre espèces. D'une part parce que selon l'environnement, les différentes modalités sensorielles ne sont pas toutes utiles de la même façon. D'autre part, parce qu'elles demandent souvent beaucoup d'énergie (Niven et Laughlin 2008) et un fort investissement dans un système sensoriel pourra entrainer une baisse d'investissement pour les autres systèmes, phénomène appelé compromis sensoriel.

1) L'électroperception

L'électroperception est la capacité de détecter des champs électriques. Elle peut être soit passive, soit active lorsque l'organisme est capable de générer sa propre électricité. Cette

[8] « birth-and-death model » en anglais, concept introduit par Masatoshi Nei dans les années 90 (Nei, Gu, et Sitnikova 1997)

modalité sensorielle était déjà présente chez l'ancêtre commun le plus récent (MRCA[9]) des vertébrés (Bodznick et Northcutt 1981; Modrell et al. 2011) avant d'être perdu chez l'ancêtre commun des Néopterygiens, taxon regroupant les Amiiformes, les Lepisosteiformes et les téléostéens (Figure 7). Celle-ci est ensuite réapparue dans certains groupes de téléostéens, le plus souvent de manière passive comme chez les Siluriformes (Collin et Whitehead 2004). Seulement deux groupes de téléostéens sont capables d'électroperception active, les Mormyridés (famille d'Osteoglossiformes) et les Gymnotiformes (Crampton 2019).

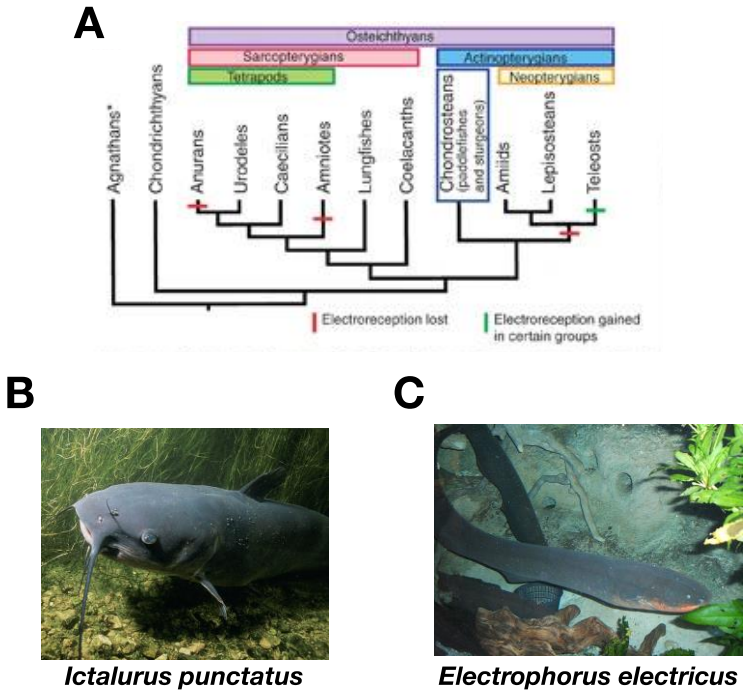


Figure 7 : A- Gains et pertes secondaires de l'électroperception chez les vertébrés. Les pertes sont indiquées par des barres rouges sur la phylogénie tandis que la barre verte indique des gains secondaires dans certains clades. Modifié d'après Modrell et al. (2011). B- Ictalurus punctatus (Siluriforme) capable d'électroperception passive C- Electrophorus electricus (Gymnotiforme), capable d'électroperception active.

L'électroperception passive est assurée par des cellules spécialisées, les électrorécepteurs ampullaires, qui peuvent détecter des fréquences électriques faibles, inférieures à 30 Hz. Leur

[9] MRCA = most recent common ancestor = ancêtre commun le plus récent

principale fonction est la reconnaissance de potentiels d'actions cellulaires (bioélectromagnétisme) d'autres organismes, qui peuvent être des proies, des prédateurs ou des congénères (Peters et al. 2002).

Les organismes capables d'électroperception active sont équipés d'électrorécepteurs tubéreux, qui détectent de fortes fréquences électriques, allant de 50 Hz à 18 kHz. Ces espèces possèdent un organe électrique, dans la queue chez les Mormyridés ou au niveau du ventre chez les Gymnotiformes (Crampton 2019). Elles sont ainsi capables d'électrolocalisation, en produisant un champ électrique et en détectant les perturbations causées par des objets biotiques ou abiotiques sur celui-ci. L'électroperception active permet l'envoi et la détection de signaux électriques sociaux, avec des implications importantes dans la reproduction et dans les comportements d'agrégation (Crampton 2019). Afin d'éviter un brouillage du sens de l'électroperception, certaines espèces pourront moduler la fréquence de leur champ électrique, pour que celle-ci soit différente des individus à proximité (Fortune et al. 2020).

2) La magnétoréception

La magnétoréception est la capacité de percevoir les champs magnétiques. Chez les téléostéens, elle a surtout été étudiée chez les poissons migrant sur de longues distances, tels que les anguilles ou les saumons.

Les anguilles d'Europe (*Anguilla anguilla*) naissent dans la mer des Sargasses[10], au nord de l'Océan Atlantique, après quoi elles migrent vers l'Europe, puis retournent dans leur mer originelle pour pondre et finir leur vie lorsqu'elles atteignent la maturité sexuelle (Durif et al. 2013). Ainsi, ces espèces doivent parcourir de très longues distances, la mer des Sargasses étant à plus de 5000 km des rivières européennes dans lesquelles elles sont aussi trouvées.

[10] Une étude récente montre qu'un deuxième site de reproduction de cette espèce pourrait exister, 2000 km à l'est de la mer des Sargasses, au niveau de la ride médio-atlantique (Chang et al. 2020).

De la même manière, les saumons, après leur naissance dans l'eau douce de rivières, partent vivre dans l'océan avant de retourner pondre dans leur rivière natale à l'âge de la reproduction (Figure 8), représentant des voyages de milliers de kilomètres.

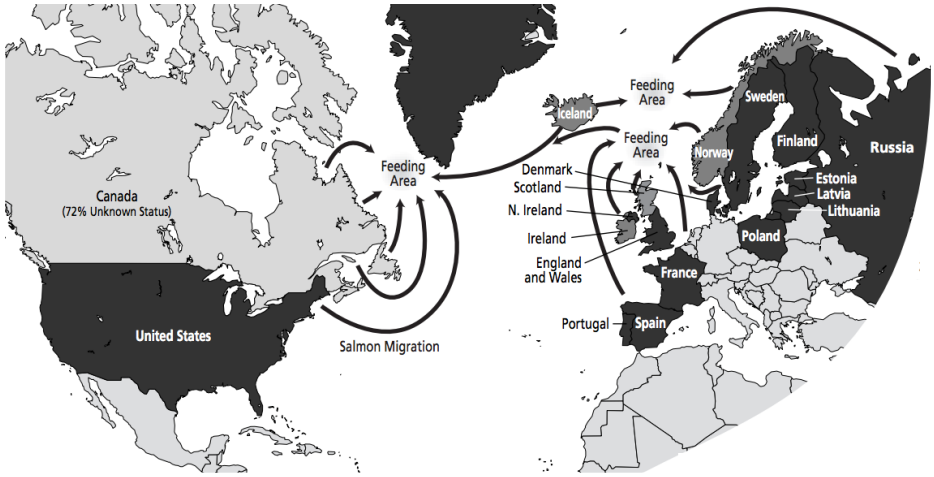


Figure 8 : Routes migratoires des saumons de l'atlantique (Salmo salar). Modifié d'après WWF (2001). Feeding area = zone d'alimentation.

La migration de ces espèces serait en grande partie permise par leur capacité à s'orienter grâce au champ magnétique terrestre, une capacité qui pourrait également être utilisée par certaines espèces non-migratrices (Formicki, Korzelecka-Orkisz, et Tański 2019). Deux mécanismes principaux seraient utilisés par les téléostéens pour détecter les champs magnétiques :

- Le déplacement des poissons dans l'eau salée, fortement conductrice, induirait des courants électriques grâce au champ magnétique terrestre[11]. Les poissons équipés d'électrorécepteurs seraient capables de détecter ces courants à des fins de prédation ou d'évitement. De plus, un tel courant électrique produit par une espèce qui possède des électrorécepteurs pourrait lui indiquer la direction dans laquelle elle nage (Formicki, Korzelecka-Orkisz, et Tański 2019) ;

[11] Ce phénomène vient de la loi de Faraday : en présence du champ magnétique terrestre, une force locomotrice dans l'eau induit une force électromotrice (Nature 1948).

- La production de magnétite par l'organisme, minéral magnétique agissant comme une boussole. Ce serait le mécanisme principal utilisé par les espèces migratrices (Naisbett-Jones et al. 2020).

La magnétoréception est actuellement la modalité sensorielle des poissons la moins bien comprise. C'est notamment dû au fait que contrairement aux autres modalités sensorielles, il n'a pas encore été montré qu'il existait des cellules spécialisées ou de gènes spécifiques pour la détection des champs magnétiques (Formicki, Korzelecka-Orkisz, et Tański 2019). Les principales études reposent uniquement sur des expériences comportementales.

3) La mécanoréception

La mécanoréception est la capacité de détecter des stimuli mécaniques, notamment par les systèmes de la ligne latérale et de l'ouïe.

Les organes de la ligne latérale, situées sur les flancs et la tête de l'animal, sont composés de cellules ciliées spécialisées appelées neuromastes. Ces cellules sont soit superficielles, à la surface de l'organisme, soit à l'intérieur de canaux remplis de fluide. Elles servent notamment à détecter les mouvements ou vibrations de l'eau de fréquences comprises entre ~20 et 100-150 Hz (Coombs, Braun, et Donovan 2001; Bleckmann 2008).

Les lignes latérales permettent aux poissons de localiser la présence et la force des courants, jouant un rôle majeur dans le comportement de rhéotaxie, positive quand l'individu choisit de rester dans le courant, ou négative quand l'individu l'évite. La détection des vibrations dans l'eau permet également la détection de proies et d'objets inanimés (Bleckmann 2008; Yoshizawa et al. 2010). Les lignes latérales sont partagées par plusieurs clades au sein des vertébrés (par exemple les amphibiens), ce qui indique qu'elles étaient déjà présentes chez leur MRCA. Des pertes secondaires se sont produites, notamment chez le MRCA des amniotes[12] en relation avec l'adaptation à l'environnement terrestre (Piotrowski et Baker 2014).

[12] Tétrapodes disposant d'un sac amniotique, regroupant les mammifères et les reptiles (oiseaux inclus).

L'autre volet de la mécanoréception est l'ouïe, qui est la capacité à percevoir des sons. La plupart des téléostéens ont une ouïe sensible à des fréquences situées entre ~200 et ~10 KHz mais certaines espèces, par exemple quelques représentant des Clupéidés, peuvent détecter des ultrasons (>20 KHz) (Popper et al. 2004).

Trois types de sons existent : la géophonie (sons d'origine abiotique), la biophonie (sons d'origine biotique) et l'anthropophonie (sons produits par les activités humaines). Le son se déplaçant dans l'eau cinq fois plus vite que dans l'air, il s'atténue beaucoup moins rapidement et se propage sur de plus grandes distances, permettant aux espèces marines de détecter des sons d'origines plus lointaines que les espèces terrestres (Putland, Montgomery, et Radford 2019). La perception du son chez les poissons se fait grâce à l'oreille interne et la production du son[13] se fait par différents mécanismes, par exemple par claquement de dents ou par vibration de la vessie natatoire (Putland, Montgomery, et Radford 2019). Ces sons servent d'outil lors de communications entre congénères (Ladich 2000), lors de la recherche de nourriture et de partenaires, ou lors de comportements agonistiques (Hyacinthe, Attia, et Rétaux 2019) (Figure 9). Tout comme les lignes latérales, l'oreille interne est une structure partagée par tous les clades de vertébrés indiquant une apparition chez leur MRCA (Fritzsch et Elliott 2017).

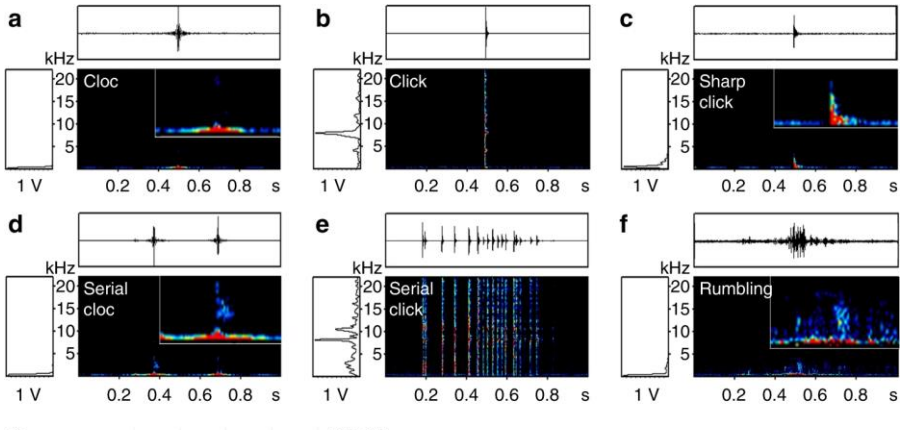


Figure 9 : Sonagrammes de 6 sons produits par l'espèce Astyanax mexicanus. Le « sharp click » joue un rôle dans les comportements agonistiques ou la recherche de nourriture, selon la population étudiée. Modifié d'après Hyacinthe, Attia, et Rétaux (2019).

[13] Les espèces capables de produire un ou plusieurs sons sont dites « soniques ».

4) La photoréception et la chémoréception

Les deux autres modalités sensorielles des actinoptérygiens sont la photoréception et la chémoréception. Contrairement aux modalités sensorielles précédentes, elles reposent sur l'existence de récepteurs moléculaires spécifiques, ce qui rend particulièrement intéressant leur étude au niveau des gènes. La photoréception permet la détection de lumière de longueurs d'ondes variées et elle peut être visuelle ou non-visuelle. La chémoréception est la capacité de détecter des signaux chimiques, par l'olfaction et la gustation (Ueda 2021).
Ces deux modalités sensorielles seront étudiées en détail dans cette thèse, avec : une étude de l'évolution des gènes de l'olfaction chez les actinoptérygiens et une étude de gènes de la photoréception chez des espèces cavernicoles, vivant partiellement ou totalement dans l'obscurité.

II) La chémoréception chez les vertébrés

A- La gustation

La gustation, est une modalité sensorielle liée à l'alimentation permettant d'évaluer la qualité de la nourriture avant de l'avaler. Cette fonction est assurée par les bourgeons gustatifs, organes en forme d'oignons composés de cellules sensorielles et de cellules de soutien. On en trouve chez tous les vertébrés à l'exception des myxines (famille faisant partie des agnathes) chez qui la gustation est assurée par les organes de Schreiner (Barreiro-Iglesias, Anadón, et Rodicio 2010). Ces derniers ont une fonction similaire mais sont apparus indépendamment.

Le nombre de bourgeons gustatifs est très variable chez les vertébrés, une vingtaine chez la mésange bleue *Cyanistes caeruleus* et près de 20 000 chez le sanglier *Sus scrofa* (Roura, Baldwin, et Klasing 2013). Ce nombre peut également être très différent au sein des téléostéens,

avec environ 2000 bourgeons chez le poisson-zèbre *Danio rerio* et près d'un million chez le poisson-chat *Ictalurus punctatus* (Kasumyan 2019) (Figure 10).

Alors que ces organes se trouvent uniquement dans la cavité buccale des tétrapodes (à l'exception de quelques amphibiens), ils peuvent être présents sur toute la surface du corps des poissons. Chez le poisson-chat, de nombreux bourgeons gustatifs sont placés sur les barbillons, permettant de localiser des sources de nourritures éloignées (Kasumyan 2019). Cela rend un peu plus difficile de distinguer les rôles relatifs de la gustation et de l'olfaction (autre volet de la chémoréception qui sera détaillé dans le prochain chapitre).

Figure 10 : Nombre de bourgeons gustatifs chez trois téléostéens (Danio rerio, Phoxinus phoxinus et Ictalurus punctatus) et trois tétrapodes (Cyanistes caeruleus, Homo sapiens et Sus scrofa).

Les cellules sensorielles des bourgeons gustatifs expriment des gènes codant des protéines à sept domaines transmembranaires qui servent de récepteurs, et qui sont classés en deux familles : les gènes *T1R* impliqués dans la reconnaissance de saveurs sucrées et umami, et les gènes *T2R* impliqués dans la reconnaissance des saveurs amères, souvent associées à la toxicité. Chez les vertébrés, le nombre de gènes *T1R* varie entre 0 chez le crapaud (*Xenopus tropicalis*) et 1 chez le poisson-zèbre (*Danio rerio*) à 3 chez l'humain et la souris, tandis que le nombre de gènes *T2R* varie entre 3 chez le poulet (*Gallus gallus*) à une cinquantaine chez le crapaud (Shi et Zhang 2006). La taille des répertoires de gènes serait associée aux comportements d'alimentation des différentes espèces. Ainsi, la proportion de plantes dans le régime alimentaire des vertébrés est corrélée avec le nombre de gènes *T2R*, peut-être parce que les plantes produisent plus souvent des toxines que les animaux (Li et Zhang 2014).

B- L’olfaction

1) Rôle de l’olfaction

La capacité à détecter les odeurs est une modalité sensorielle majeure chez les vertébrés. Elle peut jouer un rôle dans de nombreux comportements. Par exemple dans la reproduction, par la reconnaissance de phéromones sexuelles produites par les congénères. Chez les poissons, ces dernières sont majoritairement des stéroïdes gonadiques ou des prostaglandines (Lastein, Hamdani, et Døving 2006). Il a aussi été montré que les larves de poissons vivant près des récifs coralliens utilisent leur odorat pour détecter et choisir leur habitat au sein de ces environnements (Lecchini et Nakamura 2013) . L’olfaction est aussi impliquée dans la détection de nourriture et de prédateurs. Les Ostariophysi sont capables de produire une substance d’alarme dans leur épiderme, qui est relâchée dans l’eau lorsque la peau est abîmée par une attaque. Ainsi, la détection de cette substance par des congénères va induire des changements comportementaux afin de réduire la pression de prédation, tels que des agrégations en banc ou des nages erratiques (v. Frisch 1938; Maximino et al. 2019).

2) Organisation du système olfactif chez les actinoptérygiens

Les poissons ont deux cavités nasales (une de chaque côté de la tête), et chacune de ces deux cavités possède deux ouvertures : une narine antérieure (= narine pour l’inhalation) et une narine postérieure (= narine pour l’exhalation), permettant la formation d’un courant d’eau entrant et sortant de la cavité (ils possèdent donc deux paires de narines) (Figure 11, A). Quelques exceptions existent, notamment les cichlidés et certains pomacentridés qui ne possèdent qu’une narine par organe olfactif, l’eau entrant et ressortant par le même orifice (Barlow 2008).

Figure 11 : Système olfactif des actinoptérygiens
A- Photographie d'une narine antérieure (AN) et postérieure (PN) chez l'espèce Micropterus salmoides. Crédits : Kim, Yun, et Park (2019) B- Photographie de l'épithélium olfactif de l'espèce Fistularia tabacaria, composée de 17 lamelles arrangées en forme de rosette. C- Images de microscopie électronique de trois types de neurones olfactifs chez le Danio rerio. De gauche à droite : un neurone cilié, un neurone à microvillosité et un neurone dit « de crypte ». Modifié d'après Yoshihara (2014).

Dans chacune des cavités nasales se trouve un épithélium olfactif. Chez la plupart des téléostéens, celui-ci a une forme de rosette, c'est-à-dire un arrangement circulaire de lamelles (Figure 11, B). Ces lamelles permettent d'accroître la surface d'échange avec le milieu extérieur, et le nombre de lamelles au sein de l'épithélium diffère en fonction de l'espèce. Certaines espèces ont un épithélium olfactif composé de plusieurs dizaines de lamelles, d'autres de seulement une lamelle voire d'aucune lamelle, auquel cas l'épithélium est qualifié de plat. (Dymek et al. 2021).

Dans cet épithélium se trouvent des neurones sensoriels olfactifs permettant de détecter les molécules odorantes, de diverses formes et reconnaissables par leur signature moléculaire, c'est-à-dire les gènes qu'ils expriment (Figure 11, C) (Figure 12). Chez le poisson-zèbre par exemple, cinq types de neurones olfactifs existent et expriment à leur surface des récepteurs olfactifs, qui sont des protéines transmembranaires, sur lesquels se lient les molécules odorantes. La liaison d'une molécule va provoquer un changement de conformation du récepteur et ainsi activer une voie de signalisation (Wakisaka et al. 2017). Ces protéines sont souvent composées d'une partie conservée permettant leur ancrage dans la membrane, et d'une partie variable permettant de reconnaître une ou plusieurs molécules odorantes. Chaque neurone n'exprime qu'un seul récepteur olfactif parmi un répertoire dont la taille dépend du

nombre de gènes codant des récepteurs dans le génome, selon la règle « un neurone – un récepteur » (Figure 12). Cependant, certaines exceptions existent, avec des cas de co-expression de plusieurs récepteurs au sein d'un même neurone (Sato, Miyasaka, et Yoshihara 2007).

Les neurones sensoriels olfactifs projettent leur axone vers le bulbe olfactif, lui-même organisé en plusieurs régions sphériques appelées glomérules. Tous les neurones exprimant le même récepteur olfactif vont projeter leur axone vers le même glomérule. Les glomérules, après avoir traité l'information sensorielle, relaient l'information perçue vers différentes régions du cerveau. L'information y sera intégrée pour induire une réponse comportementale (Figure 12). Chez les téléostéens, quatre régions du cerveau semblent être impliquées dans ce décodage : le télencéphale ventral et dorsal, le tubercule olfactif et l'habénula (Yoshihara 2014).

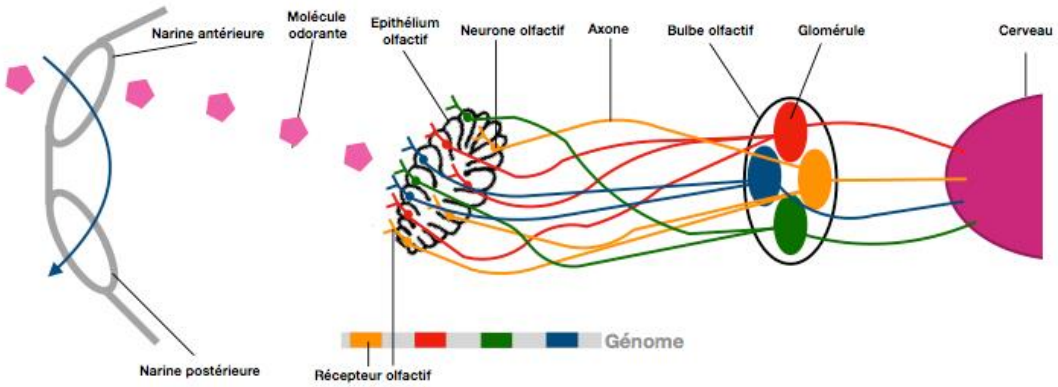


Figure 12 : Schéma du système olfactif chez les téléostéens

3) Reconnaissance des molécules odorantes par les récepteurs olfactifs

a) Spécificité des récepteurs olfactifs

La diversité des odeurs reconnaissables par une espèce dépend largement du nombre de récepteurs olfactifs, qui sont codés par des gènes appartenant à quatre familles : les *OR*, les *TAAR*, les *V1R* et les *V2R*. La nomenclature de ces gènes est principalement issue des études

chez la souris. Par exemple, la lettre « V » dans *V1R* et *V2R* est l'initiale de « voméronasal », un organe présent chez les tétrapodes et spécialisé dans la reconnaissance des phéromones, mais absent chez les actinoptérygiens. Pour ce groupe d'espèces, ces deux familles de gènes sont nommées respectivement *ORA* et *OlfC*.

Un des principaux défis dans les études de l'olfaction est de réussir à « dé-orphaniser » ces récepteurs, c'est-à-dire déterminer les molécules odorantes détectées par ces derniers.

Chez les téléostéens, uniquement deux récepteurs, un dans la famille *ORA* et un dans la famille *TAAR* ont pu être dé-orphanisés en utilisant l'organisme modèle *Danio rerio*. Chez cette espèce, on sait que le gène *ORA1* détecte l'acide 4-hydroxyphénylacétique, et cette molécule semble induire un comportement de ponte chez la femelle. Cela indiquerait que tout comme les gènes *V1R* des tétrapodes, les gènes *ORA* des téléostéens sont impliqués dans la reconnaissance de phéromones sexuelles (Ahuja et Korsching 2014). Le gène *TAAR13c* a une grande affinité pour la cadavérine, molécule relâchée lors de la putréfaction d'un cadavre. La reconnaissance de cette molécule induit un comportement de fuite chez le poisson-zèbre (Hussain et al. 2013). Cependant, le *Danio rerio* étant équipé d'environ 150 *OR*, 6 *ORA*, 50 *OlfC* et 100 *TAAR*, la spécificité moléculaire des récepteurs olfactifs demeure inconnue pour la très grande majorité.

b) Variabilité du nombre de récepteurs olfactifs

Chez les vertébrés, une grande variabilité du nombre de récepteurs olfactifs existe, avec parfois des liens avec le milieu et mode de vie des différentes espèces. Je présenterais les connaissances acquises sur les différentes familles de récepteurs olfactifs dans les sections suivantes, avant de présenter les résultats de mes travaux sur l'évolution de ces familles chez les actinoptérygiens.

OR

La plus grande famille de récepteurs olfactifs chez les vertébrés est la famille des *OR* (Figure 13). Pour cette raison, elle est la famille la plus étudiée, également dû au fait que ces gènes, contrairement à ceux des autres familles, ne sont composés que d'un seul exon, ce qui facilite beaucoup leur extraction des séquences génomiques.

Le nombre moyen d'*OR* chez les mammifères est de 800, mais une très grande variabilité existe. Chez le dauphin *Tursiops truncatus*, seulement 58 *OR* ont été trouvés alors que ce nombre est de 2 514 chez l'éléphant *Loxodonta africana* (Hughes et al. 2018). Chez les téléostéens, le répertoire d'*OR* a été étudié chez un nombre très réduit d'espèces et il semble qu'il y ait, en moyenne, 100 *OR* fonctionnels par espèce (Niimura 2009).

Cette famille multigénique a été divisée en sept sous familles : six sous familles de type I (α, β, δ, ε, γ, ζ) et une sous famille de type II (η), qui se différencient notamment par leurs relations phylogénétiques. Chez les tétrapodes, les deux sous-familles majoritaires sont les familles α et γ , tandis que les téléostéens ont complètement perdu la famille α et n'ont que quelques gènes dans la famille γ. Ces derniers ont une majorité de gènes δ et η. Cette observation a mené à l'hypothèse selon laquelle les familles α et γ seraient impliquées dans la reconnaissance de molécules odorantes dans l'air tandis que les autres familles seraient plus spécialisées pour les molécules odorantes solubles dans l'eau (Niimura et Nei 2005).

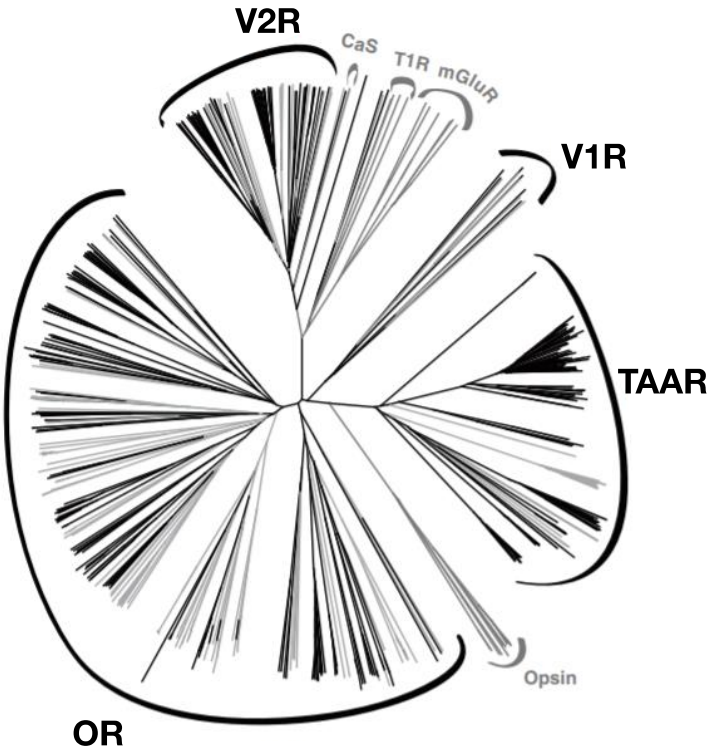


Figure 13 : Arbre phylogénétique des quatre familles de récepteurs olfactifs, extraites du poisson-zèbre (Danio rerio) et du Fugu (Takifugu rubripes).
Les clades dont les noms sont en gris correspondent à des gènes codant des protéines à sept domaines transmembranaires mais qui ne codent pas des récepteurs olfactifs et qui sont utilisées ici en tant que groupes externes. Modifié d'après S. Korsching (2008).

TAAR

Contrairement à la famille OR, les gènes *TAAR* semblent beaucoup plus nombreux chez les téléostéens que chez les tétrapodes, avec une centaine de *TAAR* fonctionnels chez le *Danio rerio* contre uniquement 15 chez la souris ou encore 6 chez l'humain. Le dauphin n'a aucun gène *TAAR* fonctionnel (Eyun et al. 2016). La famille *TAAR* a été divisée en trois sous-familles, nommées classe I, classe II (non monophylétique) et classe III.

Bien que la majorité des *TAAR* ne possède qu'un exon chez les vertébrés, certains téléostéens possèdent une grande proportion de *TAAR* avec deux exons (Azzouzi, Barloy-Hubler, et Galibert 2015).

V1R* et *V2R

Les gènes des familles *V1R* et *V2R* ont un nombre plus élevé d'exons, de 1 à 4 pour les *V1R* et de 6 à 13 pour les *V2R*.

Ces gènes sont également présents en nombre extrêmement variable chez les vertébrés, avec une absence totale de *V1R* chez les oiseaux, contre presque 300 gènes chez l'ornithorynque *Ornithorhynchus anatinus* (Nikaido 2019). Un groupe de 6 gènes *V1R* (*ORA*) semble conservé chez la plupart des téléostéens, avec de rares cas de gains et pertes indépendants chez certaines espèces (Zapilko et Korsching 2016). Il en est de même pour les *V2R*, avec aucun gène trouvé chez l'humain contre environ 250 gènes chez le crapaud *Xenopus tropicalis*. Chez les téléostéens, bien qu'un très faible nombre de gènes *OlfC* ait été caractérisé dans la littérature, une grande variabilité semble également exister avec 4 gènes chez le tétraodon *Dichotomyctere nigroviridis* contre 44 gènes chez le *Danio rerio* (Shi et Zhang 2007).

c) Mécanismes d'expansion et de contraction du répertoire de récepteurs olfactifs

En tant que familles multigéniques, l'expansion et la contraction du nombre de récepteurs olfactifs suivent un modèle de gains et pertes de gènes récurrents (par duplications et par pseudogénisations/délétions respectivement), également appelé « modèle naissance-et-mort » (birth-and-death model décrit précédemment, Nei, Gu, et Sitnikova 1997). Le nombre de gènes peut varier de façon aléatoire, par fixation et perte de copies sous le seul effet de la dérive génétique, mais également sous l'effet de la sélection, par exemple pour adapter les capacités olfactives en fonction d'un nouvel environnement ou de nouveaux comportements.

Figure 14 : Exemples d'espèces ayant subi une expansion ou une contraction du répertoire de gènes OR.

Le nombre d'OR fonctionnels observés dans les génomes de ces espèces est indiqué. 2 514 pour Loxodonta africana, 1 310 pour Sus scrofa, 247 pour Megaderma lyra et 58 pour Tursiops truncatus.

Les cétacés ont perdu la grande majorité de leurs gènes *OR* suite au retour à une vie aquatique, il y a environ 50 millions d'années (Liu et al. 2019, Figure 14). En effet, les tétrapodes n'ont gardé majoritairement que les familles permettant de détecter les molécules odorantes dans l'air. Celles-ci sont devenues inutiles lorsque les cétacés sont retournés dans l'eau, ce qui a conduit à leur perte, suite à une absence de sélection pour leur maintien. Cette transition écologique étant assez récente, il est fascinant d'observer que de nombreuses traces de gènes *OR* sont encore identifiables dans les génomes de ces espèces marines sous la forme de pseudogènes. Par ailleurs, et même si les baleines à dents, comme les dauphins, ont un système olfactif dégénéré (Kishida et al. 2015), il reste encore quelques gènes *OR* fonctionnels. Nous ne savons pas s'ils sont encore présents par chance ou s'ils sont nécessaires et maintenus par sélection.

De nombreuses pertes ont également été observées chez les chauves-souris, probablement en relation avec un compromis sensoriel entre l'olfaction et l'écholocalisation utilisée par certaines de ces espèces (Tsagkogeorga et al. 2017, Figure 14). De la même façon, un faible nombre de récepteurs olfactifs est observé chez les primates ayant acquis la vision trichromatique, bien

que celle-ci ne semble pas totalement expliquer la contraction du répertoire de gènes *OR* (Matsui, Go, et Niimura 2010). Au contraire, et pour des raisons encore inexpliquées, il y a eu chez les éléphants et les sangliers une grande expansion de ce répertoire (Nguyen et al. 2012 ; Niimura, Matsui, et Touhara 2014) (Figure 14).

Contrairement aux tétrapodes chez qui de nombreuses études comparatives ont été menées avec quelques exemples cités ci-dessus, très peu de publications se sont intéressées à l'évolution des différentes familles de récepteurs olfactifs chez les actinoptérygiens. Dans notre étude qui sera présentée dans les chapitres suivants, nous avons tiré profit du nombre croissant de génomes de poissons disponible dans les bases de données publiques pour mettre à jour les connaissances sur l'évolution de l'olfaction dans ce groupe, au niveau moléculaire, mais également au niveau morphologique.

III) Résultats - Évolution des quatre familles de récepteurs olfactifs chez les actinoptérygiens

Les résultats de mes travaux présentés dans les chapitres suivants abordent l'évolution de toutes les familles de récepteurs olfactifs chez les actinoptérygiens. Seulement une sous-partie de ce travail, qui concerne la famille *OR* et focalisée sur son évolution chez les téléostéens, est pour l'instant publiée.

A- Génomes utilisés et méthode de détection des gènes

1) Récupération et sélection de génomes depuis les bases de données publiques

Nous avons construit un jeu de données de 335 génomes d'actinoptérygiens disponibles avant avril 2021 et téléchargés depuis ENSEMBL et NCBI. Parmi ces 335 génomes, 328 sont des génomes de téléostéens et sept sont des génomes de non-téléostéens représentants les quatre ordres :

- Le bichir gris *Polypterus senegalus* (Polypteriformes)
- Le poisson roseau *Erpetoichthys calabaricus* (Polypteriformes)
- L'esturgeon *Acipenser ruthenus* (Acipenseriformes)
- Le spatulaire *Polyodon spathula* (Acipenseriformes)
- L'amie *Amia calva* (Amiiformes)
- Le lépisostée tacheté *Lepisosteus oculatus* (Lepisosteiformes)
- Le garpique alligator *Atractosteus spatula* (Lepisosteiformes)

Pour chaque génome, la taille attendue a été obtenue sur le site http://www.genomesize.com/ (Gregory 2021). Cette taille attendue correspond à l'estimation de la quantité d'ADN par cellule. Ce résultat en picogramme (pg) est alors convertible en nombre de paires de bases avec la formule suivante :

$$Nombre\ de\ paires\ de\ bases\ = Masse\ en\ pg\ \times\ 0{,}978\ \times\ 10^{9}$$

De plus, la complétude de chaque génome a été évaluée avec le programme BUSCO (Waterhouse et al. 2018).

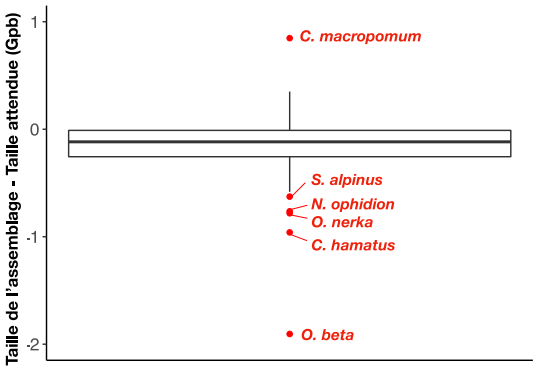
BUSCO

BUSCO (Benchmarking Universal Single-Copy Orthologs) est un logiciel qui va rechercher des gènes dits « de ménage » dans un génome assemblé. Ces gènes sont supposés être présents en une seule copie chez tous les organismes d'un groupe prédéfini, et permettent donc de donner une mesure quantitative de la complétude d'un génome. Chez les actinoptérygiens, ces gènes universels sont au nombre de 4 584.
Ce logiciel va classer ces 4 584 gènes en 3 catégories :

- Gènes trouvés complets (C), soit en copie unique (S), soit en plusieurs copies (D) ;
- Gènes trouvés fragmentés (F), c'est-à-dire que seulement une partie de ces gènes est trouvée dans le génome ;
- Gènes non trouvés (M).

Six génomes ont tout d'abord été retirés de l'analyse, car leur taille était incohérente avec la taille attendue (Figure 15).

A



B

Espèce	Taille du génome assemblé (pb)	Taille du génome attendue (pb)	Différence (pb)
Chionodraco hamatus	829,466,164	1,789,740,000	-960,273,836
Colossoma macropomum	2,314,107,020	1,467,000,000	847,107,020
Nerophis ophidion	976,361,201	1,740,840,000	-764,478,799
Oncorhynchus nerka	1,927,125,257	2,709,060,000	-781,934,743
Opsanus beta	1,028,783,780	2,934,000,000	-1,905,216,220
Salvelinus alpinus	2,169,536,488	2,797,080,000	-627,543,512

Figure 15 : Filtrage des génomes selon leur taille d'assemblage.
A- Distribution de la différence entre la taille de l'assemblage des génomes et leur taille attendue par mesure du poids de l'ADN. La différence est exprimée en giga paires de bases (Gpb). Les espèces retirées de l'analyse sont en rouge. B- Détail des tailles de génomes assemblés et des tailles attendues pour les espèces retirées de l'analyse, avec la différence entre ces deux tailles (pb : paires de bases).

Ensuite, les génomes de téléostéens pour lesquels moins de 90% de gènes BUSCO étaient retrouvés complets ont été retirés. Ce seuil a été choisi pour obtenir des estimations fiables du nombre de récepteurs olfactifs, tout en gardant un nombre de génomes assez important. Ainsi, 178 génomes de téléostéens ont été conservés (Figure 16).

Les génomes des autres actinoptérygiens n'ont pas été filtrés sur leur complétude car seulement sept génomes de non-téléostéens sont actuellement disponibles et leur position phylogénétique permet d'avoir une bien meilleure idée de l'évolution des récepteurs olfactifs à l'échelle de l'ensemble des poissons à nageoires rayonnées[14].

[14] Résultats BUSCO pour les génomes de non-téléostéens : *Polypterus senegalus* : 82% ; *Erpetoichthys calabaricus* : 81% ; *Acipenser ruthenus* : 91% ; *Polyodon spathula* : 88% ; *Amia calva* : 92% ; *Lepisosteus oculatus* : 92% ; *Atractosteus spatula* : 93%.

Au total, 185 génomes ont été conservés, dont 178 téléostéens (qui représentent 40 ordres différents parmi les 64 existants[15]) et sept espèces de non-téléostéens (qui représentent les quatre ordres existants).

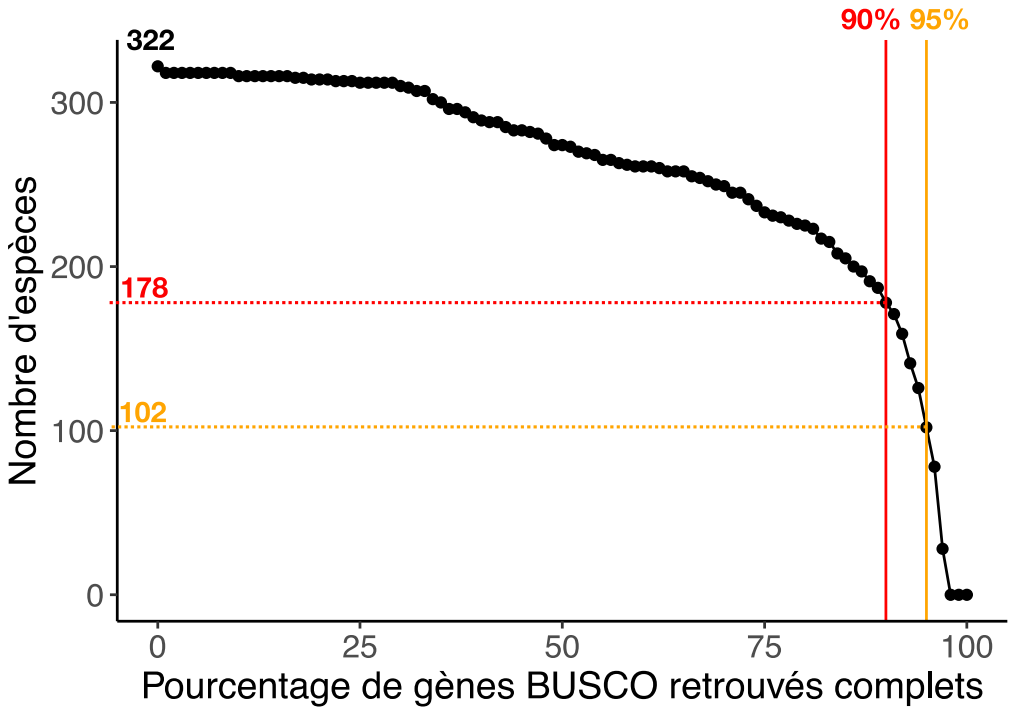


Figure 16 : Nombre de génomes en fonction d'un seuil de pourcentage de gènes BUSCO trouvés.
Parmi 322 génomes de téléostéens, 178 ont au moins 90% de gènes complets contre uniquement 102 si l'on monte ce seuil à 95%.

La phylogénie de ces 185 espèces a été obtenue sur le site https://fishtreeoflife.org/ (Rabosky et al. 2018) sur lequel se trouve la phylogénie d'actinoptérygiens la plus complète à ce jour, contenant près de 12000 espèces.

[15] Selon la base de données https://fishtreeoflife.org/

2) Méthodes de récupération des séquences de récepteurs olfactifs

Deux méthodes ont été appliquées[16], une première pour récupérer les *OR* qui sont des gènes composés d'un seul exon, et une deuxième pour récupérer les autres familles de gènes avec de multiples exons.

OR

Un *tblastn* (Altschul et al. 1990) a été réalisé avec des séquences connues d'OR (Niimura 2009) contre les génomes d'intérêt avec une e-value seuil de $1e^{-10}$, afin d'identifier les régions génomiques contenant ces gènes. Ces régions ont été extraites et étendues de 1 000 paires de bases en amont et en aval avec le programme *samtools* (Li 2011). Les phases ouvertes de lecture (= ORF) plus grandes que 750 paires de bases (taille minimum d'un gène *OR*) ont ensuite été extraites de ces régions à l'aide du programme *EMBOSS getorf* (Rice, Longden, et Bleasby 2000).

Toutes ces ORFs ont été soumises à un *blastp* contre la base de données *UniProt* (The UniProt Consortium 2021) et seules les séquences pour lesquelles le meilleur résultat correspondait à un gène *OR* connu ont été retenues. Un deuxième filtre a alors été appliqué en alignant ces séquences avec quelques gènes *OR* connus et des gènes non *OR* servant de groupe externe[17] avec le programme *MAFFT* (Katoh et Standley 2013). Un arbre phylogénétique en maximum de vraisemblance a été calculé avec *IQ-TREE* (Nguyen et al. 2015) et seules les séquences qui étaient bien groupées avec les gènes *OR* connus ont été retenues pour la suite. Lorsque deux ou plus de séquences étaient identiques à 100% au niveau nucléotidique, une seule a été conservée à l'aide du programme *CD-HIT* (Fu et al. 2012), car elles sont souvent des duplications artificielles, résultats d'un mauvais assemblage du génome.

Toutes les séquences récupérées après ces étapes ont été classées en tant qu'*OR* fonctionnels.

[16] Disponibles sur le lien suivant : https://github.com/MaximePolicarpo/Olfactory_receptor_genes.git

[17] Une séquence externe (= séquence outgroup) est une séquence plus éloignée de tous les gènes étudiés que ces gènes entre eux. Ici, les gènes outgroup utilisés sont des gènes qui possèdent sept domaines transmembranaires, de la même façon que les OR, mais qui ne sont pas des OR. On peut également parler d'espèce outgroup lorsque l'on fait une phylogénie d'espèces et que l'on racine l'arbre avec une espèce plus éloignée des espèces étudiées qu'elles ne le sont entre elles.

Ces gènes ont ensuite été utilisés pour un second *tblastn* contre les génomes, avec une e-value seuil plus contraignante de $1e^{-20}$. Pour chaque génome, les régions identifiées et contenant un gène OR fonctionnel ont été retirées ; les régions restantes correspondaient à des gènes *OR* non fonctionnels, classés en trois catégories :

- Gènes tronqués, correspondants à des gènes incomplets mais sans mutations perte de fonction apparentes dans le résultat de *blast* ;
- Pseudogènes, correspondants à des gènes qui contiennent une ou plusieurs mutations perte de fonction ;
- Gènes de bordure, correspondants à des gènes tronqués, localisés à moins de 30 paires de bases d'un début ou fin de contig, probablement résultant d'un artefact d'assemblage.

Les gènes fonctionnels, tronqués et de bordure ont été alignés avec des gènes *OR* d'actinoptérygiens connus afin de réaliser une phylogénie et de les classer dans les différentes sous-familles, selon leur position dans l'arbre (α, β, δ, ε, γ, ζ, η). Les pseudogènes, qui sont plus difficile à aligner, notamment lorsqu'il y a un décalage du cadre de lecture, ont quant à eux été classés par rapport à l'*OR* fonctionnel le plus similaire trouvé avec un *blastx*.

TAAR, ORA, OlfC

Tout comme pour les gènes *OR*, un *tblastn* avec une e-value seuil de $1e^{-10}$ a été utilisé avec des séquences connues de *TAAR*, *ORA* et *OlfC* contre les génomes d'intérêt. Les régions génomiques correspondantes à ces gènes ont alors été extraites et étendues de 5 000 paires de bases en amont et en aval avec *samtools*. Les gènes avec la meilleure e-value de *tblastn* contre chacune de ces régions ont été alignés contre ces dernières avec le programme *Exonerate* (Slater et Birney 2005). Ce programme aligne une protéine contre une séquence d'ADN en permettant la présence de mutations perte de fonction et d'introns. Les séquences prédites par ce programme ont été soumises à un *blastx* contre une base de données contenant des gènes *OR*, *TAAR*, *OlfC*, *ORA* et d'autres familles de gènes à sept domaines transmembranaires.

Les séquences pour lesquelles le meilleur résultat était bien un récepteur olfactif d'une des trois familles ont été retenues, et la prédiction réalisée par *Exonerate* vérifiée manuellement. Des

phylogénies ont été calculées afin de ne garder que les séquences groupées avec des récepteurs olfactifs connus. Si des séquences était 100% identiques au niveau nucléotidique, une seule a été conservée et les gènes ont été classés en 4 catégories (gènes fonctionnels, gènes tronqués, gènes de bordure, pseudogènes).

3) Évaluation des méthodes utilisées

Afin d'évaluer les méthodes utilisées, des comparaisons avec les études précédentes sur les récepteurs olfactifs d'actinoptérygiens ont été réalisées. Pour toutes les familles, nous retrouvons systématiquement un plus grand nombre de gènes fonctionnels que dans la littérature (Tableau 1). La seule exception concerne le nombre de gènes *ORA* chez l'épinoche *Gasterosteus aculeatus,* chez laquelle le gène *ORA1* fonctionnel dans la version du génome utilisée par l'étude de Zapilko et Korsching (2016) est maintenant sous forme de pseudogène dans la version plus récente du génome utilisé dans notre étude.

OR	Notre étude			Etudes précédentes			Référence
	F	I	P	F	I	P	
Danio rerio	162	4	8	154	1	21	Niimura 2009
Gasterosteus aculeatus	107	11	33	102	5	52	Niimura 2009
Oryzias latipes	81	3	5	68	6	24	Niimura 2009
Takifugu rubripes	100	17	12	47	39	39	Niimura 2009

TAAR	Notre étude			Etudes précédentes			Référence
	F	I	P	F	I	P	
Danio rerio	112	1	9	109	N/A	10	Hashiguchi and Nishida 2007
Gasterosteus aculeatus	50	2	15	50	N/A	15	Azzouzi et al. 2015
Oryzias latipes	39	2	4	27	N/A	7	Azzouzi et al. 2015
Takifugu rubripes	24	0	3	13	N/A	6	Hashiguchi and Nishida 2007

OlfC	Notre étude			Etudes précédentes			Référence
	F	I	P	F	I	P	
Danio rerio	54	3	2	53	1	1	Yang et al. 2019
Gasterosteus aculeatus	22	2	2	13	1	1	Yang et al. 2019
Oryzias latipes	24	0	2	16	1	2	Yang et al. 2019
Takifugu rubripes	27	0	4	16	2	3	Yang et al. 2019

ORA	Notre étude			Etudes précédentes			Référence
	F	I	P	F	I	P	
Danio rerio	7	0	0	7	N/A	N/A	Zapilko and Korsching 2016
Gasterosteus aculeatus	5	0	1	6	N/A	N/A	Zapilko and Korsching 2016
Oryzias latipes	7	0	0	7	N/A	N/A	Zapilko and Korsching 2016
Takifugu rubripes	5	0	1	5	N/A	N/A	Zapilko and Korsching 2016

Tableau 1 : Comparaison du nombre de gènes fonctionnels (F), incomplets (I) ou pseudogènes (P) dans notre étude et dans les études précédentes chez quatre espèces de téléostéens.
Comparaison réalisée sur les gènes OR, TAAR, V2R et V1R (de haut en bas). N/A indique que le nombre de gènes dans cette catégorie n'est pas rapporté dans l'étude.

La plus grande différence concerne le polyptère *Polypterus senegalus*, pour lequel une étude de 2021 estimait que son répertoire d'*OR* était composé de 245 gènes fonctionnels (Bi et al. 2021), contre 606 dans nos analyses (séquences extraites de la même version du génome dans les deux études).

Nous avons vérifié que ces séquences additionnelles étaient bien des récepteurs olfactifs en les plaçant au sein d'arbres phylogénétiques contenant les séquences des études précédentes. Ces arbres montrent que ces nouveaux gènes sont bien groupés avec les récepteurs olfactifs déjà connus, comme l'illustre l'exemple ci-dessous.

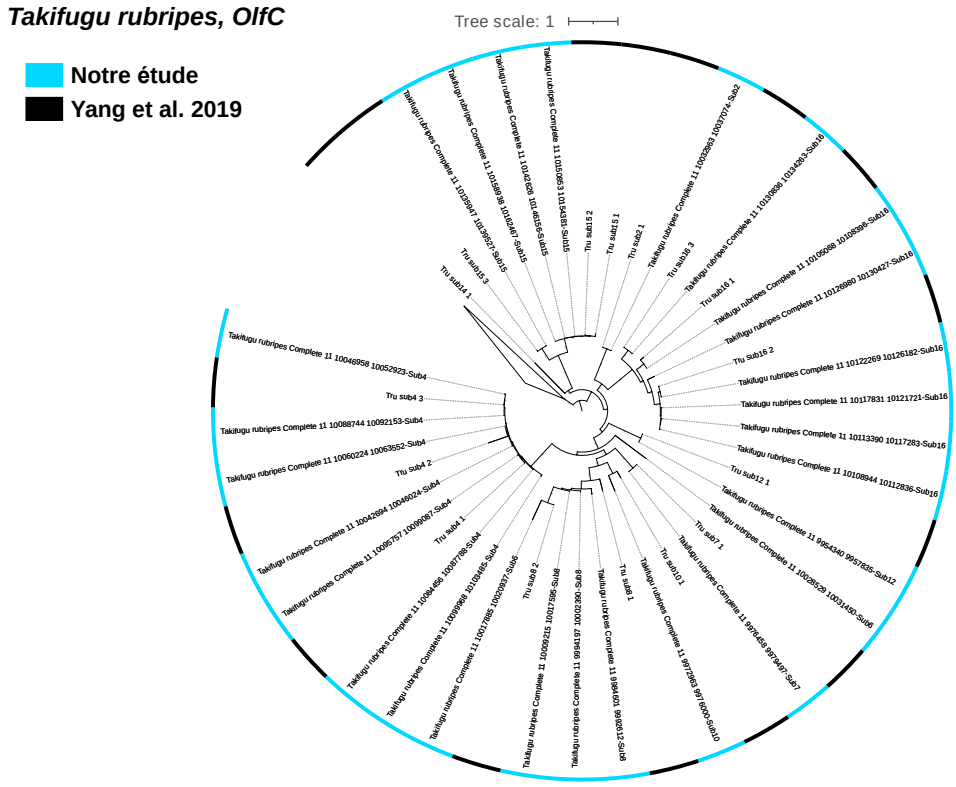


Figure 17 : Comparaison phylogénétique des gènes OlfC chez Takifugu rubripes. Arbre phylogénétique en maximum de vraisemblance réalisé avec IQ-TREE. Les séquences de gènes OlfC (=V2R) de Takifugu rubripes sont comparées entre notre étude (séquences en bleu) et l'étude de Yang et al. 2019 (séquences en noir). L'arbre est enraciné avec des gènes T2R (récepteurs gustatifs, triangle gris) et les valeurs de bootstrap sont rapportées sur chaque nœud.

B- Nombre de récepteurs olfactifs chez les actinoptérygiens

Nous avons trouvé un nombre très variable de récepteurs olfactifs chez les différentes espèces d'actinoptérygiens. La famille la plus abondante est la famille des *OR*, avec une moyenne de 126 récepteurs par poisson (Figure 18), tandis que les familles *TAAR* et *OlfC* sont un peu plus petites, avec respectivement une moyenne de 50 et 40 récepteurs par espèce. La famille *V1R* semble de son côté beaucoup moins variable, avec la majorité des espèces de téléostéens ayant 6 gènes, de façon cohérente avec la littérature.

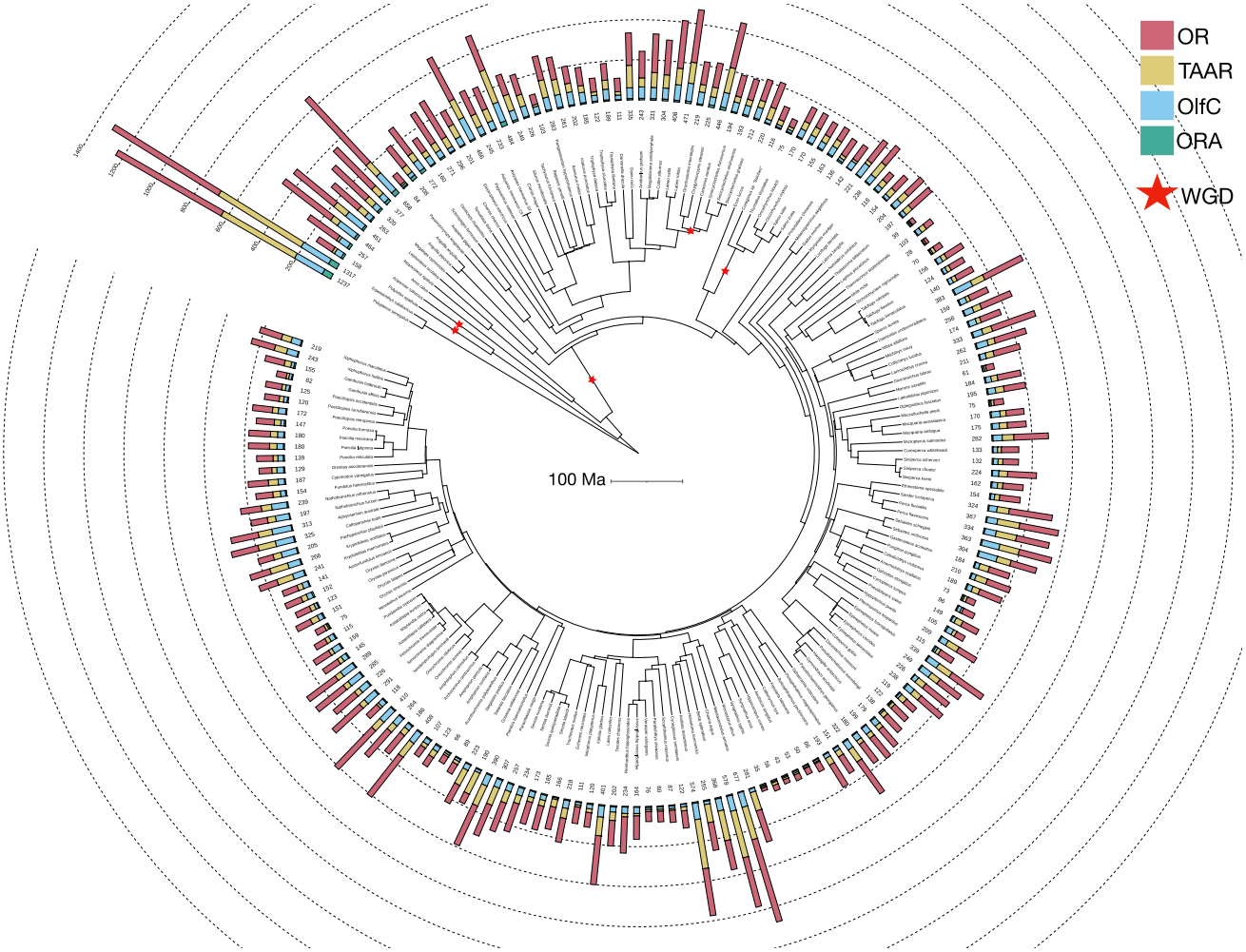


Figure 18 : Phylogénie des 185 espèces d'actinoptérygiens étudiés avec leur nombre de récepteurs olfactifs fonctionnels dans chaque famille.
Les duplications entières de génomes (WGD) sont indiquées par des étoiles rouges. Le nombre total de récepteurs olfactifs fonctionnels trouvé chez chaque espèce est indiqué au-dessus de leur nom. Figure réalisée avec iTOL

Pour toutes les grandes familles de gènes (*OR*, *TAAR* et *OlfC*), le nombre de pseudogènes est corrélé avec le nombre de gènes tronqués, indiquant que ces derniers sont probablement issus de vraies délétions génomiques et correspondent à des gènes inactifs, contrairement aux gènes de bordure pour qui aucune corrélation n'est trouvée et qui résultent d'artefacts d'assemblage (Figure 19).

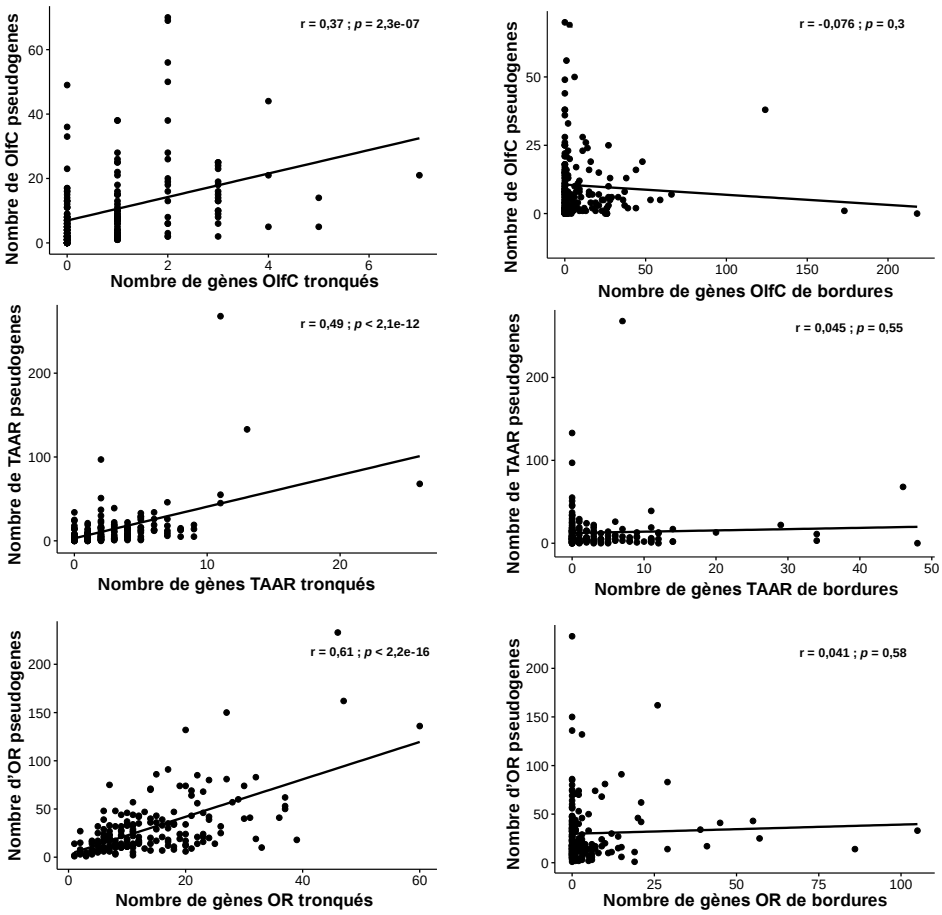


Figure 19 : Corrélations du nombre de pseudogènes et du nombre de gènes tronqués ou gènes de bordure par famille.
Corrélations de Pearson entre le nombre de pseudogènes et le nombre de gènes tronqués ou entre le nombre de pseudogènes et le nombre de gènes de bordure pour les trois grandes familles de récepteurs olfactifs.

De plus, et également pour ces trois familles, les nombres de récepteurs fonctionnels et les nombres de gènes inactifs (somme des pseudogènes et gènes tronqués) sont corrélés (tests de

Pearson, r = 0,47, r = 0,76, r = 0,2 pour les *OR*, *TAAR* et *OlfC* respectivement avec des p-values de 10^{-12}, 2×10^{-16} et 0,005). Il n'y a cependant pas de corrélation entre le nombre de gènes fonctionnels et le nombre de gènes de bordure, ces derniers ne pouvant pas être clairement attribués à une des deux principales catégories, gène fonctionnel ou pseudogène.

Par ailleurs, le nombre de gènes fonctionnels et le nombre de pseudogènes ont un signal phylogénétique fort. Afin de prendre en compte ce signal qui biaise les résultats de corrélations qu'on mesure, il faut appliquer une correction phylogénétique. Après ces corrections, les nombre de gènes fonctionnels et de gènes inactifs sont toujours corrélés pour les familles *OR* et *TAAR* (r = 0,28 et p-value = 8.57×10^{-5} pour les *OR*, r = 0,24 et p-value = 0,001 pour les *TAAR*). Cela est cohérent avec le modèle naissances-décès des gènes, qui est un processus lent de remodelage du répertoire de récepteurs olfactifs : avec un même taux de mort de gènes, une espèce avec un grand nombre de récepteurs fonctionnels aura une probabilité plus forte que ces gènes meurent (= se pseudogénisent ou disparaissent par délétions).

Signal phylogénétique

Un signal phylogénétique apparaît dans les données quand deux espèces proches tendent à se ressembler davantage que deux espèces prises de façon aléatoire dans l'arbre. Par exemple, dans notre cas, deux espèces proches auront tendance à avoir des nombres proches de gènes fonctionnels et de pseudogènes, car elles partagent une histoire évolutive commune plus longue qu'avec des espèces plus éloignées. Ainsi, les deux observations sur ces espèces proches ne seront pas des observations indépendantes. Afin de quantifier ce signal phylogénétique, plusieurs mesures existent, et la plus utilisée est le lambda de Pagel (Pagel's λ). La valeur de cet indice varie entre 0 et 1. Plus la valeur du lambda de Pagel est proche de 0, plus le signal phylogénétique est faible, c'est-à-dire que deux espèces proches ne sont pas plus similaires que deux espèces prises de façon aléatoire dans l'arbre. Au contraire, une valeur proche de 1 indique un signal phylogénétique fort.

Afin de prendre en compte ces covariances de valeurs dans l'arbre, notamment pour réaliser des régressions linéaires ou des régressions logistiques, plusieurs méthodes de correction existent et sont implémentées dans divers programmes. Dans cette étude, ces corrections seront réalisées avec le package R « caper » (Orme et al. 2013).

1) *OR*

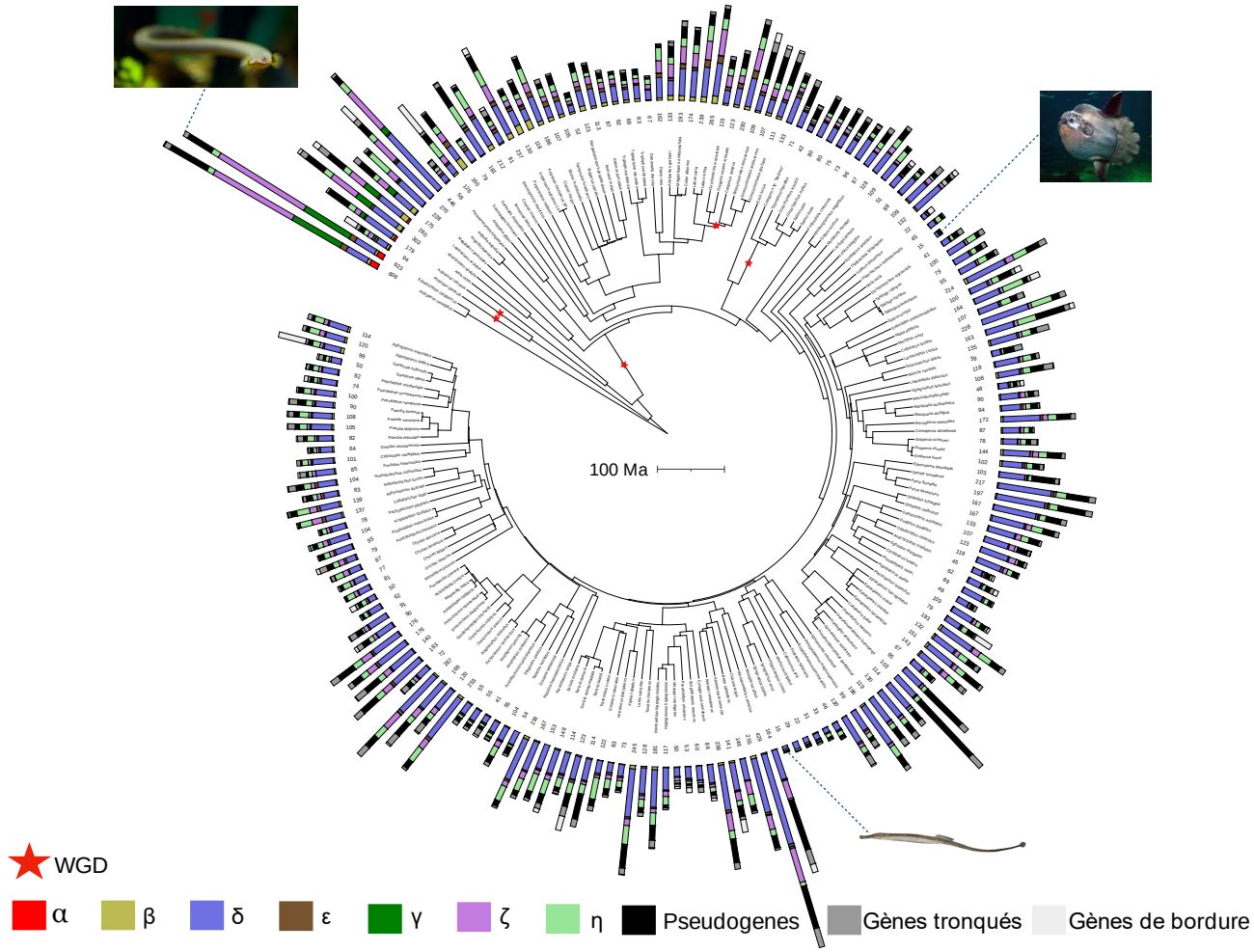


Figure 20 : Phylogénie des 185 espèces d'actinoptérygiens étudiés avec leur nombre de récepteurs olfactifs fonctionnels dans la famille OR.
Les gènes sont classés en 7 familles monophylétiques. Le nombre total d'OR fonctionnels trouvé chez chaque espèce est indiqué au-dessus de leur nom. Les duplications entières de génomes (WGD) sont indiquées par des étoiles rouges. Figure réalisée avec iTOL. Les positions phylogénétiques des espèces avec le plus (E. calabaricus) et le moins de gènes (M. mola et S. typhle) sont indiquées par des lignes pointillées reliant leur nom à leur photographie.

Le nombre maximum de gènes *OR* fonctionnels trouvés est 623 chez le polyptère *E. calabaricus* tandis que seulement 15 gènes sont présents chez le syngnathe *Syngnathus typhle* et le poisson lune *Mola mola* (Figure 20). Contrairement aux téléostéens, qui ont complètement perdu la famille α et qui n'ont plus que quelques représentants de la famille γ, les polyptères possèdent une grande proportion de ces gènes. Comme on l'a vu précédemment, ces deux familles semblent être dédiées à la détection de molécules odorantes dans l'air. Or, les polyptères, pourvus de poumons, sont capables de respirer dans l'air. Ainsi, ces espèces sont régulièrement en contact avec le milieu aérien dans lequel elles pourraient alors également utiliser leur odorat, ce qui pourrait expliquer la présence d'un grand nombre de ces récepteurs olfactifs.

Il existe également des téléostéens, qui bien que dépourvus de poumons, sont capables de respirer dans l'air, notamment certains Gobiiformes tels que l'espèce *Boleophthalmus pectinirostris*. Ceux-ci passent souvent du milieu aquatique au milieu terrestre, notamment lors de la période de reproduction, pendant laquelle les mâles creusent des trous où les femelles viennent pondre leurs œufs. Cependant, ces espèces n'ont aucun gène dans la famille α, qui a complètement été perdue chez l'ancêtre commun des téléostéens, ni dans la famille γ, disparue chez l'ancêtre commun des Gobiiformes et des Syngnathiformes.

Chez les téléostéens, deux familles sont majoritairement présentes : δ avec presque 60% des gènes et η avec presque 20% (Figure 21).

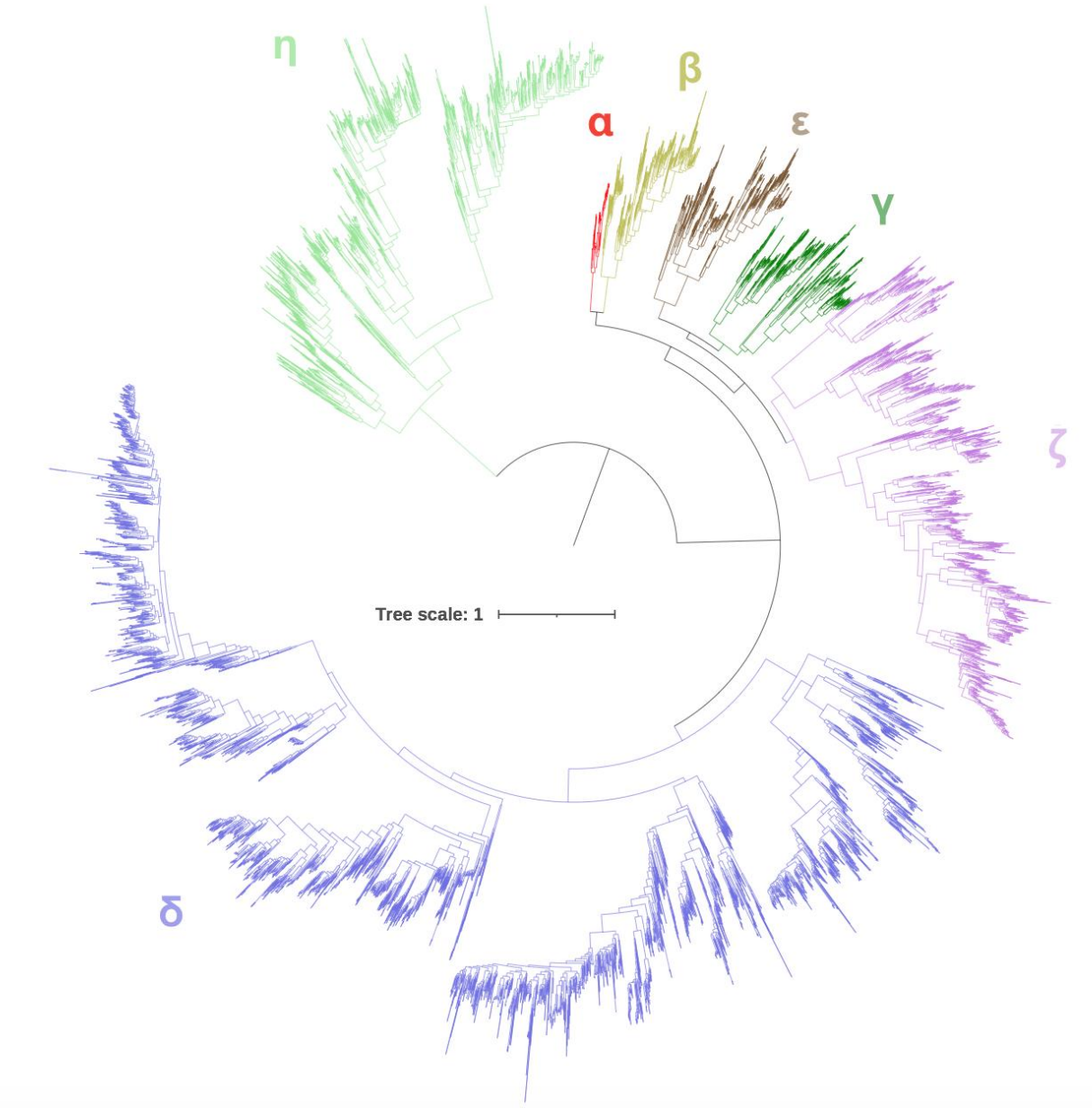


Figure 21 : Phylogénie en maximum de vraisemblance des gènes OR issus de 44 actinoptérygiens (représentant 44 ordres).
L'arbre est enraciné avec la famille η.

2) *TAAR* et *OlfC*

De la même façon que pour les gènes *OR*, *E. calabaricus* est l'espèce qui possède le plus grand nombre de *TAAR* et *OlfC* fonctionnels, avec 486 et 161 gènes respectivement. Les espèces possédant le moins de récepteurs fonctionnels sont *Callionymus lyra* avec uniquement trois gènes *TAAR* et *Mola mola* avec deux gènes *OlfC* (Figure 22 et Figure 23).
Des phylogénies en maximum de vraisemblance ont été réalisées à partir de toutes les séquences *TAAR* et *OlfC* et sont représentées respectivement dans les Figure 24 etFigure 25.

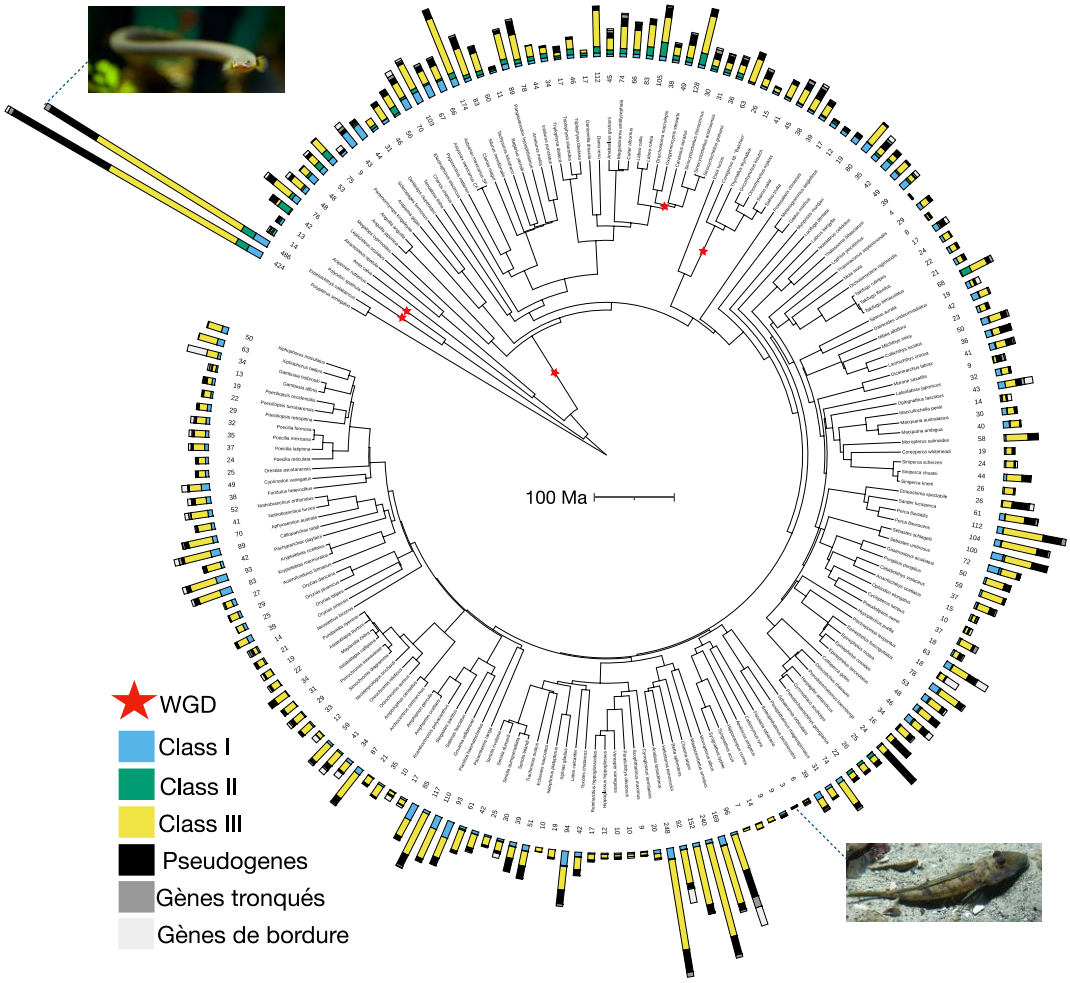


Figure 22 : Phylogénie des 185 espèces d'actinoptérygiens étudiés avec leur nombre de récepteurs olfactifs fonctionnels dans la famille TAAR.
Les gènes sont classés en 3 familles. Le nombre total de TAAR fonctionnels trouvé chez chaque espèce est indiqué au-dessus de leur nom. Les duplications entières de génomes (WGD) sont indiquées par des étoiles rouges. Figure réalisée avec iTOL. Les positions phylogénétiques des espèces avec le plus (E. calabaricus) et le moins de gènes (C. lyra) sont indiquées par des lignes pointillées reliant leur nom à leur photographie.

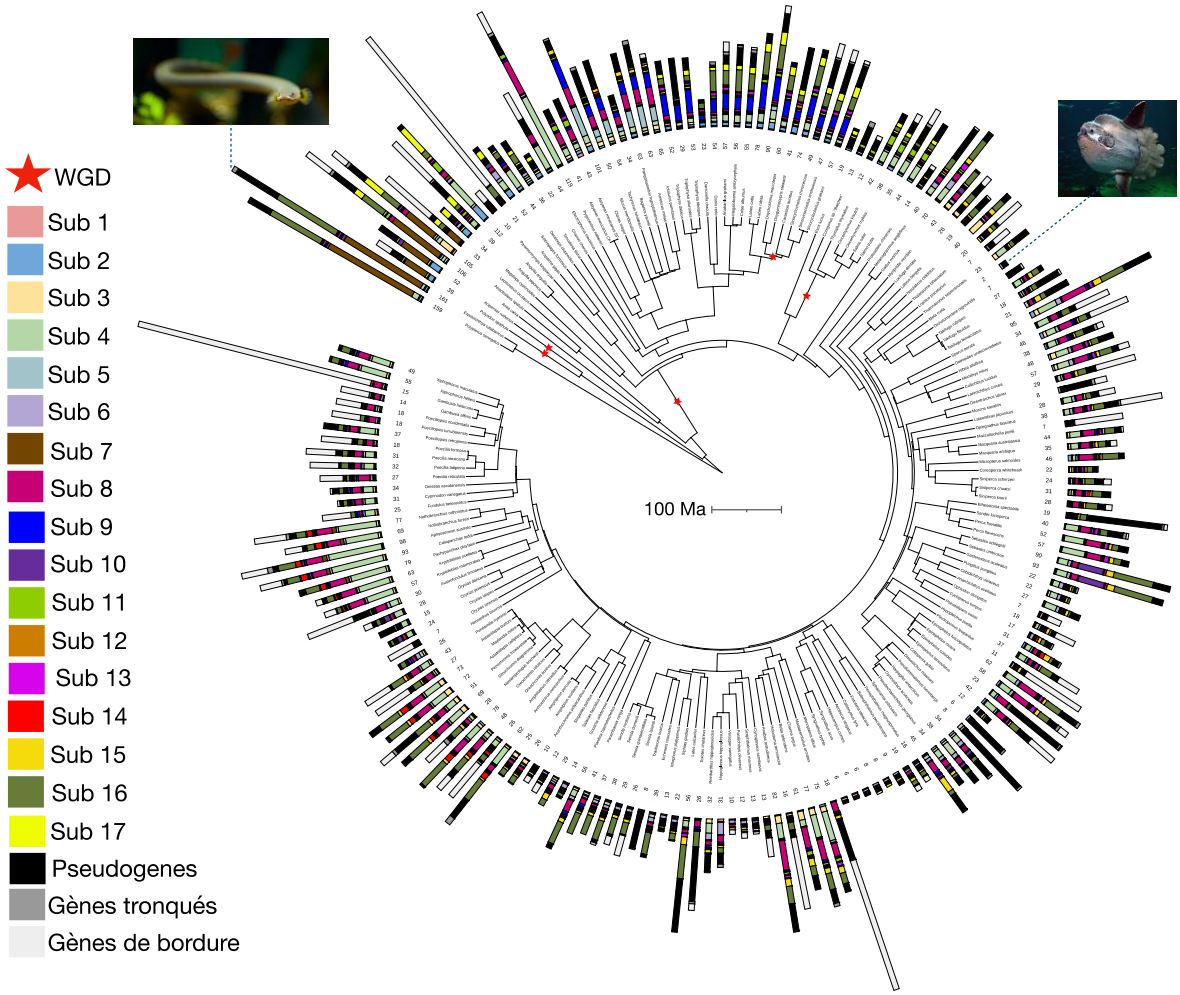


Figure 23 : Phylogénie des 185 espèces d'actinoptérygiens étudiées avec leur nombre de récepteurs olfactifs fonctionnels dans la famille OlfC.
Les gènes sont classés en 17 familles. Le nombre total d'OlfC fonctionnels trouvé chez chaque espèce est indiqué au-dessus de leur nom. Les duplications entières de génomes (WGD) sont indiquées par des étoiles rouges. Figure réalisée avec iTOL. Les positions phylogénétiques des espèces avec le plus (E. calabaricus) et le moins de gènes (M. mola) sont indiquées par des lignes pointillées reliant leur nom à leur photographie.

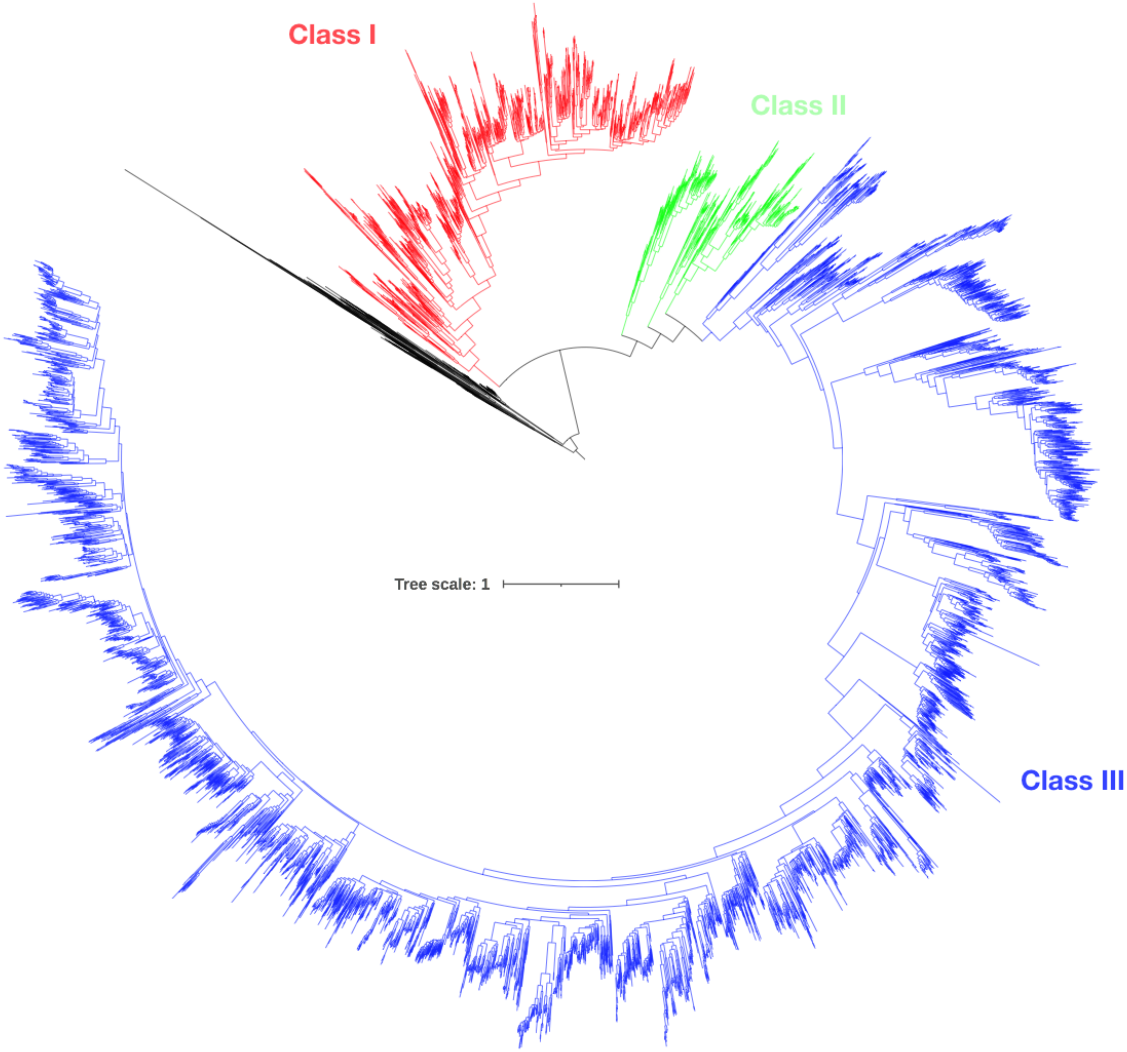


Figure 24 : Phylogénie en maximum de vraisemblance des gènes TAAR issus des 185 actinoptérygiens de l'étude.
L'arbre est enraciné avec des gènes à 7 domaines transmembranaires non TAAR. En cohérence avec les classifications faites dans la littérature, les TAAR de classe II ne forment pas un groupe monophylétique.

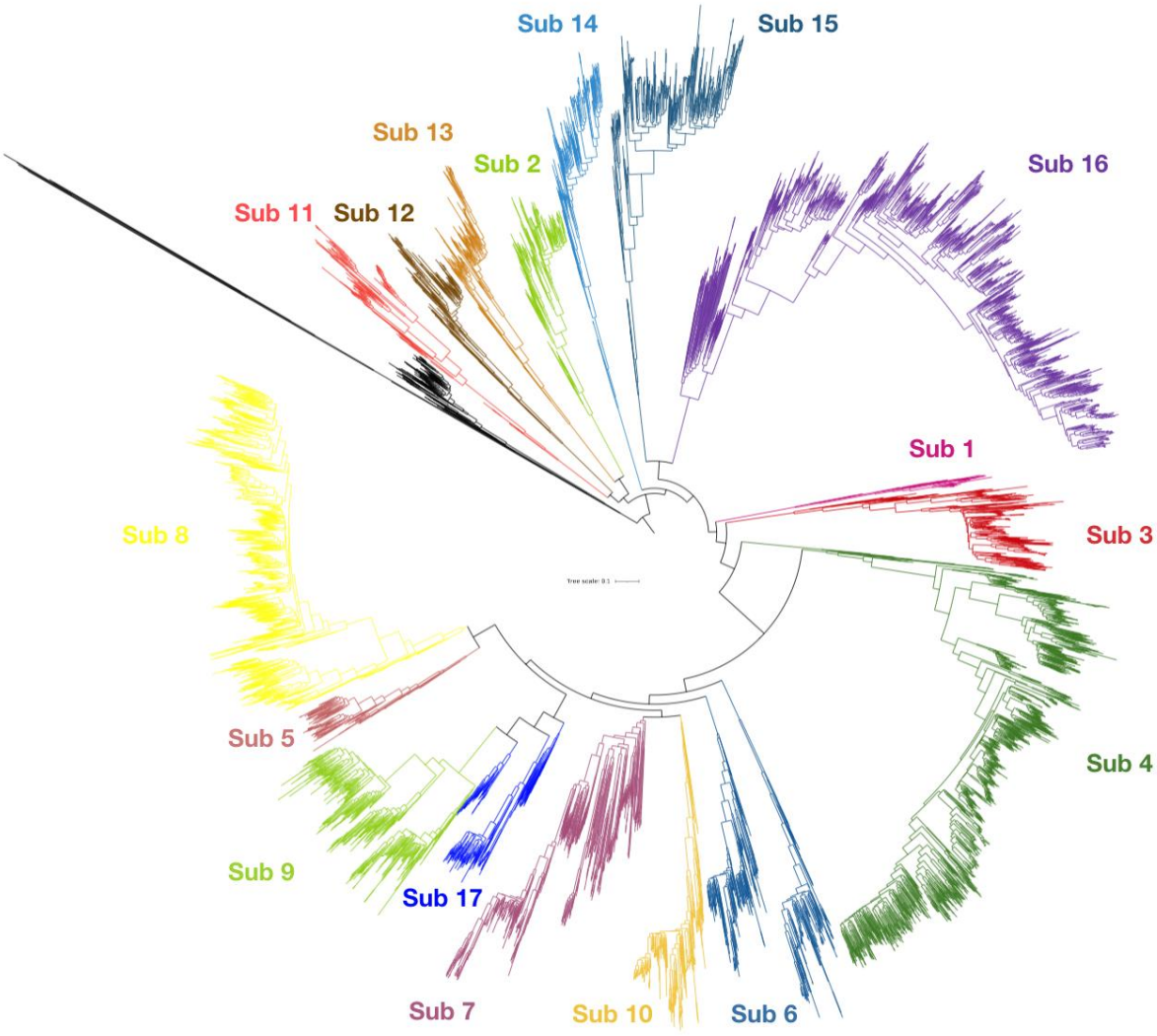


Figure 25 : Phylogénie en maximum de vraisemblance des gènes OlfC issus des 185 actinoptérygiens de l'étude.
L'arbre est enraciné avec des gènes T1R, CasR et V2R2.

3) *ORA*

Comme indiqué précédemment, la grande majorité des téléostéens possède les six gènes orthologues de la famille *ORA* : *ORA1* à *ORA6* (Figure 26). Certaines espèces ont cependant subi des pertes de quelques-uns de ces gènes. Par exemple, le *Mola mola* possède *ORA2*, *ORA3* et *ORA5* sous forme fonctionnelles, mais *ORA1*, *ORA4* et *ORA6* sous forme de pseudogènes, indiquant qu'ils ont été perdus relativement récemment. Tout comme pour la perte de la famille α des gènes *OR*, il y a eu une perte complète de la famille *ORA7* chez l'ancêtre commun des téléostéens. De plus, il est intéressant de noter que les non-téléostéens, les élopomorphes et les ostéoglossomorphes ont de nombreuses copies dans la famille *ORA8*, tandis que cette dernière a complètement été perdue avant l'ancêtre commun le plus récent des clupéocéphales (Figure 27).

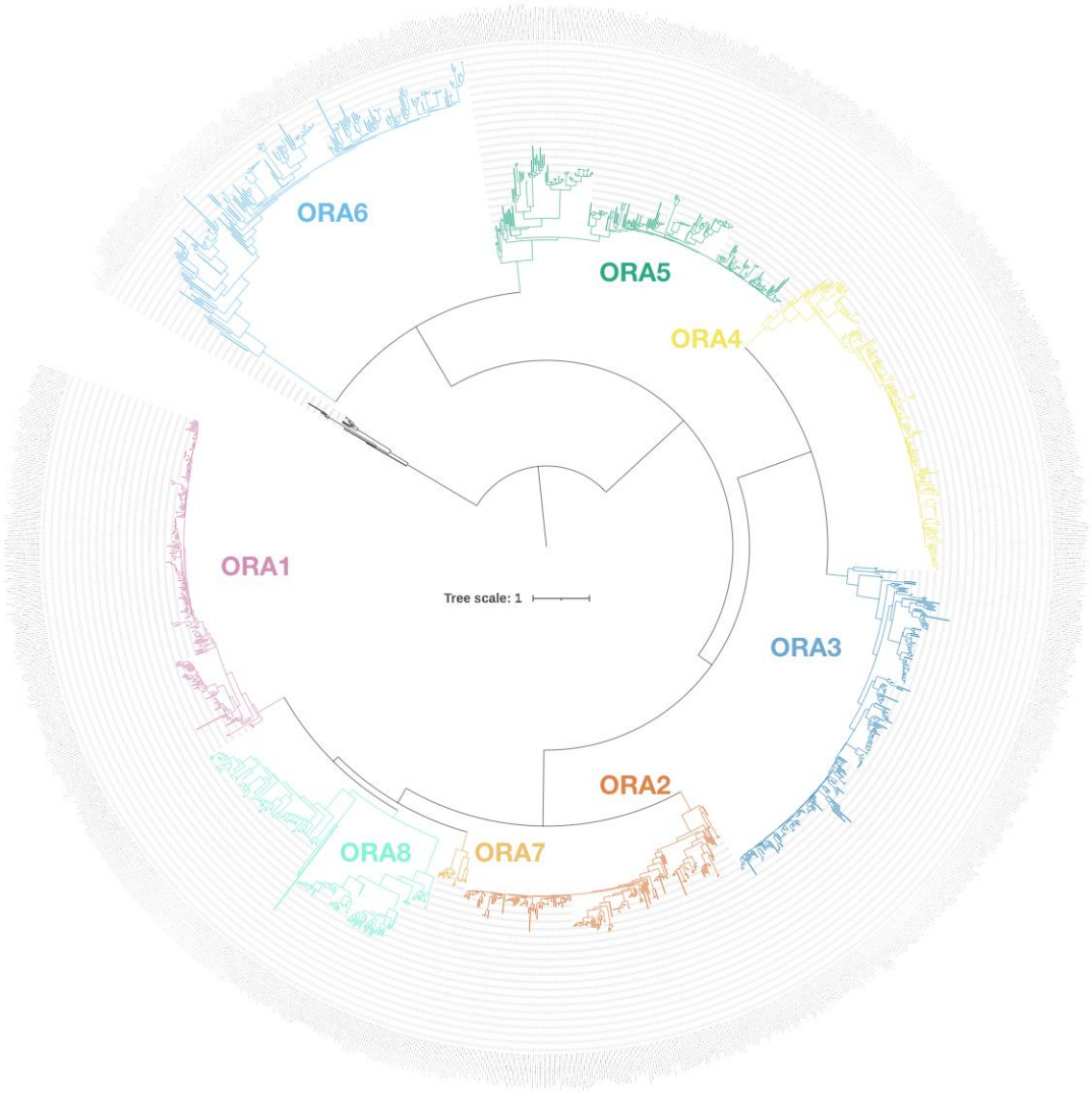


Figure 26 : Phylogénie en maximum de vraisemblance des gènes ORA issus des 185 actinoptérygiens de l'étude.
L'arbre est enraciné avec des gènes T2R.

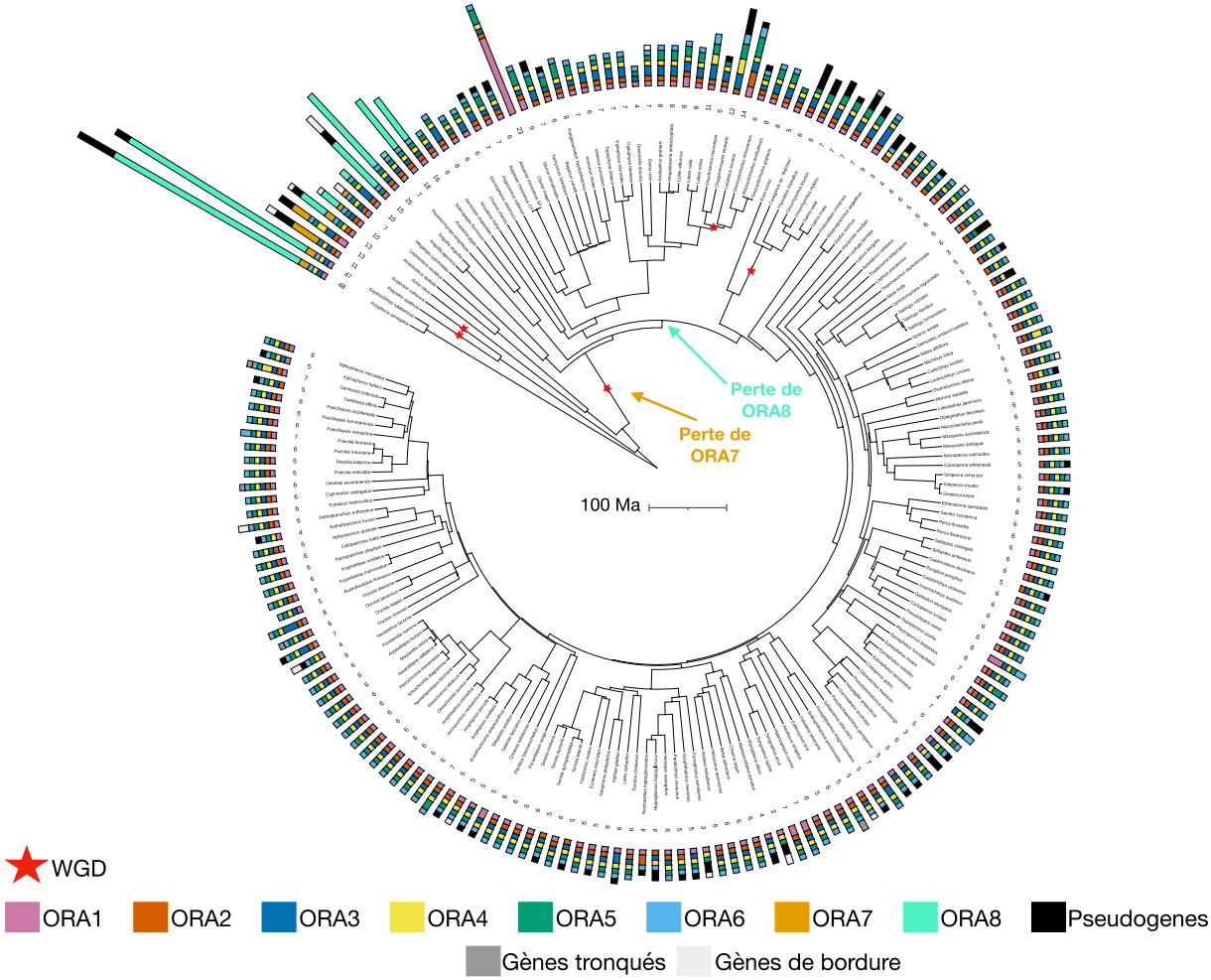


Figure 27 : Phylogénie des 185 espèces d'actinoptérygiens étudiés avec leur nombre de récepteurs olfactifs fonctionnels dans la famille ORA.
Les gènes sont classés en 8 familles. Le nombre total d'ORA fonctionnels trouvé chez chaque espèce est indiqué au-dessus de leur nom. Les duplications entières de génomes (WGD) sont indiquées par des étoiles rouges et les pertes des familles ORA7 et ORA8 sont indiquées par des flèches. Figure réalisée avec iTOL.

4) Les récepteurs olfactifs : une coévolution de trois familles de gènes

De façon surprenante, nous avons observé une forte corrélation entre les nombres de gènes fonctionnels des trois grandes familles (*OR*, *TAAR* et *OlfC*) (Figure 28). De plus, une corrélation est observée entre les nombres de pseudogènes de ces trois familles, et également entre les proportions de pseudogènes (nombre de pseudogènes / (nombre de pseudogènes + nombre de gènes fonctionnels)) (Figure 28).

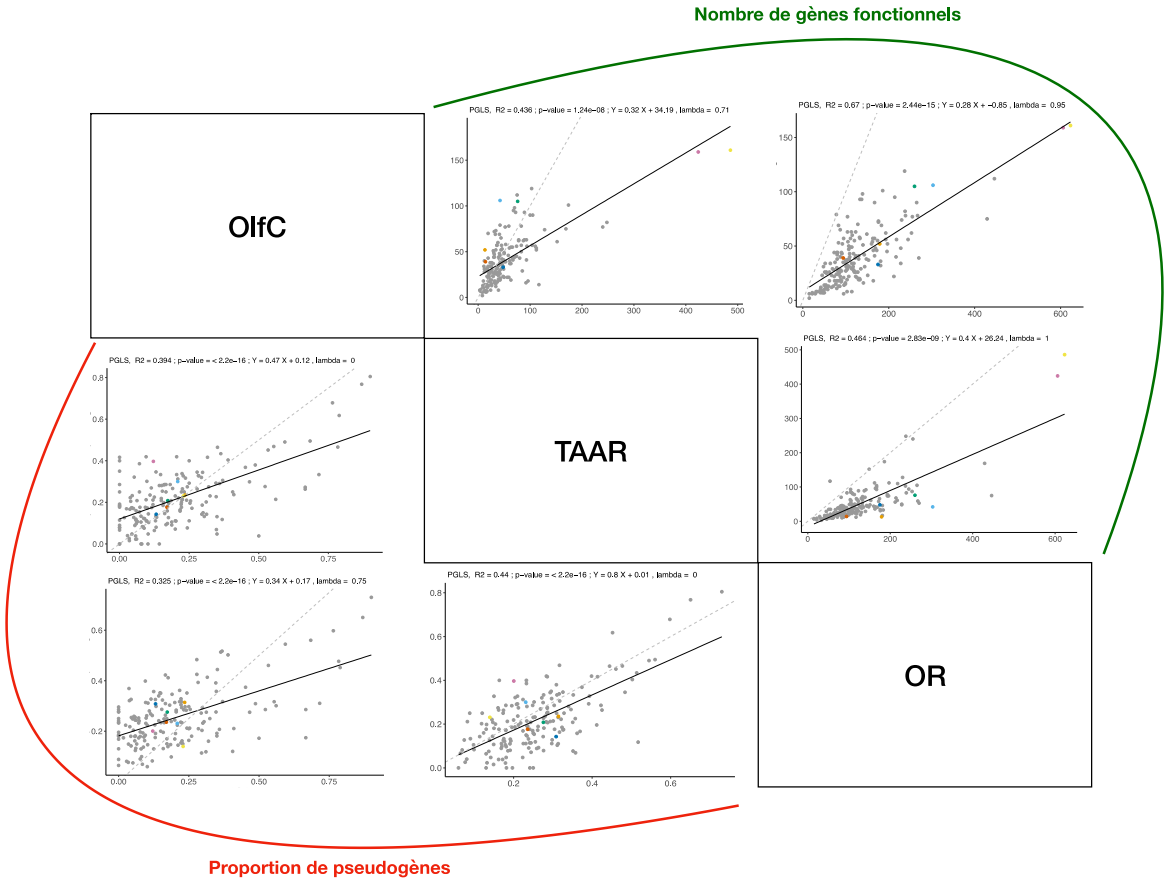


Figure 28 : Régressions linéaires phylogénétiques entre le nombre de gènes fonctionnels et la proportion de pseudogènes chez les 185 espèces d'actinoptérygiens.
Les droites pleines représentent la régression linéaire tandis que les droites pointillées représentent la droite x=y.

Ces résultats indiquent que ces familles co-évoluent, et qu'elles semblent se comporter comme une seule unité évolutive plutôt que comme trois familles indépendantes. Cela remet en question la catégorisation faite dans la littérature de ces différents gènes : les *TAAR* et *OlfC* sont considérés comme récepteurs d'acides aminés tandis que les *OR* seraient des récepteurs d'acides biliaires (Yoshihara 2014). D'après cette catégorisation, s'il y a sélection pour pouvoir détecter un plus grand nombre de molécules odorantes liées à la nourriture, une amplification du répertoire de gènes *TAAR* est attendue, mais il n'y aurait pas de raison d'observer une amplification du répertoire d'*OR*. Or d'après les résultats ci-dessus, il semblerait que ce processus de sélection ne se fasse pas par famille mais sur toutes les familles en même temps.
Ainsi, une espèce qui sera sous sélection pour détecter un grand nombre de molécules odorantes, quelles que soient leur nature, subira une amplification de toutes les familles de récepteurs olfactifs et non pas spécifiquement d'une famille.
Par ailleurs, la coévolution des familles ne supporte pas l'hypothèse que la dynamique de la taille des familles est essentiellement neutre. Si c'était le cas, on devrait observer une décorrélation de la taille des différentes familles au cours du temps. Des simulations d'évolution neutre du nombre de gènes dans les différentes familles, avec un modèle « naissance-et-mort », devront être réalisées afin d'affiner notre connaissance des attendus sous cette hypothèse.

Il est intéressant d'observer de plus près l'histoire évolutive de ces gènes dans la phylogénie des actinoptérygiens, afin de pouvoir détecter des évènements extrêmes de contraction ou d'expansion des répertoires de récepteurs olfactifs et aussi afin d'avoir une représentation générale du mode d'évolution de ces familles.

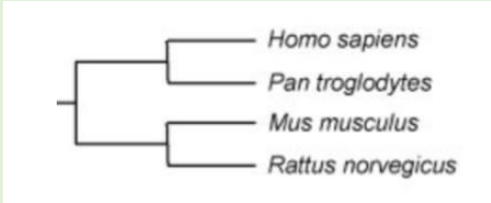
C- Histoire évolutive des récepteurs olfactifs

1) Méthode d'inférence de gains et pertes de gènes

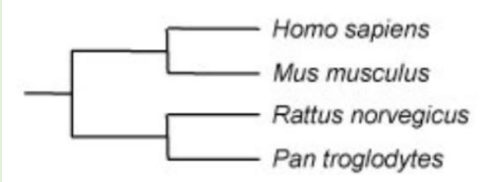
Afin de reconstruire la dynamique évolutive du nombre de récepteurs olfactifs chez les actinoptérygiens, nous avons utilisé la méthode de réconciliation des arbres des gènes avec l'arbre des espèces, permettant d'estimer le nombre de gains ou de pertes de gènes dans toutes les branches de l'arbre.

Réconciliation arbres des gènes – arbre des espèces

Soit un arbre des espèces composé de l'homme (*Homo sapiens*), du chimpanzé (*Pan troglodytes*), de la souris (*Mus musculus*) et du rat (*Rattus norvegicus*) (Penel et al. 2009) :

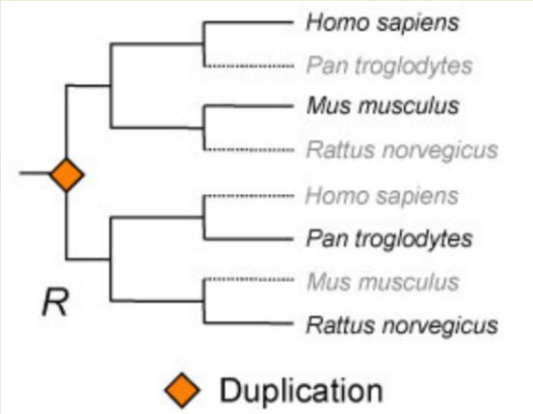


En réalisant une phylogénie avec un gène trouvé dans les génomes de ces quatre espèces, on trouve un arbre incongruent avec l'arbre des espèces :



Réconciliation arbres des gènes – arbre des espèces

L'objectif de la méthode de réconciliation est de trouver le scénario le plus parcimonieux (= impliquant le moins d'évènements possibles) qui pourrait expliquer un tel arbre de gènes, en faisant l'hypothèse que l'arbre des espèces est vrai. Ici, le scénario le plus parcimonieux implique une duplication chez l'ancêtre commun de ces quatre espèces, puis une perte indépendante d'une des deux copies du gène chez chacune d'entre elles comme suit :



Dans un cas peu complexe comme ci-dessus, avec quatre espèces et quatre gènes, il est assez simple de trouver le bon scénario sans appliquer un algorithme. Cependant, l'utilisation de logiciels tels que NOTUNG (Chen, Durand, et Farach-Colton 2000), ou TreeRecs (Comte et al. 2020) devient indispensable avec beaucoup de séquences, par exemple dans notre cas avec un arbre de 185 espèces et des dizaines de milliers de gènes.

Afin de tester la fiabilité des estimations avec cette méthode, nous avons vérifié la congruence des nombres de pertes de gènes estimés dans les branches terminales de l'arbre avec les nombres de gènes inactifs observés dans ces mêmes branches. Il est à noter que des gènes

perdus ne sont pas nécessairement trouvés sous forme inactive dans le génome, leur séquence ayant accumulée trop de mutations pour qu'il soit possible de les reconnaître, ou étant éliminée du génome par une grande délétion. Afin de quantifier le temps après lequel un gène évoluant de façon neutre n'est plus reconnu par *blast*, 10000 simulations d'évolution neutre d'une séquence de gène *OR* ont été réalisées. Avec un taux de mutation de 10^{-8} par site par génération, la moitié des gènes *OR* n'est plus reconnue après un temps d'évolution neutre de 40 millions de générations (Figure 29).

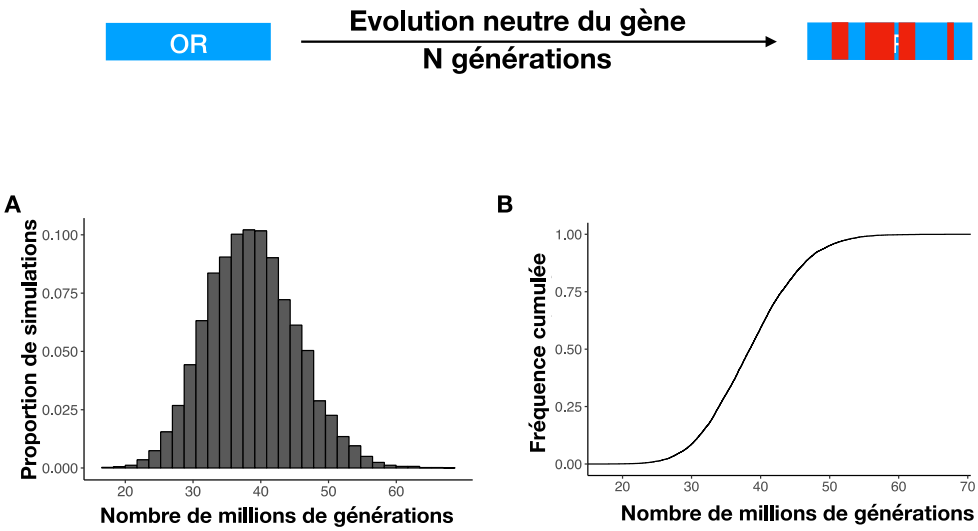


Figure 29 : Résultats des simulations d'évolution neutre d'un gène OR pendant N générations.
A- Distribution du nombre de générations après lesquelles le gène n'est plus reconnu dans le génome par tblastn B- Fréquence cumulée du nombre de générations après lesquelles le gène n'est plus reconnu dans le génome par tblastn.

En fait, ces gènes devraient ne plus être reconnus en moins de temps, car les simulations réalisées ne prennent pas en compte la possibilité de grandes délétions. Or, les gènes codant les récepteurs olfactifs étant souvent regroupés en clusters[18] dans le génome, ces délétions peuvent éliminer un ou plusieurs *OR* en même temps. Ainsi, il devrait y avoir une corrélation entre les nombres de pertes estimées et les nombre de gènes inactifs observés seulement en prenant les

[18] Gènes placés les uns à côté des autres.

branches terminales de petites tailles, dans lesquelles les pseudogènes n'ont pas eu le temps d'accumuler beaucoup de mutations et de délétions.

Ce résultat attendu est effectivement observé. En ne retenant que les branches terminales de taille inférieure à 40 millions d'années, la corrélation de Pearson entre les nombres de gènes inactifs observés et les nombres de gènes perdus inférés est forte, avec un $r > 0,5$ et une p-value $< 0,05$. En prenant progressivement des branches de plus en plus longues, la corrélation disparaît et devient rapidement inférieure à 0,2 (Figure ci-dessous).

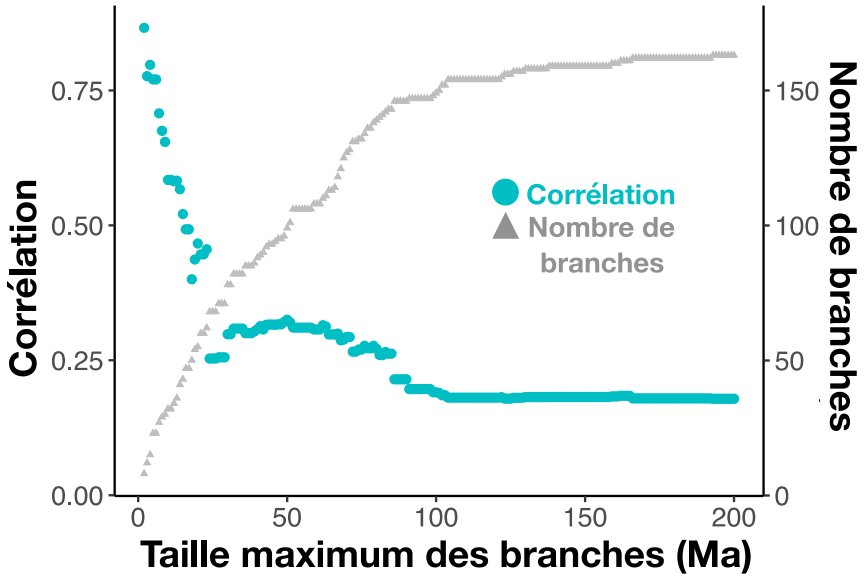


Figure 30 : Variation du coefficient de corrélation de Pearson, entre le nombre de gènes perdus estimé par la méthode de réconciliation de l'arbre de gènes et l'arbre des espèces, et le nombre de gènes inactifs observé.
Le coefficient de corrélation en fonction de la taille maximum des branches échantillonnées est représenté par des ronds verts. Les triangles gris indiquent le nombre de branches échantillonnées.

Ce résultat indique que la méthode de réconciliation entre arbre de gènes et arbre des espèces semble fiable, avec une bonne corrélation entre ces deux mesures indépendantes du nombre de gènes perdus.

2) Taux de gains et de pertes de gènes dans l'arbre des espèces

Les nombres absolus de gains et de pertes estimées sur chaque branche de l'arbre des espèces sont peu informatifs, notamment parce que ces branches ont des tailles très variables. En effet, si 10 pertes sont estimées sur une branche courte d'un million d'années, cela correspond à un processus de mort des gènes beaucoup plus rapide que si ces 10 pertes sont estimées sur une branche de 20 millions d'années.

Ces nombres absolus ont donc été normalisés pour prendre en compte la longueur des branches et obtenir des taux de gain et de perte, exprimés en gènes par million d'années. Pour cela, nous avons utilisé les équations de l'étude de Niimura, Matsui, et Touhara (2014), avec β = taux de naissance et δ = taux de perte, en faisant l'hypothèse que ces taux sont constants le long d'une branche. Au temps initial de la branche (t =0), le nombre de gènes est noté A_0. Au bout de la branche, après un temps t (t = T = taille de la branche), le nombre de gènes sera égal à $A_0 + G - L$ avec $G =$ nombre de gènes gagnés dans la branche et L = nombre de gènes perdus dans la branche. Les taux sont calculés comme suit :

$$\beta = \frac{G}{(G-L)T} \times \ln\left(1 + \frac{G-L}{A_0}\right)$$

$$\delta = \frac{L}{(G-L)T} \times \ln\left(1 + \frac{G-L}{A_0}\right)$$

Dans les rares cas ou $G = L$, ces équations ne sont plus applicables (division par zéro), et les taux de gains et de pertes seront calculés comme suit (Vieira, Sánchez-Gracia, et Rozas 2007; Niimura, Matsui, et Touhara 2014) :

$$\beta = \frac{G}{(T\ x\ A_0)}$$

$$\delta = \frac{L}{(T\ x\ A_0)}$$

Les taux moyens de naissance dans les familles *OR*, *TAAR* et *OlfC* sont égaux à 0,0071, 0,0079 et 0,007 par gène par million d'année respectivement. Les taux de perte sont à peu près égaux à ces taux de naissance : respectivement 0,0071, 0,01 et 0,006 par gène par million d'année (Figure 31). Ainsi, bien que plus petite que la famille *OR*, la famille *TAAR* semble être la famille de récepteurs olfactifs la plus dynamique, que ce soit en termes de gains ou de pertes.

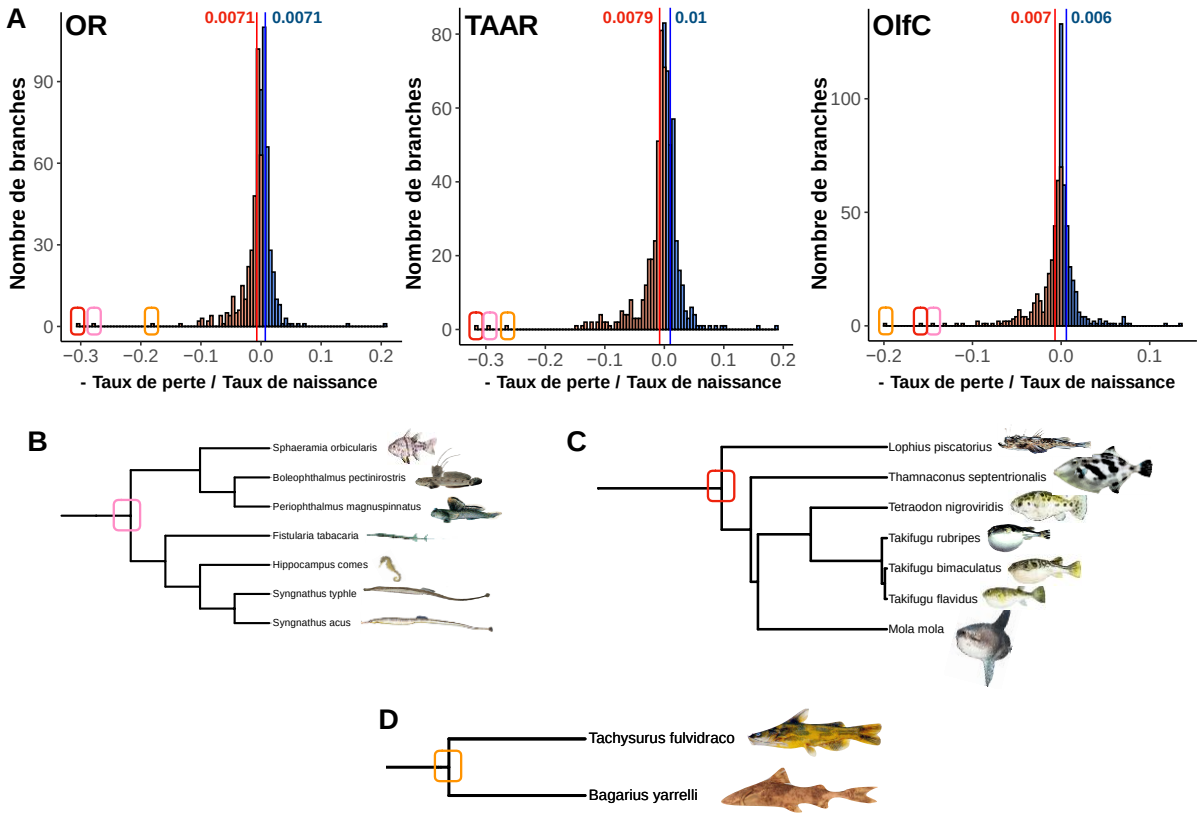


Figure 31 : Taux de naissance et de perte des récepteurs olfactifs estimées sur toutes les branches de l'arbre des espèces de l'étude (368 branches)
A- Distribution des taux de perte et taux de naissance pour les familles OR, TAAR et OlfC. Les droites verticales rouges et bleues représentent les taux moyens de perte et de naissance dans l'arbre respectivement. Les trois taux de mort les plus importants pour chaque famille sont encadrés et leur correspondance dans l'arbre est indiquée dans les figures B, C et D.

Le taux de perte des gènes *OR* le plus élevé, égal à 0,3, se situe chez l'ancêtre commun le plus récent des Tétraodontiformes et des Lophiiformes (ordre représenté par l'espèce *Lophius piscatorius*). De façon remarquable, le taux de perte le plus élevé des gènes *TAAR* et le deuxième taux de perte le plus haut des gènes *OlfC* se situent également au niveau de cet ancêtre commun.

De la même façon, un taux de perte très élevé dans les trois familles est retrouvé chez le MRCA des Kurtiformes (ordre représenté par l'espèce *Sphaeramia orbicularis*) et des Syngnathiformes mais également chez le MRCA de *Bagarius yarrelli* et *Tachysurus fulvidraco*.

Les nombres de naissances de gènes des différentes familles de récepteurs olfactifs sont fortement corrélés sur l'ensemble de l'arbre des espèces, tout comme les nombres de pertes (Figure 32). Cette observation confirme notre hypothèse de coévolution de ces différentes familles à la manière d'une seule unité évolutive.

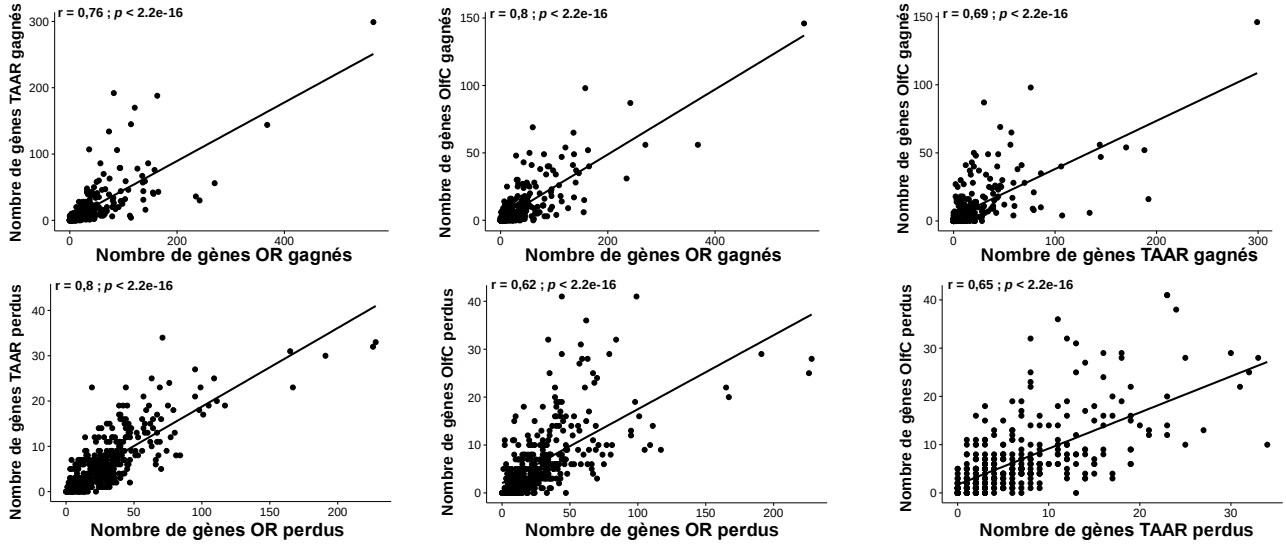


Figure 32 : Corrélations de Pearson entre le nombre de gains ou nombre de pertes de gènes des familles OR, TAAR et OlfC sur tout l'arbre phylogénétique.
Les analyses incluent les branches terminales et internes, pour un total de 368 branches.

Afin de mieux comprendre nos résultats d'analyses génomiques, nous avons cherché à identifier si des variations de facteurs environnementaux, comportementaux ou morphologiques sont liées aux variations de la taille du répertoire olfactif chez les actinoptérygiens.

D- Influence de facteurs environnementaux et morphologiques sur le répertoire de gènes olfactifs

Afin de tester si le nombre de récepteurs olfactifs des poissons pouvait être lié à leur milieu de vie, 97 paramètres écologiques de ces espèces ont été extraits de la base de données fishbase (Froese et Pauly 2021 ; https://www.fishbase.se/search.php).
Pour chaque paramètre, il a été testé s'il y avait une corrélation avec le nombre de gènes dans chaque grande famille (*OR*, *TAAR* et *OlfC*) ou le nombre de récepteurs olfactifs total (*OR* + *TAAR* + *OlfC* + *ORA*). Des régressions linéaires phylogénétiques ont été calculées pour les traits quantitatifs et des régressions logistiques phylogénétiques ont été calculées pour les traits binaires. Parmi les paramètres écologiques testés, nous trouvons par exemple la salinité de l'eau (eau salée ou eau douce), la profondeur de vie moyenne, la température et le pH de l'eau ou encore le niveau trophique des espèces, c'est-à-dire leur position dans la chaine alimentaire. Pour prendre en compte la multiplicité des tests réalisés, la méthode FDR (false discovery rate) a été utilisée (Benjamini et Hochberg 1995).

Parmi tous les paramètres écologiques testés, un seul est significativement corrélé avec les nombres de récepteurs olfactifs dans chaque famille ou avec le nombre total de récepteurs : les espèces vivant dans les cours d'eau et les courants océaniques (d'après le glossaire fishbase, appelé « stream » en anglais) possèdent significativement plus de récepteurs olfactifs que les autres espèces (R2 = 0,0703, FDR p-value = 0,018) (Figure 33). En appliquant une méthode de correction statistique plus conservative, la méthode de BonFerroni (Bonferroni 1936), ce résultat n'est cependant plus significatif.

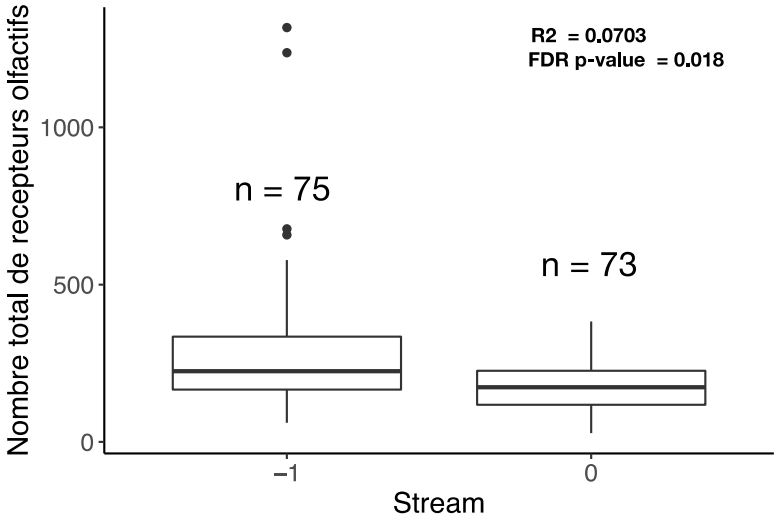


Figure 33 : Corrélation entre le nombre total de récepteurs olfactifs et la vie en cours d'eau (-1) ou non (0).
Le nombre d'espèces dans les deux catégories sont indiquées (n), ainsi que les valeurs du R2 et de la p-value corrigée par la méthode FDR. Régression logistique calculée avec le package R « caper ».

Le même type d'analyse a été réalisé avec des caractéristiques morphologiques adultes des espèces étudiées telles que la taille et le poids moyen de l'espèce ou encore la taille relative de l'œil. Aucune de ces caractéristiques n'est corrélée avec le nombre de récepteurs olfactifs. Cependant, la base de données fishbase ne contient pas de données morphologiques directement reliées à l'odorat, telles que la taille relative du bulbe olfactif ou de l'épithélium olfactif. Or, il est connu, chez les primates, que le nombre de gènes OR est corrélé avec la forme du nez (nez chez Haplorrhiniens contre truffe chez les Strepsirrhiniens) et avec la taille relative du bulbe olfactif (Niimura, Matsui, et Touhara 2018).

E- Évolution de l'épithélium olfactif des actinoptérygiens

Comme mentionné dans l'introduction, l'épithélium olfactif des actinoptérygiens a souvent une forme de rosette composée de plusieurs lamelles, mais peut également ne posséder qu'un très petit nombre de lamelles voire aucune lamelle.
Nous avons fait l'hypothèse que le nombre de lamelles est un bon indicateur de l'investissement relatif d'une espèce dans l'olfaction. Ainsi, un poisson équipé d'un épithélium composé de beaucoup de lamelles devrait avoir un nombre plus grand de neurones, permettant d'exprimer un nombre plus grand de récepteurs olfactifs (d'après la règle 1 neurone – 1 récepteur). Cependant, comme il n'existe pas encore de base de données regroupant cette information pour les actinoptérygiens, nous avons réalisé un travail bibliographique afin de documenter le nombre de lamelles chez le plus grand nombre possible d'espèces appartenant à ce groupe.

1) Récupération de données dans la littérature

Nous avons divisé, sur la base de travaux précédents (Hansen et Zielinski 2005), les épithéliums olfactifs en deux catégories : les épithélium à plusieurs lamelles (appelés ML, avec plus de 2 lamelles et communément appelés rosettes dans la littérature) et les épithéliums avec très peu ou pas de lamelles (appelés non-ML, avec 2 ou moins de lamelles) (Figure 34).

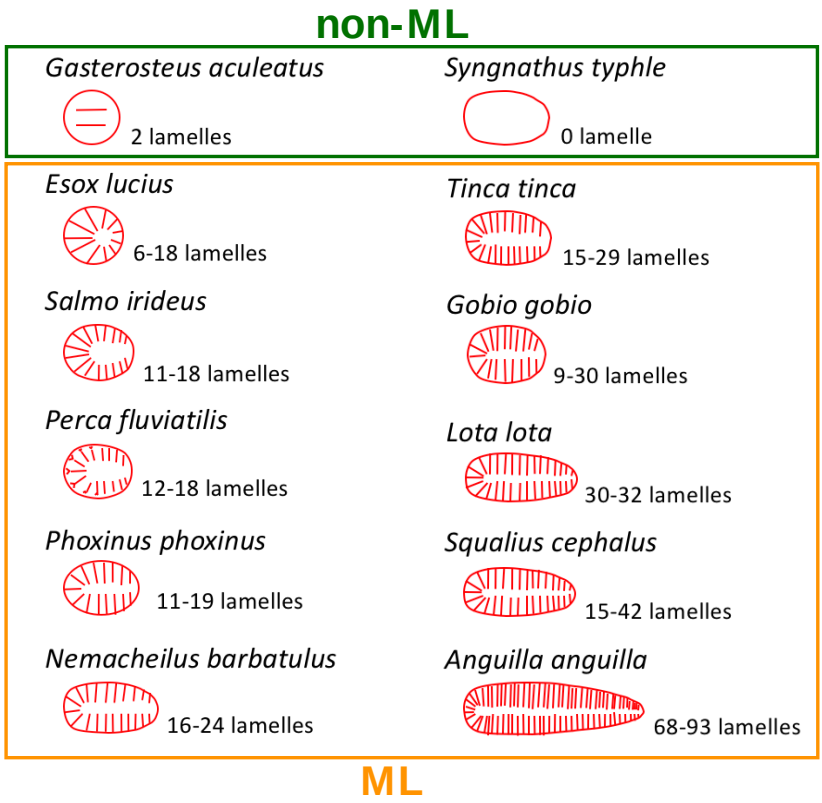


Figure 34 : Catégories utilisées pour définir les épithéliums olfactifs et exemples du nombre de lamelles chez quelques téléostéens.

Cadre vert : Épithélium non-ML (0 à 2 lamelles). Cadre orange : Épithélium ML (3 ou plus de lamelle).

Afin de connaître le type (ML ou non-ML) et le nombre de lamelles chez différentes espèces d'actinoptérygiens, plusieurs méthodes de recherche dans la littérature ont été utilisées : recherche à partir de mots clefs (par exemple « Olfactory epithelium » + « teleost » ou « Actinopterygii » ou encore le nom de différentes espèces) sur Google, PubMed et Google Scholar. Le plus souvent, lorsqu'une publication contenant une information sur l'épithélium olfactif d'une ou plusieurs espèces était trouvée, une recherche des autres articles publiés par les auteurs ainsi que de la littérature citée dans la publication a été réalisée. Ces recherches ont également été faites avec les mêmes mots clefs en français ou en allemand.

Un total de 138 articles, revues, livres et thèses contenant des informations sur l'épithélium olfactif d'un ou plusieurs actinoptérygiens ont été trouvés. L'article le plus ancien date de 1909

(Burne 1909) tandis que le plus récent date de 2021 (Dymek, Kuciel, et Żuwała 2021). Dans de nombreux articles, il est indiqué que l'épithélium olfactif est en forme de rosette (composé de plusieurs lamelles) mais sans précision sur le nombre de lamelles. Ainsi, nous avons récupéré une information sur la forme de l'épithélium pour 423 espèces mais un nombre de lamelles pour 339 d'entre elles.

Il est difficile d'évaluer la variabilité du nombre de lamelles chez les individus d'une espèce. Nous avons des données pour seulement un individu pour la grande majorité des espèces (249), tandis que pour les autres, les données sont souvent limitées à deux ou trois individus. Ces données suggèrent qu'il y a une faible variabilité du nombre de lamelles entre individus adultes d'une même espèce, mais que le nombre de lamelles augmente avec l'âge et la taille du poisson. Pour compléter ces données, nous avons établi une collaboration avec Katherine E. Bemis et James C. Tyler de la Smithsonian Institution de Washington, afin de disséquer et compter le nombre de lamelles d'espèces d'intérêt pour notre étude, notamment celles avec très peu de récepteurs olfactifs telles que le poisson lune *Mola mola* ou l'hippocampe *Hippocampus comes*. Cette collaboration nous a permis de connaitre le nombre de lamelles de 14 espèces supplémentaires.

Ainsi, nous connaissons la forme globale de l'épithélium olfactif (ML ou non-ML) pour 437 espèces dont 353 pour lesquelles nous connaissons également le nombre moyen de lamelles. Afin de mieux comprendre l'évolution de cet épithélium, une phylogénie de 380 de ces espèces a été réalisée, avec 363 de ces dernières présentes dans la phylogénie de https://fishtreeoflife.org/ et 17 pour lesquelles la position phylogénétique a été inférée grâce à la présence d'une espèce du même genre. Les autres espèces ne sont pas incluses car absentes de cette base de données et sans espèces du même genre présente.

2) Reconstruction de la forme ancestrale de l'épithélium olfactif

Une reconstruction de l'histoire évolutive de la forme de l'épithélium olfactif a été réalisée avec le programme PastML (Ishikawa et al. 2019). Ce programme permet d'estimer l'état d'un

caractère dans toutes les branches internes de la phylogénie, soit par maximum de vraisemblance, soit par parcimonie.

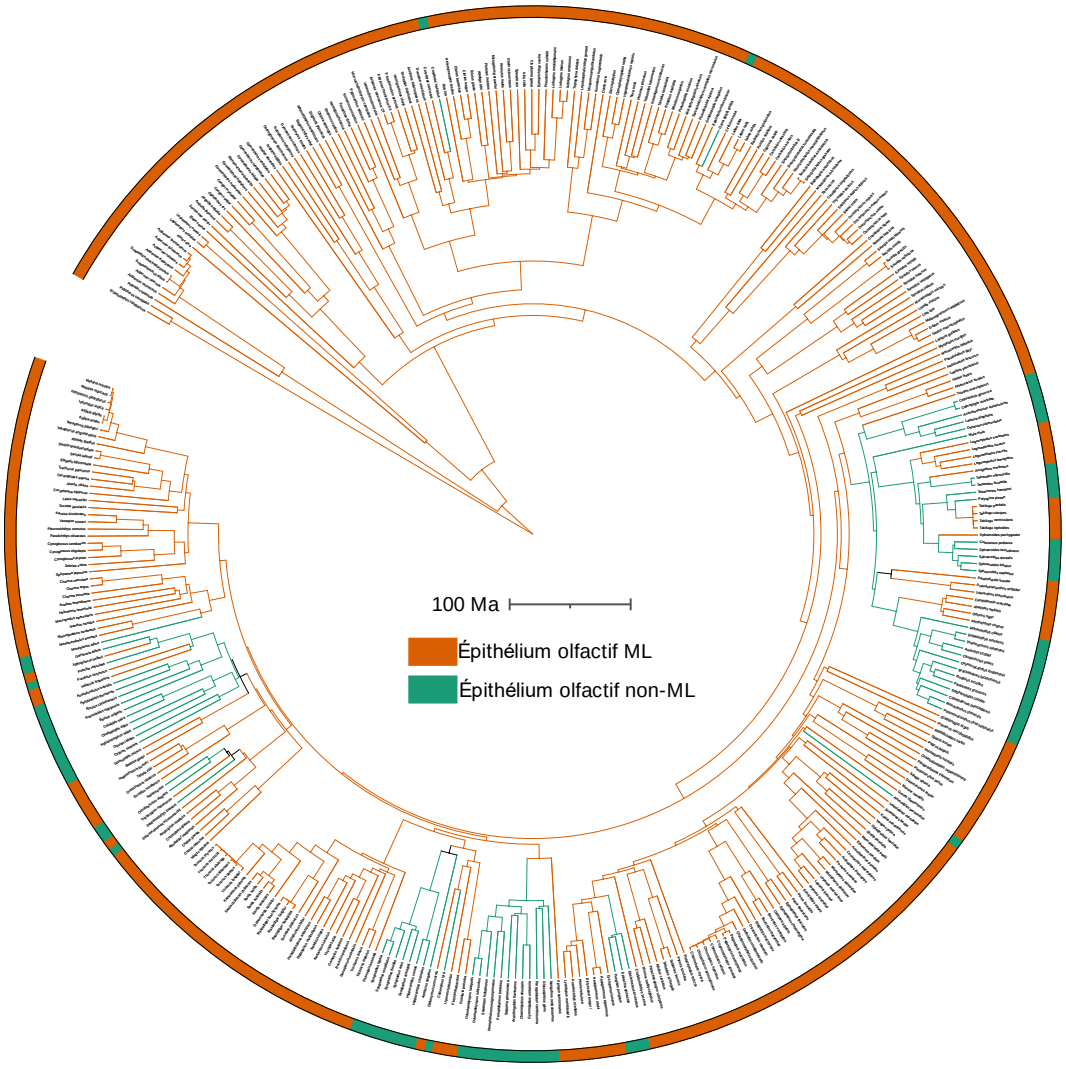


Figure 35 : Reconstruction de l'état ancestral de l'épithélium olfactif. Phylogénie basée sur fishtreeoflife.org avec les espèces pour lesquelles l'information sur la forme de l'épithélium olfactif (ML ou non-ML) est disponible. La forme de l'épithélium pour les branches internes a été inférée avec PastML. La méthode de reconstruction par maximum de vraisemblance (MPPA + EFT = « marginal posterior probabilities approximation + estimate-from-tips ») donne le même résultat que la reconstruction par parcimonie. Branches vertes = épithélium non-ML, Branches oranges = épithélium ML, Branches grises = État ambigu, avec une probabilité égale des deux caractères.

La reconstruction de la forme de l'épithélium olfactif montre que ce dernier est composé de plusieurs lamelles dans l'ancêtre commun des actinoptérygiens et également dans la grande majorité des branches internes et externes de la phylogénie (Figure 35). De nombreuses transitions d'épithélium ML à non-ML ont eu lieu, notamment chez les tétraodontiformes, les gobiiformes et les syngnathiformes. Les transitions dans le sens inverse semblent plus rares. Un gain de lamelles est notamment arrivé dans le MRCA du genre Takifugu (tétraodontiformes) (Figure 35). Cette transition inférée est supportée par les observations faites par Katherine E. Bemis chez l'espèce *Takifugu rubripes*. Effectivement, l'épithélium ML de cette dernière n'a pas une forme de rosette, mais est composé de lamelles disposées de façon parallèles (Figure 36 - A). Il sera donc intéressant d'observer l'épithélium des autres espèces ayant subie une telle transition afin de voir si l'organisation des lamelles est également différente.

Il faut noter que les épithéliums olfactifs des polyptères sont également différents de la plupart des autres espèces, avec un gain de complexité dans ce clade. Ils sont équipés d'un épithélium olfactif composé de 5 sections organisées de façon circulaire et d'une section supplémentaire placée à côté (Pfeiffer 1968; Theisen 1970). Chacune de ces sections est composée de plusieurs lamelles (Figure 36 - B), ce qui en fait les espèces avec le plus de lamelles parmi les actinoptérygiens, avec 300 lamelles chez *Polypterus senegalus* et 150 lamelles chez *Erpetoichthys calabaricus*.

A B

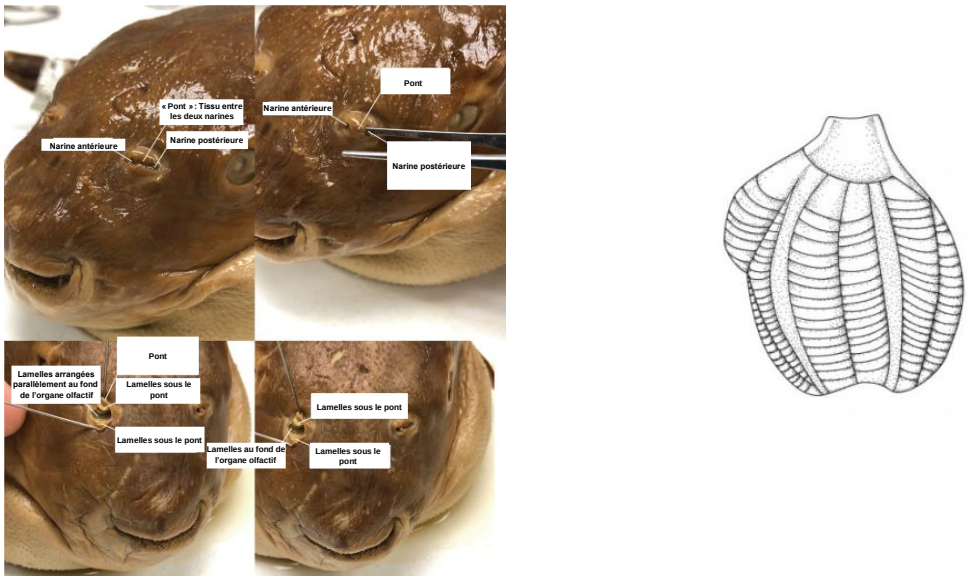


Figure 36 : Épithélium à plusieurs lamelles sans forme de rosette.
A- Photographies de l'appareil olfactif de Takifugu rubripes B- Schéma des cinq secteurs organisés de façon circulaire + un diverticule (à gauche sur le dessin) constituant l'épithélium olfactif de Polypterus senegalus. Photographies et schéma réalisés par Katherine E. Bemis.

3) Corrélation de la richesse du répertoire olfactif avec la complexité morphologique de l'épithélium olfactif

Parmi les 353 espèces pour lesquelles le nombre de lamelles est connu, 72 ont des génomes de très bonne qualité pour lesquels nous avons déterminé le nombre de gènes codant des récepteurs olfactifs (Figure ci-dessous).

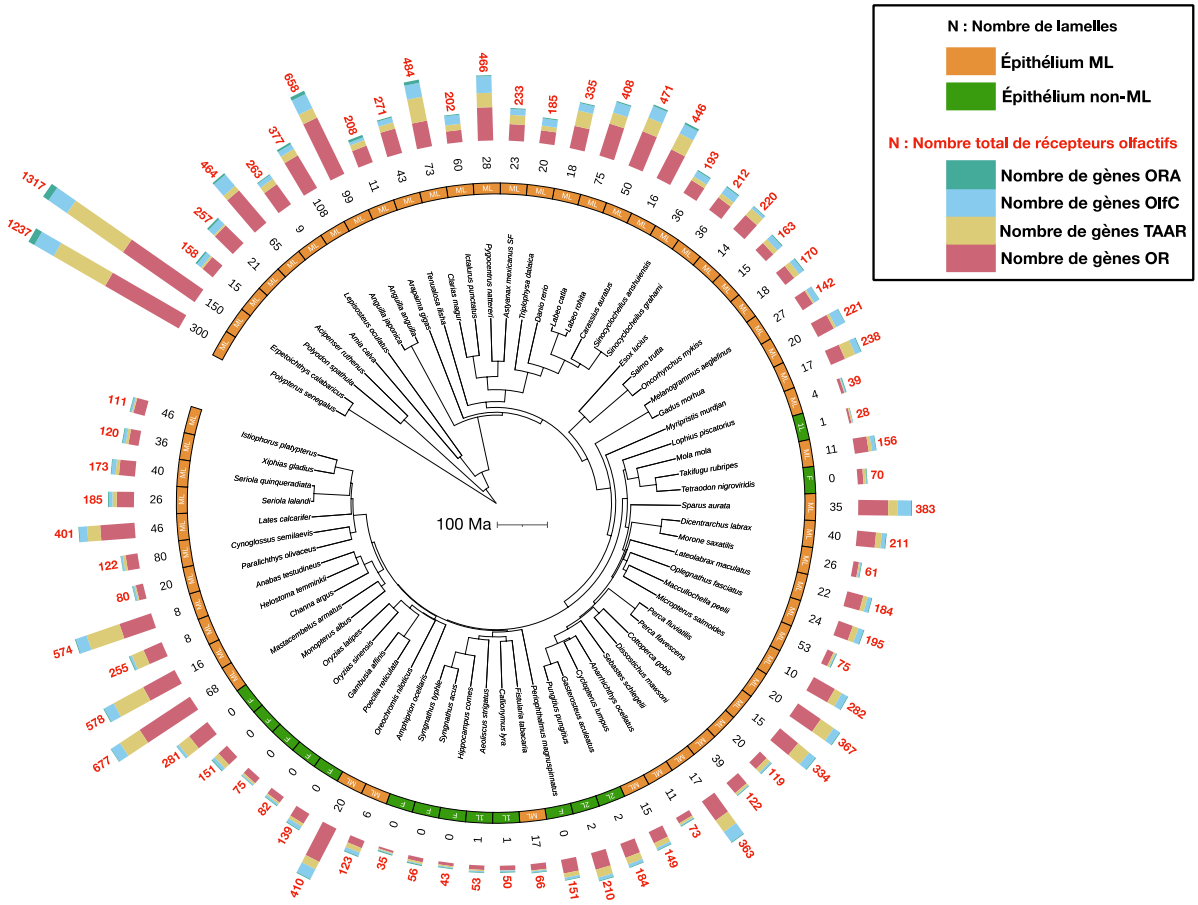


Figure 37 : Arbre phylogénétique des 72 espèces pour lesquelles des données sont disponible pour la morphologie de l'épithélium olfactif et pour le nombre de récepteurs olfactifs.

En réalisant des régressions linéaires phylogénétiques, nous observons que le nombre de lamelles dans l'organe olfactif est corrélé avec le nombre de gènes *OR* fonctionnels (R2 = 0,574 , p-value = $1{,}38e^{-14}$, Figure 38) mais également avec le nombre de gènes *OlfC* fonctionnels (R2 = 0,213 , p-value = $4{,}55e^{-5}$) et le nombre total de récepteurs olfactifs fonctionnels (R2 = 0,131 , p-value = 0,00176, Figure 38). Cependant, pas de corrélation significative n'est observé avec le nombre de gènes *TAAR* fonctionnels (R2 = 0,002, p-value = 0,726).

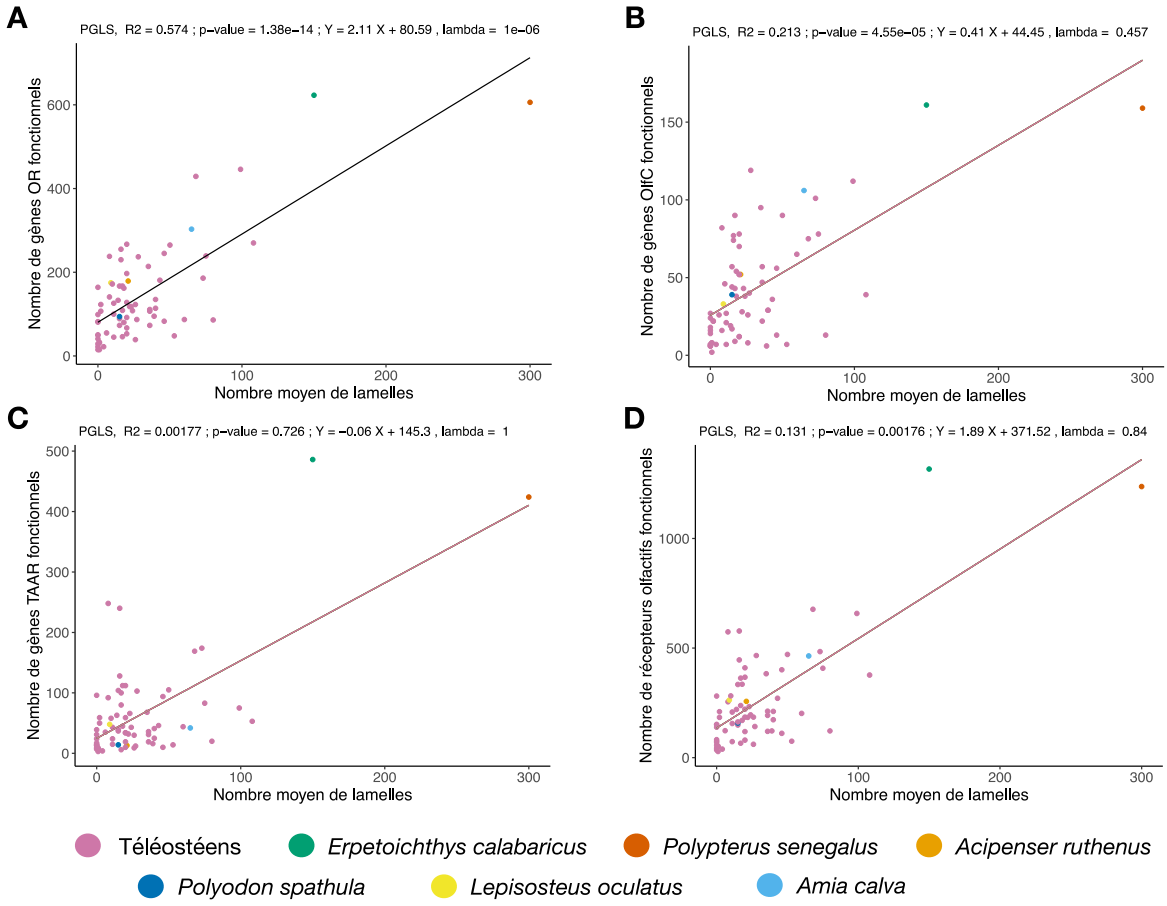


Figure 38 : Régressions linéaires phylogénétiques entre le nombre de lamelles et A- le nombre de gènes OR fonctionnels B- le nombre de récepteurs olfactifs fonctionnels C- le nombre de gènes TAAR fonctionnels D- le nombre total de récepteurs olfactifs fonctionnels. Dans chaque cas, sont indiquées la valeur du R2, de la p-value, de la valeur du lambda de Pagel et de l'équation de la droite de régression.

Ces résultats indiquent qu'une espèce avec de nombreuses lamelles aura tendance à avoir un grand répertoire de récepteurs olfactifs, tandis qu'une espèce avec peu de lamelles aura peu de récepteurs. Les lamelles permettant d'accroître le nombre de neurones olfactifs, un plus grand nombre de récepteurs vont alors pouvoir être exprimés dans ces neurones. Cette analyse mériterait d'être affinée, en prenant le nombre de neurones olfactifs ou bien la surface de l'épithélium olfactif, mais ces données ne sont disponibles que pour très peu d'espèces, telles

que les poissons-lézards (Fishelson et al. 2010) ou l'espèce modèle *Danio rerio* (Korsching et al. 1997).

F- Discussion

1) Dynamique évolutive des récepteurs olfactifs

Avec près de 200 espèces d'actinoptérygiens analysées, cette étude des récepteurs olfactifs est la plus grande effectuée à ce jour chez les vertébrés, la deuxième étant l'étude de Hughes et al. (2018) avec une analyse d'une seule des quatre familles de récepteurs olfactifs, la famille *OR*, chez 58 mammifères.

Nous avons mis en lumière la grande variabilité de la taille de répertoire de gènes olfactifs chez les actinoptérygiens, allant de 28 gènes fonctionnels chez le poisson lune *Mola mola* à 677 gènes chez l'anguille épineuse *Mastacembellus armatus* (qui est le téléostéen avec le plus grand nombre de récepteurs olfactifs), et 1317 gènes chez le polyptère *Erpetoichthys calabaricus* (qui est l'actinoptérygiens avec le plus grand nombre de récepteurs olfactifs).
Les différences entre ces nombres de récepteurs olfactifs semblent dépendre principalement des processus lents de pertes et gains de gènes, bien que des pertes massives et rapides semblent avoir eu lieu dans quelques lignées de téléostéens. Ces dernières se trouvent notamment à trois positions dans la phylogénie : chez le MRCA des Syngnathiformes et des Kurtiformes, chez le MRCA des Tétraodontiformes et des Lophiiformes et chez le MRCA de *Bagarius yarrelli* et *Tachysurus fulvidraco.*
Ces pertes massives sont relativement anciennes, datant de 112, 102 et 66 millions d'années respectivement, ce qui explique que très peu de pseudogènes sont retrouvés dans les génomes de ces espèces. Effectivement, nous avons montré à l'aide de simulations que le nombre de pseudogènes est un bon estimateur du nombre de gènes perdus chez une espèce, seulement si ces pertes sont récentes (inférieures à 40 millions d'années), après quoi ces séquences sont trop dégénérées pour être reconnues.

De grandes expansions du répertoire de gènes olfactifs semblent aussi avoir eu lieu, notamment dans la lignée avant l'ancêtre commun le plus récent des polyptériformes, groupe qui contient les espèces d'actinoptérygiens possédant le plus de gènes, mais l'absence de groupe ou d'espèce proches ne permet pas de préciser le rythme de cette amplification.

En analysant les gènes appartenant aux quatre familles de récepteurs olfactifs présentes chez les vertébrés, nous avons trouvé une corrélation forte et inattendue entre les nombres et les dynamiques de gains et de pertes des *OR*, *TAAR* et *OlfC*. Ainsi, ces trois familles ne semblent pas évoluer indépendamment les unes des autres. Cette observation remet en question la spécialisation des différentes familles dans la détection des molécules odorantes. Nous pouvons aussi penser que ces familles de gènes coévoluent suite à une contrainte partagée, par exemple liée à la morphologie de l'épithélium olfactif. Une augmentation du nombre de neurones sensoriels olfactifs mènerait à une extension de ces répertoires en même temps, tandis qu'une réduction du nombre de neurones mènerait à leur contraction.

La famille *ORA* semble avoir subi une réduction de taille majeure au cours de l'évolution des actinoptérygiens : une première suite à la perte de *ORA7* chez l'ancêtre commun le plus récent des téléostéens et une deuxième suite à la perte de *ORA8* chez le MRCA des clupéocéphales. Les gènes *ORA8* représentent la grande majorité des gènes *ORA* chez la plupart des non-téléostéens (87% chez les polyptériformes, 46% chez *Amia calva* et 40% chez *Atractosteus spatula*), des élopomorphes (80% chez les anguilles) et des osteoglossiformes (60% chez *Scleropages formosus*). Cette famille semble également avoir été perdue chez l'ancêtre commun des acipenseriformes[19]. Suite à ces différentes pertes, la plupart les téléostéens n'a conservé qu'un groupe de 6 gènes, bien que quelques exceptions existent. Par exemple, l'espèce *Clarias magur* a subi une expansion récente de la sous famille *ORA1* (16 gènes gagnés dans la branche terminale menant à cette espèce). Au contraire, des pertes secondaires ont eu lieu chez quelques espèces, par exemple chez le poisson lune (*Mola mola*) ou chez l'aiglefin

[19] Comme indiqué dans la Figure 27, aucun gène ORA8 fonctionnel n'a été retrouvé chez *Lepisosteus oculatus*. Cependant, deux gènes dans le génome de cette espèce, qui semblent être proches d'ORA8, sont classé en gènes de bordures, car incomplet et proches de fins de contig. Il reste donc compliqué de tirer une conclusion quant à une perte ou non de ORA8 chez cette espèce.

(*Melanogrammus aeglefinus*), qui ont perdus trois de ces gènes relativement récemment puisqu'on les retrouve sous forme de pseudogènes dans leur génome.

La dynamique de coévolution des différentes familles de récepteurs olfactifs pourra être mieux comprise lorsque plusieurs récepteurs seront déorphanisés, ce qui permettra de vérifier l'hypothèse de la non-spécificité des différentes familles et indiquera les paramètres écologiques influençant l'apparition ou la disparition de ces gènes.

2) Coévolution moléculaire et morphologique du système olfactif

L'épithélium des actinoptérygiens peut avoir une forme très complexe, avec plusieurs centaines de lamelles chez les polyptères ou une rosette composée de dizaines de lamelles chez la plupart des téléostéens. La rosette semble être la forme ancestrale de l'épithélium olfactif, également présente chez les chondrichtyens (Ferrando et al. 2017) . Celle-ci s'est complexifiée chez les polyptères, chez lesquels la morphologie de l'épithélium ressemble à un regroupement de 5+1 rosettes (Pfeiffer 1968; Theisen 1970) (Figure 39). Au contraire, des pertes de complexité de l'épithélium olfactif se sont produites plusieurs fois indépendamment au cours de l'évolution des téléostéens. Ainsi, chez certaines espèces, cet épithélium n'est composé que de quelques lamelles voire d'aucune lamelle.

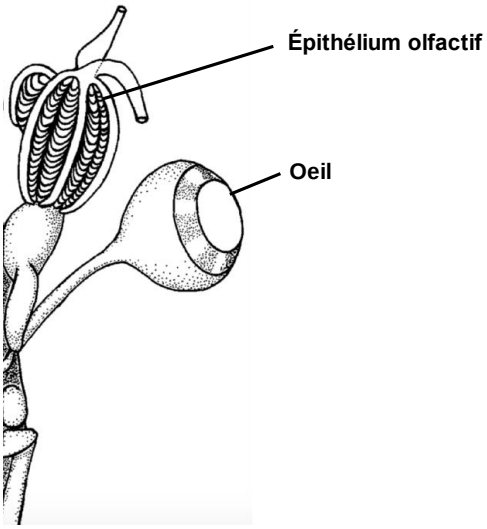


Figure 39 : Schéma de l'organe olfactif et de l'œil de Polypterus senegalus. Modifié d'après Pfeiffer 1968.

Nous avons vu que le nombre de lamelles est positivement corrélé avec la taille du répertoire olfactif. Les espèces qui ont un investissent fortement dans l'olfaction auront plus de lamelles permettant d'avoir une surface d'échange plus grande avec le milieu extérieur. Cette augmentation de surface (et probablement du nombre de neurones) permettrait à l'espèce d'exprimer un plus grand nombre de récepteurs olfactifs. Cependant, il ne faut pas confondre une grande richesse du répertoire olfactif, qui permet certainement de détecter un grand nombre d'odeurs différentes, avec une grande sensibilité olfactive. Chez les mammifères ou certaines espèces de téléostéens, il n'existe pas de corrélation entre le nombre de neurones olfactifs ou le nombre de récepteurs olfactifs et la sensibilité olfactive (Wackermannova, Pinc, et Jebavý 2016; Blin et al. 2018).

De façon intéressante, certains clades semblent avoir réacquis plusieurs lamelles après les avoir perdues, notamment dans le genre *Takifugu*.

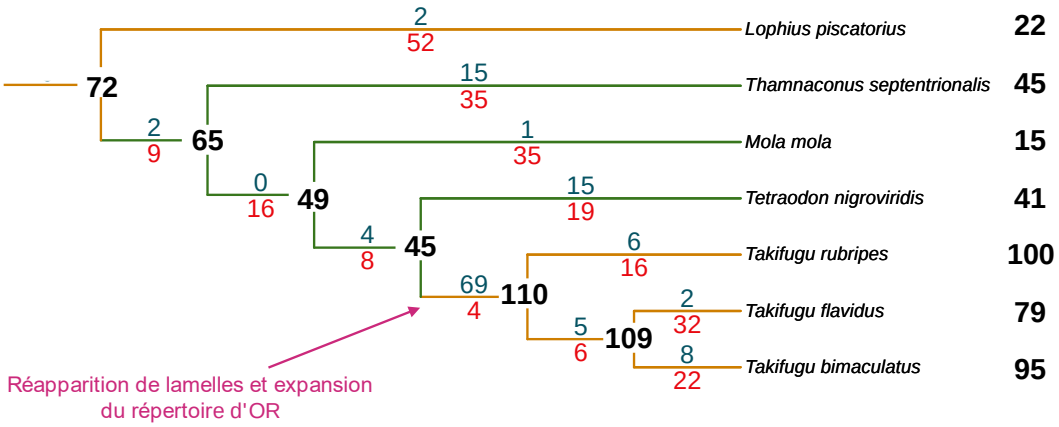


Figure 40 : Phylogénie des tétraodontiformes avec Lophius piscatorius en outgroup. Le nombre de gènes OR dans les espèces et nœuds internes est indiqué en noir. Les nombres de pertes et de gains de gènes dans chaque branche sont indiqués en rouge et bleu respectivement. Les branches sont colorées en fonction de la forme de l'épithélium olfactif : orange pour un épithélium ML, et vert pour un épithélium non-ML.

Cette réapparition estimée par une méthode de reconstruction des états ancestraux par maximum de vraisemblance semble être supportée par l'observation d'une organisation particulière des lamelles chez ces espèces. Effectivement, contrairement aux autres téléostéens, les lamelles ne sont pas organisées en rosette mais de façon parallèle, représentant un cas de convergence morphologique. Cette évolution morphologique s'est accompagnée d'une expansion du répertoire olfactif dans ce genre (Figure 40). De la même façon, chez les polyptères, la complexification de la morphologie de leur organe olfactif, permettant la présence d'un très grand nombre de lamelles, est associée à une forte expansion de leur répertoire de récepteurs olfactifs.

Cependant, il demeure incertain si l'amplification ou la disparition des récepteurs olfactifs se produit avant ou après l'apparition de modifications morphologiques. Les données sur le nombre de lamelles pourraient être complétées par des données plus précises, notamment sur le nombre de neurones olfactifs au sein de l'épithélium ou la taille relative des bulbes olfactifs et le nombre de glomérules dans ces derniers. Très peu de données sur les bulbes olfactifs existent pour les actinoptérygiens, mais certaines observations semblent cohérentes avec notre étude, notamment chez le poisson lune qui aurait des bulbes olfactifs très réduits (Burr 1928).

3) Caractéristiques écologiques liées à la taille du répertoire olfactif

Un enjeu majeur lors de l'étude des systèmes sensoriels est de relier l'importance relative de ces systèmes au mode de vie des espèces. Ici, nous n'avons pas trouvé de facteur écologique majeur déterminant la taille du répertoire olfactif des actinoptérygiens : il ne semble pas y avoir de lien avec la profondeur de vie de ces espèces, ni avec le fait de vivre en eau salée ou en eau douce. De plus, il est difficile d'expliquer pourquoi certaines espèces, vivant pourtant dans le même environnement, ont des répertoires olfactifs si différents. Par exemple, le téléostéen qui a le plus grand nombre de récepteurs, *Mastacembellus armatus*, vit dans le lac Tanganyika en Afrique où l'on trouve également un grand nombre de cichlidés, qui ont un répertoire relativement petit (72 gènes *OR*, 12 gènes *TAAR* et 28 gènes *OlfC* pour l'espèce *Neolamprologus brichardi*).

Il ne semble pas y avoir non plus de relation claire entre l'olfaction et la vision : aucune corrélation n'a été retrouvée entre le nombre de récepteurs olfactifs et la taille relative de l'œil[20]. Des analyses supplémentaires devront être réalisées afin de mieux comprendre les facteurs environnementaux ou comportementaux ayant un impact sur l'importance relative de l'olfaction. Par exemple, les poissons électriques de notre étude (*Electrophorus electricus* et *Paramormyrops kingsleyae*) ont un répertoire de récepteurs olfactifs réduit par rapport aux espèces les plus proches, ce qui pourrait indiquer une baisse d'investissement dans l'olfaction au profit de l'électroperception. D'une façon similaire, un faible nombre de récepteurs a été trouvé chez l'espèce *Danionella dracula* (111 gènes) par rapport à l'espèce la plus proche *Danio rerio* (335 gènes). Cela pourrait s'expliquer par le fait que cette espèce utilise très largement l'audition (Schulze et al. 2018).

[20] (Diamètre de l'œil / longueur standard) du poisson.

IV) La photoréception : modalité sensorielle obsolète dans les grottes

A- La photoréception visuelle

1) L'œil des vertébrés

La photoréception visuelle est assurée par un organe, l'œil, qualifié « d'une perfection extrême » par Charles Darwin dans l'Origine des espèces (1859). Il en vient même à douter du fait que seule la sélection naturelle ait pu produire cet organe : « *To suppose that the eye, with all its inimitable contrivances for adjusting the focus to different distances, for admitting different amounts of light, and for the correction of spherical and chromatic aberration, could have been formed by natural selection, seems, I freely confess, absurd in the highest possible degree* »[21]. Pour Charles Darwin, la seule façon d'expliquer l'existence de l'œil des vertébrés est de faire l'hypothèse qu'à partir d'une organisation d'abord très simple, celui-ci ait subi une complexification graduelle au cours de l'évolution.

La morphologie de l'œil est relativement similaire chez tous les vertébrés, avec les mêmes composantes principales (Figure 41 - A) :

- La **sclérotique**, membrane de protection qui permet de maintenir la forme de l'œil ;
- La **rétine sensorielle**, composée de photorécepteurs capables de détecter la lumière. Deux types de photorécepteurs existent, les cônes et les bâtonnets. Les cônes permettent la vision photopique (fortes intensités de lumière) et en couleur tandis que les bâtonnets permettent la vision scotopique (faibles intensités de lumière). Elle comprend également de multiples types d'interneurones et des cellules ganglionnaires qui projettent leur axone vers le cerveau ;

[21] « Supposer que l'œil, avec tous ses dispositifs inimitables pour ajuster le foyer à différentes distances, pour admettre différentes quantités de lumière, et pour corriger l'aberration sphérique et chromatique, ait pu être formé par sélection naturelle, semble, je l'avoue librement, absurde au plus haut point. »

- La **rétine pigmentaire,** monocouche cellulaire qui permet de stocker les rétinoïdes, composés chimiques indispensables au cycle de la vision, et les transporter vers la rétine sensorielle ;
- La **cornée**, couche transparente, jouant un rôle de protection et permettant de faire converger les rayons lumineux vers le cristallin ;
- Le **cristallin**, lentille qui permet de faire converger les rayons lumineux qu'il reçoit depuis la cornée pour projeter ces derniers sur la rétine ;
- Le **nerf optique** qui transmet les informations perçues par la rétine vers le cerveau ;
- L'**humeur vitrée**, qui assure un rôle protecteur de la rétine en absorbant une partie des rayons ultraviolets et qui permet un maintien de la rigidité et de l'élasticité de l'œil.

On estime que l'œil est apparu au moins 40 fois indépendamment chez les métazoaires (Schwab 2018) et que l'œil des vertébrés est apparu il y a environ 500 millions d'années (Lamb, Pugh, et Collin 2008). La morphologie de celui-ci a ensuite varié en fonction des différents milieux et modes de vies des organismes, les espèces nocturnes ayant par exemple une proportion de bâtonnets dans la rétine beaucoup plus élevée que les espèces diurnes (Kaskan et al. 2005).

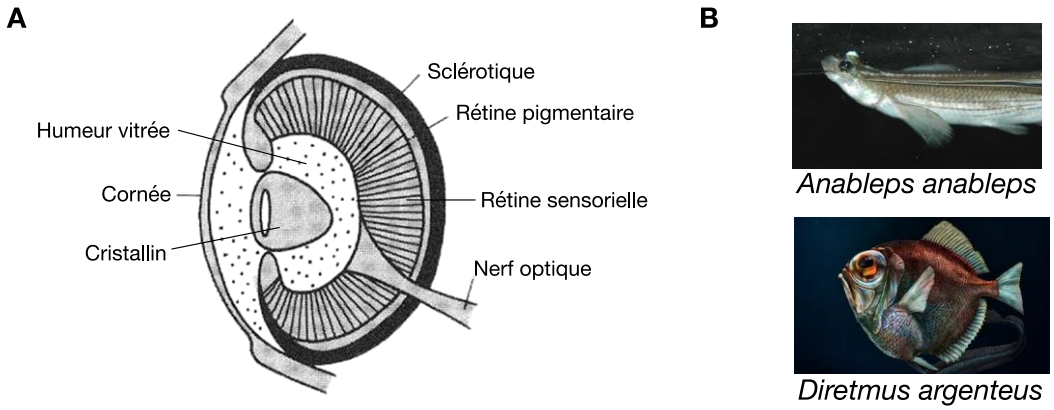


Figure 41 : Morphologie de l'œil chez les vertébrés.
A- Schéma d'un œil de vertébré., B- Photographies d'Anableps anableps et du poisson abyssal Diretmus argentus.

De plus, l'œil des espèces terrestres est souvent mal adapté à la vision aquatique car l'indice de réfraction de l'eau est plus élevé que l'indice de réfraction de l'air. Ainsi, la cornée ne peut plus

correctement dévier les rayons lumineux et les faire converger vers le cristallin. Chez les espèces marines, la cornée a une forme plus sphérique, lui permettant d'avoir un indice de réfraction plus élevé et plus proche de celui de l'eau (Marshall 2017). Cette différence est spectaculairement visible chez les poissons du genre Anableps (Cyprinodontiformes) qui sont pourvus d'yeux divisés en deux horizontalement : la partie dorsale permettant une vision dans l'air possède une cornée plus plate que la partie ventrale qui assure la vision dans l'eau (Swamynathan et al. 2003) (Figure 41 - B).

La seconde caractéristique particulière de l'eau est la forte absorption de la lumière. Seulement 1% de la lumière du soleil persiste à 100m de profondeur, et cette lumière disparaît complètement au-delà de 1000 mètres (Webb 2020). Les espèces aquatiques ont donc souvent besoin d'avoir des yeux plus sensibles à la lumière, notamment chez les espèces vivant en grande profondeur. Celles-ci ont souvent des yeux de très grande taille et une augmentation du nombre de photorécepteurs dans la rétine, leur permettant de détecter la faible lumière produite par certains organismes vivants (bioluminescence) dans les profondeurs (Figure 41).

2) La pigmentation : déjouer et attirer l'attention

La vision permet aux espèces d'avoir une représentation de l'environnement les entourant, et de reconnaître les organismes vivants ou éléments abiotiques qui ont des formes, des couleurs et des mouvements caractéristiques.

Afin de réduire la prédation, dans laquelle la vision joue souvent un rôle majeur, la pigmentation peut être un grand atout pour se camoufler. L'exemple certainement le plus connu chez les vertébrés est la présence de rayures sur les zèbres. Des études récentes montrent que ces rayures ne servent pas ou très peu, contrairement à la croyance populaire, de camouflage contre les mammifères prédateurs tel que les lions ou les hyènes (Melin et al. 2016) mais permettent de réduire significativement l'atterrissage et la piqûre de tabanidés (famille des mouches), vecteurs de maladies mortelles (Caro et al. 2019).

Le camouflage peut servir à ne pas être détecté mais également à ne pas être reconnu. Dans l'eau, la pigmentation joue un rôle majeur dans les stratégies de camouflage. Les lumières de

grandes longueurs d'ondes étant absorbées en premières, de nombreuses espèces de poissons apparaîtront noires grâce à leur pigmentation rouge (Marshall et al. 2019).

Hormis l'utilisation de la pigmentation, d'autres stratégies de camouflage existent au sein des poissons, notamment par l'acquisition de formes ou de mouvements particuliers (quelques exemples de stratégies de camouflage sont présentés dans la figure ci-dessous).

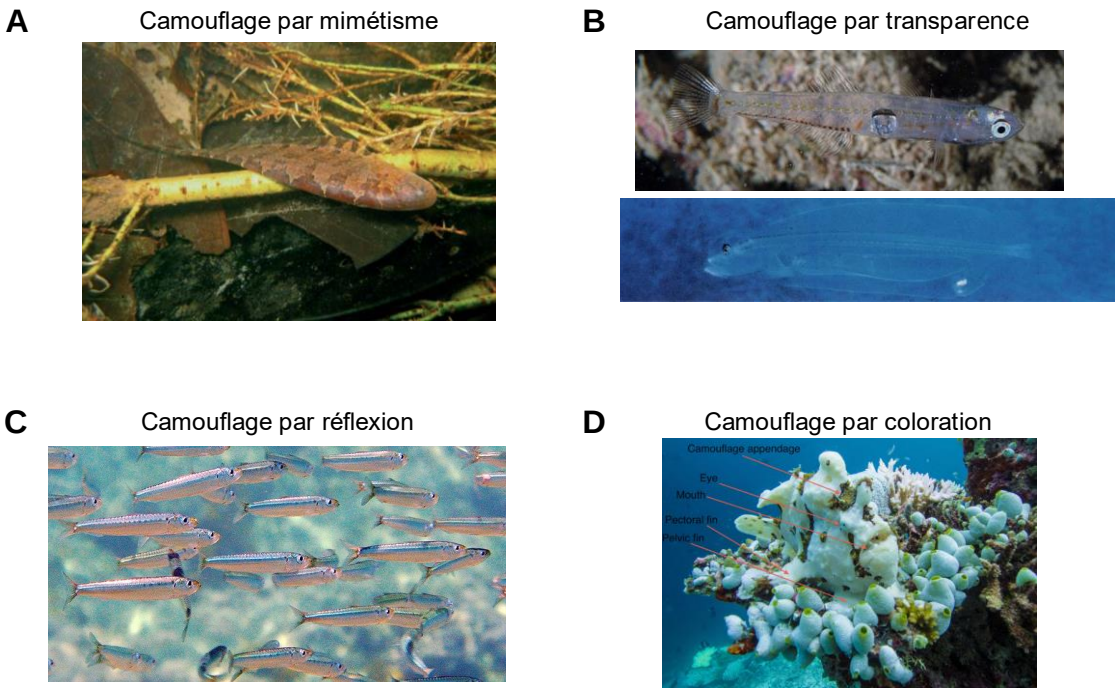


Figure 42 : Exemples de stratégies de camouflage utilisées par les téléostéens. A- Le camouflage par mimétisme est l'imitation d'une autre espèce ou d'un objet abiotique. Ici, l'espèce Steatogenys duidae imite une feuille morte. B- Le camouflage par transparence est assez rare chez les adultes mais plus commun chez les larves des espèces de téléostéens. En haut, un gobie transparent adulte (Aphia minuta) et en bas, une larve d'une espèce de Bathophilus. C- Le camouflage par réflexion permet à de nombreuses espèces de réfléchir la lumière du soleil, notamment grâce à des cellules spécialisées, les iridophores. Ici, une photo d'un groupe de sardines (Sardina pilchardus). D- Le camouflage par coloration permet à certaines espèces d'adapter leur pigmentation pour passer inaperçu. Ce changement de couleur peut être très rapide, en quelque secondes chez l'espèce de poisson plat Bothus mancus, ou en quelques semaines comme ici chez le poisson-grenouille Antennarius maculatus, qui est blanc pour se camoufler dans des coraux blanchis.

Le camouflage peut également servir à augmenter le succès de prédation, lorsque celui-ci est utilisé pour ne pas être repéré par les potentielles proies. C'est le cas pour de nombreuses espèces de poissons qui chassent à l'affût, tels que le carrelet ocellé *Bothus ocellatus* qui peut changer de couleur en quelques secondes (Ramachandran et al. 1996), ou le poisson-scorpion *Rhinopias aphanes* qui imite les algues par sa forme et sa couleur (Marshall et al. 2019).

En plus du camouflage, la pigmentation joue souvent un rôle central dans les comportements reproducteurs, les différentes couleurs et leur disposition sur l'animal permettant notamment d'attirer les congénères du sexe opposé. Ainsi, de nombreuses espèces de téléostéens possèdent un dimorphisme sexuel au niveau de leur coloration qui peut être couplé ou non à des différences comportementales. Par exemple, chez l'espèce *Xiphophorus pygmaeus*, les mâles peuvent être bleu ou jaune. Les individus jaunes vont tendre à avoir un comportement plus agressif envers les femelles tandis que celles-ci vont préférer les individus bleus (Kingston, Rosenthal, et Ryan 2003). De plus, chez l'espèce d'épinoche *Gasterosteus aculatus*, l'apparition de motifs noirs temporaires chez la femelle est associée à une attirance plus forte des mâles (Rowland, Baube, et Horan 1991). Afin d'éviter un conflit sexuel, qui arrive quand un trait est avantageux pour un sexe mais désavantageux pour l'autre, les gènes régulant cette pigmentation sont souvent localisés sur des régions non-recombinantes des chromosomes sexuels (Kottler et Schartl 2018).

Chez les poissons, six grandes classes de cellules pigmentaires, également appelées chromatophores, existent : les mélanophores (couleur noire), les iridophores (spécialisées dans la réflexion de la lumière), les xantophores (couleur jaune / orange), les érythrophores (couleur rouge), les leucophores (couleur blanche) et enfin les cyanophores (couleur bleue) (Irion et Nüsslein-Volhard 2019).

3) Gènes impliqués dans la photoréception visuelle

La vision est possible grâce aux différentes structures de l'œil et aux gènes qui y sont exprimés, qui assurent le bon déroulement de la cascade de phototransduction (Figure 43), c'est-à-dire la transformation du stimulus lumineux en signal électrique transmis au cerveau.

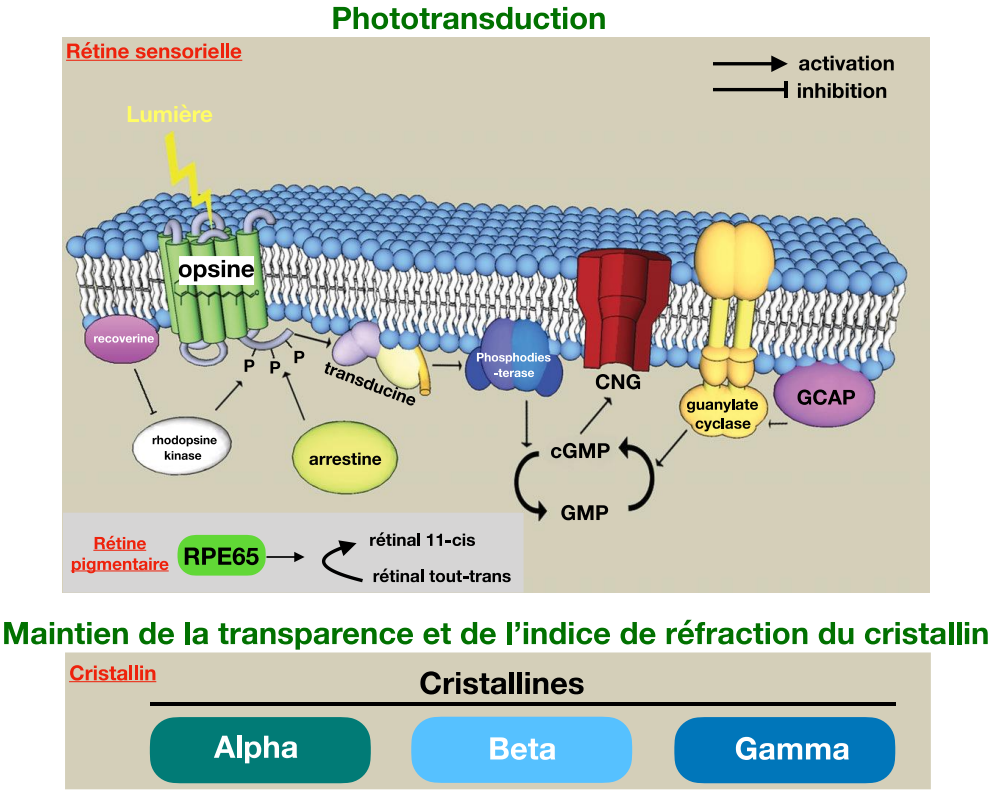


Figure 43 : Schéma récapitulant les différents gènes impliqués dans la phototransduction et dans le maintien du cristallin.
Modifié d'après Larhammar, Nordström, et Larsson (2009).

La lumière, focalisée par la cornée puis par le cristallin sur la rétine, sera absorbée au niveau des protéines présentes dans les photorécepteurs : les **opsines**. Ces protéines transmembranaires sont couplées au rétinal 11-*cis* (composé chimique de la famille des rétinoïdes) qui va changer de conformation lors de l'absorption d'un photon pour devenir du rétinal tout-*trans* et activer

l'opsine. Selon sa séquence en acides aminés, l'opsine sera optimisée pour absorber certaines longueurs d'ondes de la lumière, et chez les vertébrés, cinq familles d'opsines couplée au rétinal 11-*cis* existent (Yokoyama 2000) (Tableau ci-dessous).

Famille d'opsine	Longueurs d'ondes absorbées
LWS	510-560 nm
RH1	500 nm
RH2	470-510 nm
SWS2	440-460 nm
SWS1	360-430 nm

Tableau 2 : Longueurs d'ondes absorbées par les différentes familles d'opsines présentes chez les vertébrés.

L'opsine activera les **transducines** qui vont elles-mêmes activer les **phosphodiesterases**. Ces dernières permettent de convertir le nucléotide cyclique cGMP en GMP et cette baisse en concentration de cGMP entrainera la fermeture de canaux ioniques, les protéines **CNG**, puis l'hyperpolarisation de la cellule photoréceptrice et l'excitation des neurones rétiniens.
L'arrêt du cycle sera enclenché par les **rhodopsines kinases**, qui vont phosphoryler les opsines et permettre aux **arrestines** de s'y attacher pour les inhiber. Enfin, la pérennité du cycle visuel est assurée par les **guanylate cyclases** qui retransformeront le GMP en cGMP , par les **recoverines** qui inhiberont l'action des rhodopsines kinases (Larhammar, Nordström, et Larsson 2009) et par les protéines **RPE65** (exprimées dans la rétine pigmentaire) qui recycleront le rétinal tout-*trans* en rétinal 11-*cis*.

En plus de ces gènes indispensables au bon déroulement du cycle de phototransduction, la transparence et l'indice de réfraction du cristallin sont primordiaux pour assurer la propagation des rayons lumineux jusqu'à la rétine. Cette fonction est assurée par les **cristallines**, que l'on peut classer en trois familles : alpha, beta et gamma. Certaines cristallines alpha, qui sont des protéines de choc thermique (HSPs) avec une fonction de chaperonne, jouent en plus un rôle dans la prévention de l'apoptose des cellules du cristallin et assurent la solubilité des autres cristallines (Andley 2009).

B- La photoréception non-visuelle

La photoréception non-visuelle, également appelée photoréception extra-oculaire, est la capacité de détecter les variations de luminosité grâce à des photorécepteurs dits « profonds ». Ces derniers sont localisés dans le cerveau et sous la boite crânienne, principalement dans la glande pinéale et l'hypothalamus, ou à la surface du corps.
Ces cellules expriment des opsines, les opsines non-visuelles, présentes chez tous les vertébrés mais que l'on trouve en plus grand nombre chez les téléostéens, avec 32 opsines non-visuelles chez le poisson-zèbre contre une dizaine seulement chez l'humain (Davies et al. 2015) L'utilité d'un si grand répertoire d'opsines non-visuelles est encore mal comprise, leurs rôles distincts et leurs spectres d'absorptions étant inconnus pour la grande majorité (Fischer et al. 2013; Foulkes et al. 2016).

Afin de s'adapter aux changements environnementaux cycliques de la Terre, les organismes vivants ont développé des horloges biologiques régulées par différents facteurs (également appelés « *zeitgeber* ») tels que la lumière, la température ou la disponibilité en nourriture (Foulkes et al. 2016). L'horloge circadienne ou cycle circadien, rythme entraîné par la lumière, est notamment médié par la photoréception non-visuelle.

Chez les téléostéens, le cycle circadien est régulé par une boucle de rétroaction négative.
Cette boucle est constituée des gènes *bmal* et *clock* qui forment des hétérodimères (Figure 44). Ceux-ci joueront le rôle d'activateur et enclencheront l'expression des gènes *ror* qui assureront la continuité d'expression de *bmal* et *clock*. Ces derniers enclencheront la transcription des gènes *period (per)* et des *cryptochromes (cry)*, qui pourront réprimer leur propre expression, et permettre l'initiation d'une nouvelle boucle. Le rythme de cette boucle est en relation directe avec la réception de lumière par les opsines non-visuelles : celles-ci enclencheront la transcription des gènes *tef* qui serviront de facteur de transcription pour les gènes *per* et *cry* (Vatine et al. 2011). En plus de ces gènes principaux, le cycle circadien est régulé par d'autres familles de gènes, impliqués dans le relais ou la régulation de cette boucle rétroactive (Li et al. 2013) mais également dans la phosphorylation de *per* et *cry*, qui doivent être rapidement dégradés afin d'assurer cet effet de boucle (Takahashi et al. 2008) (Figure 44).

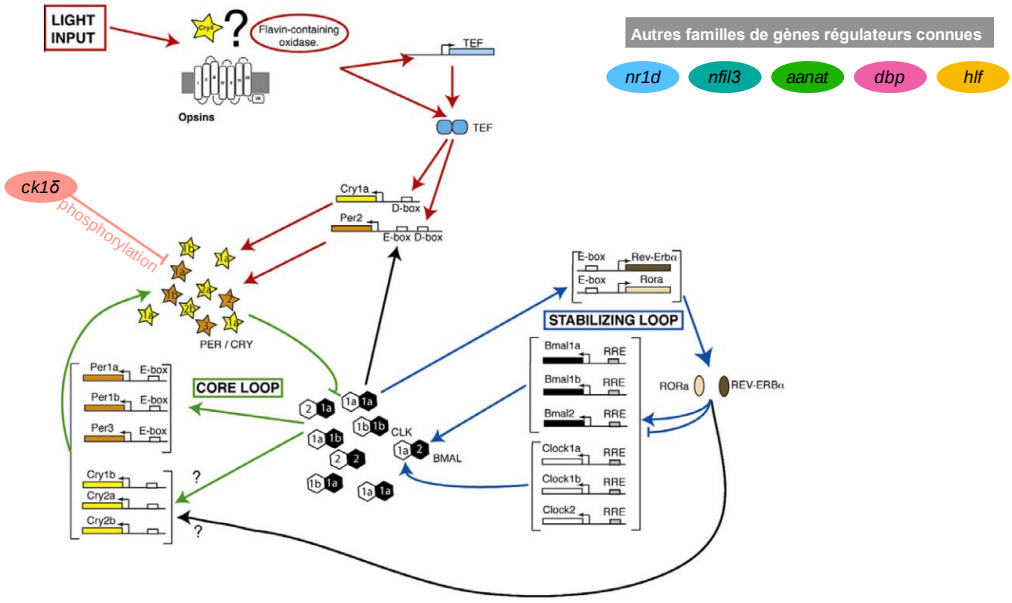


Figure 44 : Schéma représentant la boucle rétroactive du cycle circadien. Modifié d'après Vatine et al. (2011).

C- Les organismes cavernicoles

Dans l'étude précédente sur l'olfaction, les récepteurs olfactifs ont été recherchés chez le plus grand nombre d'espèces possible, pour trouver des liens entre richesse du répertoire de gènes et des caractéristiques morphologiques et écologiques. Ici, l'étude des gènes de la photoréception suivra un autre cheminement : ils seront comparés entre des espèces vivant en surface et des espèces vivant dans des environnements dépourvus de lumière, notamment les grottes, où cette modalité sensorielle est dispensable.

1) Définition d'une grotte et de sa biodiversité

Les grottes sont définies comme des ouvertures naturelles dans le sol, qui s'étendent vers une zone où la lumière du soleil est complètement absente (Davies et Morgan 1991) et qui sont totalement ou partiellement accessibles par l'Homme. Elles peuvent se former à la suite de l'érosion de roche par de l'acide naturel contenu dans l'eau souterraine (grotte karstique) ou par érosion mécanique suite à l'action des vagues (grotte marine). Elles peuvent également se former à la suite d'activités volcaniques (tunnels de lave), ou par l'écoulement d'eau sous un glacier (grotte glacière) bien que ces dernières soient souvent temporaires car très instables.

La première description d'une espèce vivant exclusivement dans les grottes remonte au 18ème siècle, lorsque le naturaliste Josephus Nicolaus Laurenti étudie et nomme la salamandre cavernicole, *Proteus anguinus*, qui vit dans des grottes karstiques des Alpes dinariques. Cette espèce était évoquée dès le 17ème siècle par le naturaliste Johann Weikhard Freiherr von Valvasor (Aljančič 2019). Au 19ème siècle, la vie dans les grottes attire l'attention de Charles Darwin (1859), qui discute la présence de formes différentes de vie cavernicoles, telles que des insectes, des amphibiens et des poissons.
Aujourd'hui, plusieurs laboratoires à travers le monde s'intéressent à l'évolution des espèces que l'on retrouve dans les grottes. On peut classer ces espèces en trois catégories (Mammola 2019) :

- Les espèces trogloxènes qui vivent de façon temporaire dans les grottes (= habitat hypogé) et qui ne peuvent compléter leur cycle de vie sans aller en surface (= habitat épigé) ;
- Les espèces troglophiles chez lesquelles on trouve des populations entièrement hypogées ou épigées, avec souvent des migrations entre les deux milieux ;
- Les espèces troglobies qui vivent exclusivement dans les grottes.

Ces espèces peuvent être trouvées dans les grottes marines (Pettibone 1985) et karstiques (Graening et al. 2010) ainsi que dans les tunnels de lave (Wynne 2011) (Figure 45).

Pelagomacellicephala iliffei *Amblyopsis rosae* *Tiphia andersoni*

Figure 45 : Photographies d'organismes cavernicoles issus, de gauche à droite, d'une grotte marine, d'une grotte karstique et d'un tunnel de lave.

On pense qu'il existe au moins 7000 espèces troglobies aquatiques, dont un peu moins de 300 téléostéens décrits (Proudlove 2021), et 21000 espèces troglobies terrestres (Juan et al. 2010). Ces nombres sont sûrement des sous estimations, la détection et l'exploration des grottes étant compliquées et souvent très incomplètes.

2) Adaptations morphologiques à la vie dans les grottes

Les animaux cavernicoles doivent souvent s'adapter à un milieu pauvre en nourriture : les grottes sont dépourvues de lumière du soleil et ne permettent pas la croissance d'organismes photosynthétiques, premiers maillons de la chaine alimentaire. Les apports de nourriture se font sous forme de guano de chauve-souris (ou fèces d'autre animaux trogloxènes), d'animaux morts, de champignons et de détritus végétaux, souvent apportés de façon irrégulière dans la grotte (Trajano 2000). Ainsi, des adaptations à une privation plus ou moins régulière de nourriture ont émergé. L'espèce *Proteus anguinus* a un métabolisme fortement réduit, impliquant une locomotion plus faible en conditions normales et encore plus faible pendant les périodes de privation alimentaire quand on la compare aux salamandres épigées. De plus, son stockage de lipides est plus important (Hervant, Mathieu, et Durand 2001) (Figure 46 - B), tout comme chez le poisson *Astyanax mexicanus*, dont certaines populations ont en plus développé un appétit insatiable en présence de nourriture (Aspiras et al. 2015). Il en va de même pour les

poissons cavernicoles de la famille des Amblyopsidae, chez lesquels une baisse du métabolisme a été observée malgré une locomotion plus élevée que chez les espèces de surface (Poulson 1963).

L'absence totale de lumière a également un impact direct sur le phénotype des organismes cavernicoles, qui n'ont plus la possibilité d'utiliser la photoréception comme modalité sensorielle. Cela entraine souvent une forte réduction de la taille des yeux (= microphtalmie) voire une absence totale de ces derniers (= anophtalmie). Chez tous les vertébrés cavernicoles, les yeux commencent cependant à se développer dans les premiers stades de vie, comme pour *Astyanax mexicanus* chez qui un développement initial quasi-normal de l'œil après fécondation est suivi par une apoptose du cristallin puis de l'œil entier 40 h après fécondation (Alunni et al. 2007). Chez le poisson *Phreatichthys andruzzii*, un développement normal de l'œil est observé jusqu'à 32 h après fécondation, après quoi l'œil entre en apoptose puis est totalement recouvert de peau entre 74 h et 98 h après fécondation (Berti et al. 2001) (Figure 46 - A). Le développement initial de l'œil observé chez tous les vertébrés cavernicoles suggère qu'il serait nécessaire à la morphogénèse du cerveau et représenterait une contrainte développementale (Rétaux et Casane 2013; Devos et al. 2021).

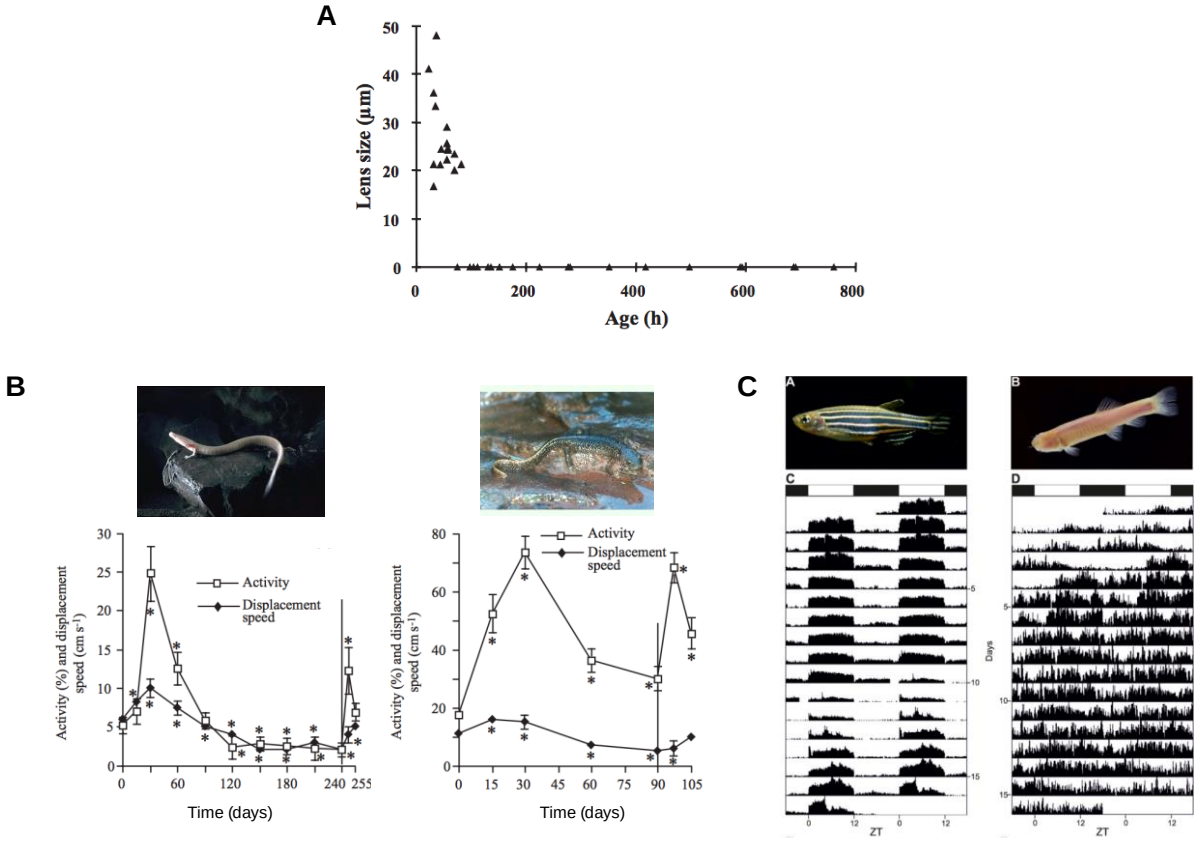


Figure 46 : Évolution phénotypique chez les organismes cavernicoles.
A- Évolution de la taille du cristallin en fonction de l'âge (en heure) chez Phreatichthys andruzzii. Modifié d'après Berti et al. (2001). B- Activité et vitesse de déplacement chez Proteus anguinus à gauche et Euproctus asper à droite. L'activité et la vitesse de déplacement en conditions normale (t = 0 jour) sont 4 fois plus faibles chez Proteus anguinus et restent plus faibles pendant une période de privation alimentaire (t = 1 jour jusqu'à 250 jours pour la salamandre cavernicole ou 90 jours pour la salamandre de surface). Modifié d'après Hervant, Mathieu, et Durand (2001) C- Actogramme du Danio rerio et de Phreatichthys andruzzii pendant un cycle 12h lumière / 12h obscurité. Modifié d'après Cavallari et al. (2011).

Cette perte de photoréception se manifeste également au niveau non-visuel, avec une perte du cycle circadien chez des organismes cavernicoles, qui a été étudiée plus en détail chez *Phreatichthys andruzzii*. Il a été montré pour cette espèce que l'activité de locomotion ne suit pas le cycle de lumière qui lui est imposé (cycle 12h lumière / 12h obscurité) contrairement aux poissons de surface, qui sont actifs en présence de lumière et non-actifs pendant la phase obscure (Figure 46 - C). Cependant, le rythme biologique de cette espèce semble toujours répondre aux variations temporelles de disponibilité de nourriture (Cavallari et al. 2011). Cela

indique que l'apport de nutriments dans les grottes pourrait, dans certains cas, être périodique (et être un *zeitgeber*), par exemple lié à l'activité des chauves-souris (Beale et al. 2013).

Il est intéressant d'observer que chez les espèces ayant perdu la photoréception, les autres modalités sensorielles sont souvent augmentées. Ainsi, le poisson cavernicole *Astyanax mexicanus,* possède un nombre plus élevé de neuromastes accompagné d'une sensibilité accrue aux vibrations de l'eau (Yoshizawa et al. 2010), une sensibilité olfactive 100 000 fois supérieure et possède deux à trois fois plus de bourgeons gustatifs que les poissons de surface (Hinaux et al. 2016; Varatharasan, Croll, et Franz-Odendaal 2009).
Enfin, un autre trait morphologique qui est partagé par la majorité des organismes cavernicoles est la perte de la pigmentation. Celle-ci n'est plus nécessaire pour protéger l'organisme des rayons UV et ne peut plus être utilisée pour la reconnaissance intra ou inter-spécifique ou pour le camouflage.

L'ensemble de ces caractéristiques morphologiques et comportementales, qui ont évolué de façon indépendante chez de nombreuses espèces cavernicoles (= convergence évolutive), est appelé troglomorphisme.

3) Datation des organismes cavernicoles

Une des principales questions concernant les espèces troglomorphes, introduite par Charles Darwin qui les qualifie de « *wreck of ancient life* » (épaves de formes de vie passées), est de savoir depuis combien de temps celles-ci ont colonisé les grottes.

a) Les méthodes de datation moléculaire

La datation moléculaire repose sur l'analyse de deux ou plusieurs séquences (ADN ou protéine) qui descendent d'un même ancêtre commun, et qui ont divergé en accumulant des mutations. On peut dater la divergence entre deux séquences issues d'une duplication au sein d'un même

génome, appelées séquences paralogues, mais également dater la divergence entre deux séquences chez des espèces différentes, appelées séquences orthologues[22].

La notion d'horloge moléculaire a été introduite dans une publication de 1965 par Emile Zuckerkandl et Linus Pauling : on peut retrouver la phrase « *There may thus exist a molecular evolutionary clock* »[23] (Zuckerkandl et Pauling 1965). Cette notion d'horloge moléculaire repose sur leur observation que le nombre de différences au sein d'une même protéine chez deux espèces est proportionnel au temps de divergence entre celles-ci. La deuxième avancée majeure dans la datation moléculaire est la théorisation de l'évolution neutre par Motoo Kimura : la majorité des mutations étant neutres, le nombre de substitutions[24] dans une lignée dépend uniquement du taux de mutation.

Effectivement, si l'on considère une population diploïde de taille N et un taux de mutation par site par génération μ, une nouvelle mutation neutre aura $2N\mu$ chances d'apparaître et $\frac{1}{2N}$ chances de se fixer dans la population. Le taux de substitution sera alors égal à $2N\mu \times \frac{1}{2N} = \mu$.

Cependant, les datations à l'aide de cette horloge moléculaire peuvent être imprécises car le taux de mutation varie entre lignées et également au sein d'un même génome. De plus, certaines séquences utilisées peuvent être sous forte pression de sélection, ce qui augmente ou réduit fortement la probabilité de fixation d'une mutation. Ainsi, de nombreuses méthodes et modèles probabilistes ont vu le jour pour prendre en compte la variation de ce taux de substitution dans les datations (Lanfear, Welch, et Bromham 2010).

[22] Une séquence A chez l'ancêtre commun ayant divergé chez ces deux espèces, pour devenir la séquence A1 chez l'espèce 1 et la séquence A2 chez l'espèce 2.
[23] « Il pourrait donc exister une horloge de l'évolution moléculaire »
[24] Mutations qui atteignent la fixation dans une espèce ou une population

b) La datation moléculaire de la colonisation des grottes

La principale difficulté pour dater l'origine des organismes cavernicoles vient du fait que leur divergence avec les espèces de surface les plus proches n'est, dans la majorité des cas, pas égale au temps de colonisation des grottes (Figure 47). Dans le cas de populations cavernicoles pour lesquelles la même espèce existe avec une forme épigée, une autre difficulté vient du fait qu'il y a de possibles migrations et flux de gènes depuis la surface vers les grottes. Ceux-ci doivent alors être pris en compte dans les modèles de datation.

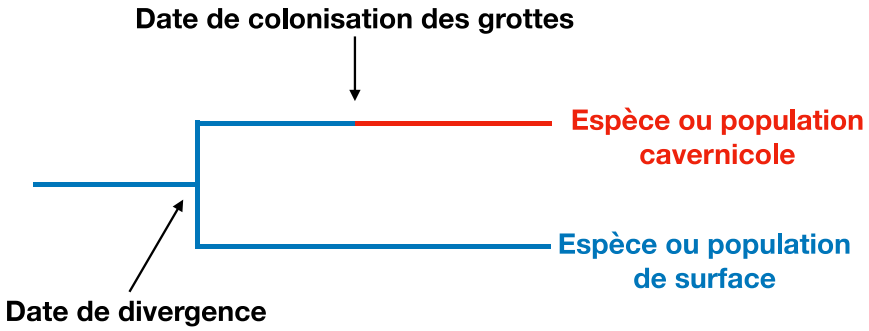


Figure 47 : Différence entre temps de divergence et temps de colonisation des grottes.

Pourtant, dans la littérature, la confusion entre divergence (entre espèces ou entre populations) et colonisation des grottes persiste. Dans un article de 2008, sur l'arthropode cavernicole *Palmorchestia hypogaea* (Villacorta et al. 2008), les auteurs estiment, à l'aide d'horloges moléculaires, le temps de divergence entre des populations cavernicoles et des populations de surface, puis interprètent directement ce temps de divergence avec la phrase « *This shows that the adaptation to cave life is relatively ancient* »[25]. Or, le temps de divergence *T* représente uniquement la limite supérieure du temps de colonisation *CG*, qui se situe entre 0 et T. Cependant, le temps de divergence peut déjà donner un ordre de grandeur du temps *CG*.

Par exemple, pour l'espèce *Astyanax mexicanus*, dont on pensait qu'elle avait colonisé les grottes il y a plusieurs millions d'années, des modélisations prenant en compte différents

[25] « Cela montre que l'adaptation à la vie cavernicole est relativement ancienne »

paramètres démographiques, tels que des migrations et des variations dans les tailles de populations, suggèrent que les populations cavernicoles auraient divergé des populations de surface il y a environ 20 000 ans (Fumey et al. 2018). Ce qui est particulièrement intéressant avec cette nouvelle hypothèse, c'est qu'elle implique des changements morphologiques, physiologiques et comportementaux très rapides associés à un changement radical d'environnement.

Une des méthodes qui existe afin d'estimer le temps *CG* consiste en l'utilisation d'un ou de plusieurs gènes qui évoluent sous sélection chez les espèces de surface mais qui évoluent de façon neutre chez les espèces cavernicoles. Le principe de cette méthode est d'estimer, à partir des mutations observées dans un gène, depuis quand celui-ci est passé d'une évolution sous sélection à une évolution neutre, ce qui devrait correspondre au temps depuis lequel l'organisme vit dans l'obscurité. Dans la publication de Calderoni et al. (2016), et après avoir observé que le cycle circadien était inactivé chez l'espèce *Phreatichthys andruzzii*, les auteurs ont comparé la séquence d'une opsine non-visuelle (*opn4m2*) de cette espèce et d'une espèce proche, *Garra barreimiae*. La date du changement de régime d'évolution du gène *opn4m2* a été estimée entre 2,5 et 9 millions d'années. Une des principales limitations dans cette étude est l'utilisation d'un seul gène, et donc d'un nombre très limité de substitutions (6 non-synonymes, 2 synonymes), seules sources d'information pour cette méthode, et qui explique la grande variance associée à cette estimation. L'application de cette approche impose d'être certain de la neutralité d'un gène. Cela est possible quand on peut observer ce dernier sous forme inactive dans le génome, c'est-à-dire sous la forme d'un pseudogène (définit dans l'introduction générale, Figure 5).

4) Analyses moléculaires chez les organismes cavernicoles

Jusqu'à présent, la majorité des mutations impliquées dans le troglomorphisme a été identifiée chez *Astyanax mexicanus*. Cette espèce a l'avantage d'exister sous la forme de deux morphotypes, le morphotype de surface non troglomorphe et le morphotype cavernicole, qui sont interfertiles. Le croisement entre individus issus de populations de surface et d'individus issus de populations cavernicoles permet de réaliser la recherche de QTL (quantitative trait

loci), c'est-à-dire de cartographier sur le génome un ou plusieurs locus avec des effets majeurs sur un changement phénotypique (Figure 48).

Cette approche a été utilisée pour la première fois chez cette espèce en 2002 (Borowsky et Wilkens 2002), mais le niveau de résolution de la cartographie ne permettait pas de trouver les gènes impliqués dans le troglomorphisme. Ce n'est que quatre ans plus tard que cette même approche, avec une meilleure résolution, a permis d'identifier, dans des populations cavernicoles, la présence de mutations « perte de fonction » dans le gène *oca2* qui est impliqué dans la pigmentation (Protas et al. 2006).
Depuis, d'autres recherches de QTL ont permis l'identification de plusieurs autres mutations impliquées dans les changements phénotypiques d'*Astyanax mexicanus*, dont un décalage du cadre de lecture et des mutations non-synonymes délétères dans le gène de pigmentation *mc1r* (Gross, Borowsky, et Tabin 2009). Cependant, la résolution reste aujourd'hui encore la principale limite de cette méthode, la plupart des études permettant seulement d'identifier un ensemble de gènes candidats ou un ensemble de régions génomiques impliqués dans un trait phénotypique.

La seconde méthode employée pour relier phénotype et modifications génomiques est l'analyse de gènes candidats, c'est-à-dire l'étude d'un ou quelques gènes choisis *a priori* sur la base de connaissances sur leur fonction. Cette méthode a par exemple permis d'identifier plusieurs mutations chez *Astyanax mexicanus* : dans le gène *mc4r* impliquées dans la résistance à la privation de nourriture (Aspiras et al. 2015), dans le gène du récepteur à l'insuline *InsR* impliqué dans les variations métaboliques (Riddle et al. 2018) ou encore une mutation dans le gène *mao* impliquée dans des différences comportementales (Elipot et al. 2014). Avec plus de 20 000 gènes dans la grande majorité des génomes d'actinoptérygiens, il est presque impossible, sans connaissances *a priori*, d'établir un lien fonctionnel entre un phénotype observé et un gène. Ainsi, il est toujours nécessaire de réaliser des cribles génétiques, tels que les analyses QTL ou d'utiliser des méthodes plus précises telles que l'insertion d'une mutation dans un gène donné afin d'évaluer sa conséquence sur le phénotype. Cela est souvent réalisées sur des organismes modèles comme la souris *Mus musculus* ou le poisson-zèbre *Danio rerio.*

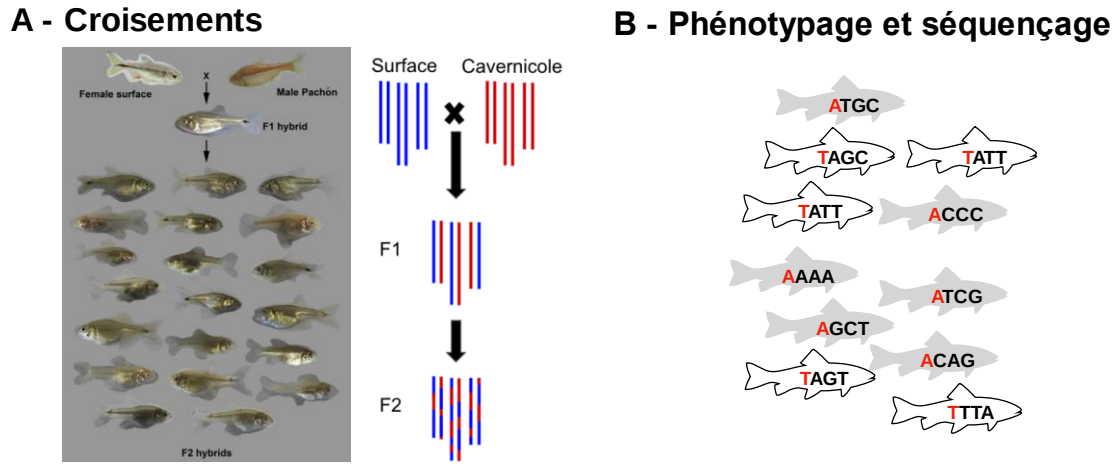


Figure 48 : Principe d'une analyse QTL chez Astyanax mexicanus. A- Croisement d'un individu de surface et d'un individu cavernicole puis obtention d'une génération F2, chez laquelle le génome est un mélange surface/cavernicole. Modifié d'après Casane et Rétaux (2016) B- Phénotypage et séquençage de tous les individus F2. Ici, le génome est simplifié à 4 nucléotides et le site 1, en rouge, est significativement associé à la pigmentation car les individus qui ont l'allèle A sont tous pigmentés tandis que tous les individus qui ont l'allèle T sont dépigmentés.

Aujourd'hui, la disponibilité de génomes d'organismes cavernicoles permet de réaliser des analyses à l'échelle de tous les gènes, également appelées « genome-wide analysis ». Ces analyses de grands jeux de données reposent sur l'annotation automatique des gènes. C'est une tâche très complexe pour les génomes eucaryotes chez qui les gènes ont plusieurs exons, entrainant de nombreuses erreurs d'annotation, même avec les programmes d'annotation les plus récents (Salzberg 2019). Aussi, les études genome-wide réalisées sur les poissons troglomorphes (McGaugh et al. 2014; Yang et al. 2016; Aardema, Stiassny, et Alter 2020) n'ont pas comporté d'analyses comparatives ni quantitatives entre espèces. A cause de ces deux facteurs limitants, peu de conclusions ont pu être tirées quant au processus de pseudogénisation ou d'accumulation de substitutions non-synonymes chez les organismes cavernicoles.

V) Résultats : Étude des gènes de la photoréception, de la pigmentation et du cycle circadien chez les poissons cavernicoles

Je présenterais dans les chapitres suivants les résultats de mes travaux sur les poissons cavernicoles. Une sous-partie de ces résultats, qui concerne les espèces *Astyanax mexicanus*, le genre *Lucifuga* et le genre *Sinocyclocheilus* est publié. L'autre sous-partie, qui concerne les espèces *Typhlichthys subterraneus* et *Lamprologus lethops* n'est actuellement disponible que sous forme de « pré-print ».

A- Hypothèses de travail

Dans notre étude, nous avons voulu analyser les gènes impliqués dans la photoréception (visuelle et non-visuelle), le cycle circadien et la pigmentation dans les différents génomes cavernicoles disponibles dans les bases de données publiques. Cette recherche s'est faite en ne nous basant non pas sur l'annotation automatique mais en réalisant des annotations manuelles. Cette étude a pour but de répondre aux questions suivantes :

- La pression de sélection s'est-elle relâchée sur les gènes impliqués dans la vision, le cycle circadien et la pigmentation chez les organismes cavernicoles ? En même proportion et sur les mêmes gènes en fonction de l'espèce ?
- Peut-on observer une différence en termes de nombre de pseudogènes entre les espèces cavernicoles et les espèces semi-cavernicoles ?
- Comment les pseudogènes peuvent-ils nous donner une indication sur la date de la colonisation des grottes par différentes espèces ?

Afin de mieux comprendre l'évolution de ces gènes, ceux-ci ont également été recherchés dans les génomes d'espèces de surface proches. Les deux prochaines sections établiront le matériel qui sera utilisé pour notre étude, c'est-à-dire les génomes utilisés ainsi que les gènes recherchés.

B- Génomes de poissons cavernicoles et de surface proches disponibles

1) *Astyanax mexicanus*

Le premier génome de poisson cavernicole à avoir été séquencé est le génome d'*Astyanax mexicanus* en 2014 (McGaugh et al. 2014).

Ce characiforme, que l'on trouve essentiellement au Mexique et dans le sud des États-Unis, est de loin l'espèce cavernicole la plus étudiée, notamment parce qu'il est possible de faire des analyses génétiques grâce à l'existence des deux morphotypes interfertiles, de surface et de grotte, comme décrit précédemment.

Au Mexique, le morphotype cavernicole est trouvé dans 34 grottes (Mitchell, Russell, et Elliott 1977; Espinasa et al. 2020) réparties en trois régions que sont la région de Guatemala, la région d'El Abra et la région de Micos (Jeffery 2020) (Figure 49). Le génome assemblé est celui d'un individu de la grotte Pachón. Une des questions majeures relative à l'existence de ces nombreuses populations est le nombre de colonisations indépendantes des grottes qui se sont produites. Une seule colonisation suivie de migrations impliquerait une période d'évolution cavernicole commune à toutes les populations suivies d'évolutions plus ou moins indépendantes. L'hypothèse de plusieurs colonisations quant à elle impliquerait l'apparition des caractères troglomorphes par convergences évolutives de ces différentes populations.

Sur ce sujet, l'article de Bradic et al. (2012) reste aujourd'hui l'étude la plus couramment citée. En génotypant 26 microsatellites chez presque 600 poissons issus de 11 grottes et de 10 points d'échantillonnage dans des rivières avoisinantes, les auteurs estiment qu'il y a eu au moins 5 colonisations indépendantes des grottes : trois colonisations « anciennes » dans la région d'El Abra, dont une dans la grotte Chica, une dans les grottes de Tinaja et Curva et une dans Pachón. Celles-ci auraient été suivies de colonisations plus récentes dans la région de Guatemala (grottes Molino et Caballo Moro) et dans la région de Micos (grotte Subterráneo).

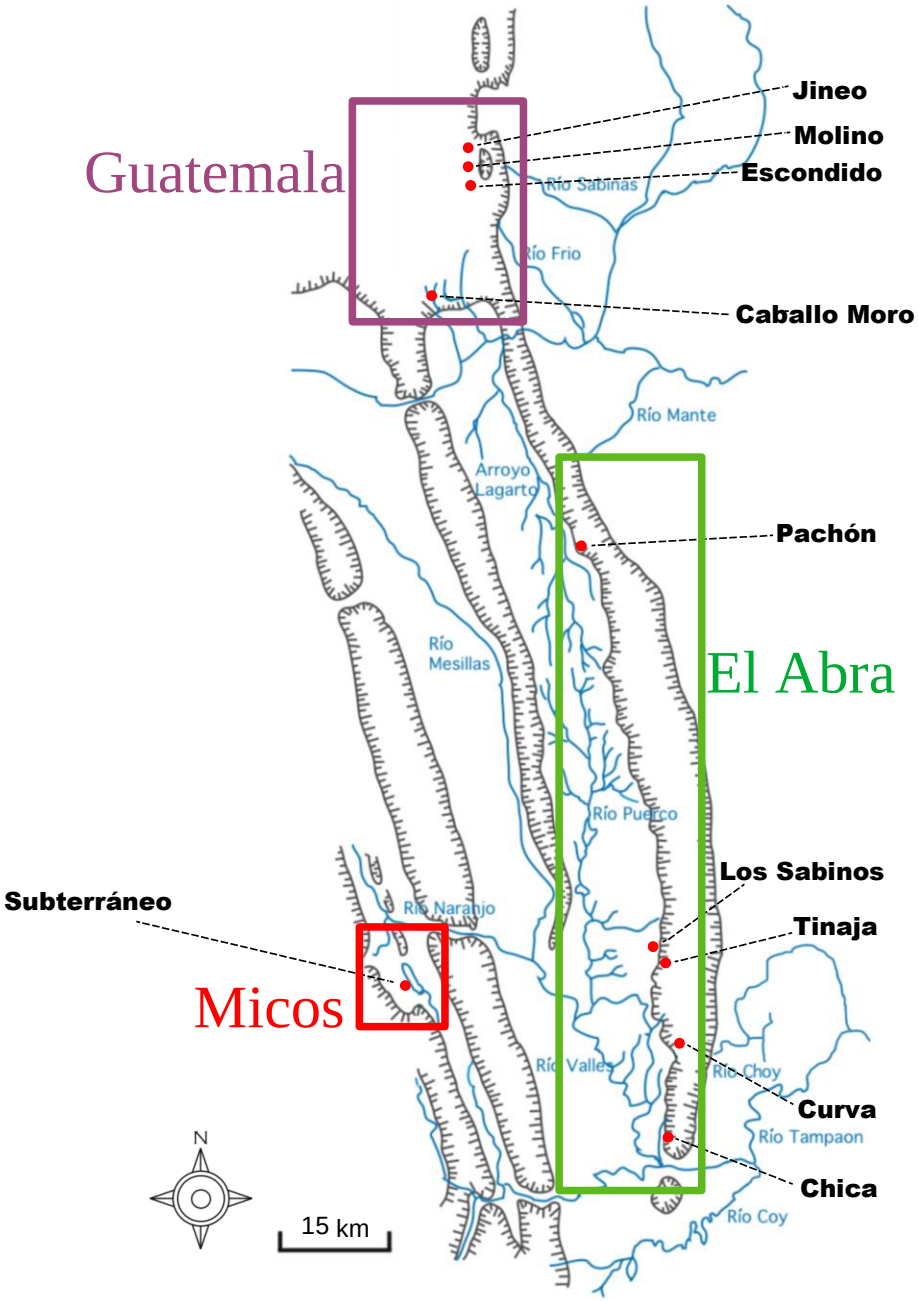


Figure 49 : Carte de répartition des populations cavernicoles d'Astyanax mexicanus. Les trois régions sont encadrées et les grottes les plus étudiées dans la littérature sont indiquées par des points rouges.

Les nombres de pseudogènes trouvés dans ce génome seront comparés aux nombres trouvés dans le génome du morphotype de surface d*'A. mexicanus* ainsi que dans le génome d'une autre espèce de Characiformes : le piranha *Pygocentrus nattereri* (*Figure 50*).

2) *Sinocyclocheilus* spp.

En 2016, deux nouveaux génomes furent séquencés : d'une espèce de poisson cavernicole (*Sinocyclocheilus anshuiensis*) et d'une espèce semi-cavernicole[26] (*Sinocyclocheilus rhinocerous*) (ordre des Cypriniformes) (Yang et al. 2016). Dans la même étude, une espèce pigmentée et avec des yeux fonctionnels du même genre a également été séquencée, *Sinocyclocheilus grahami* (Figure 50 et Figure 51).

Ces espèces qu'on trouve en Chine sont tétraploïdes, résultat d'une duplication entière de leur génome il y a environ 25 million d'années (Xu et al. 2019) *(*Figure 51*).*Avec plus de 75 espèces, ce genre est particulièrement intéressant car des espèces ont colonisé les grottes et ont perdu leur yeux indépendamment au moins trois fois (Mao et al. 2021).

3) *Typhlichthys subterraneus*

Depuis 2018, un génome de l'espèce d'amblyopsidé *Typhlichthys subterraneus* existe dans la base de données NCBI bien qu'aucune publication ne soit associée. Ce percopsiforme a la plus grande répartition géographique parmi les poissons cavernicoles connus, étant retrouvé sur une aire de plus de 14 000 km^2 à l'Est des États-Unis (Figure 51) (Niemiller, Near, et Fitzpatrick 2012). L'espèce de surface avec le génome séquencé la plus proche que nous utiliserons est *Percopsis transmontana* et ces données seront complétées avec une espèce externe, la morue *Gadus morhua* (Figure 50).

[26] Espèce avec un phénotype intermédiaire, entre celui d'un surface et d'un cavernicole. Ces espèces peuvent-être trogloxènes ou troglophiles mais il est souvent difficile d'étudier le mode de vie de ces dernières.

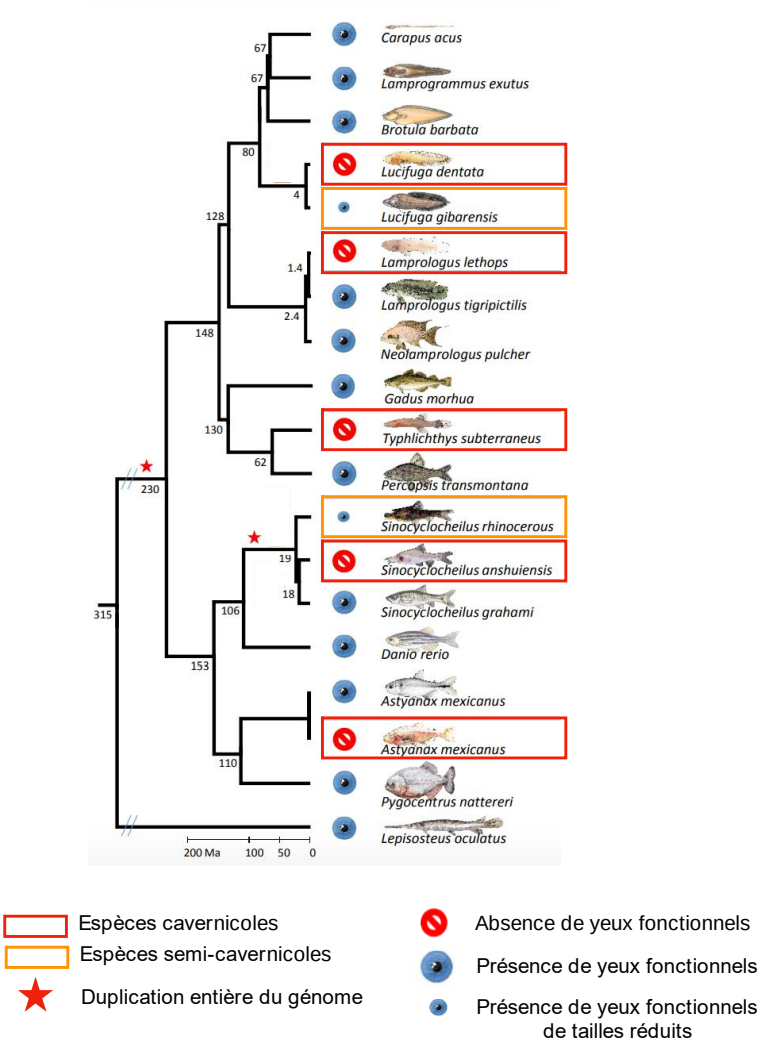


Figure 50 : Phylogénie des espèces chez lesquelles les différents gènes sont recherchés dans notre étude.

4) *Lamprologus lethops* : un poisson aveugle non cavernicole

L'espèce *Lamprologus lethops* (ordre des Cichliformes), a un génome disponible sous forme de lectures depuis 2020 (Aardema, Stiassny, et Alter 2020). Cette espèce, bien qu'aveugle et dépigmentée, n'est pas un poisson cavernicole mais a colonisé un autre environnement privé de lumière, et encore mal compris : les failles profondes du lit du fleuve Congo (Figure 51). Les données issues de ce génome seront comparées avec deux espèces proches et de surface séquencées : *Lamprologus tigripictilis* et *Neolamprologus pulcher* (Figure 50).

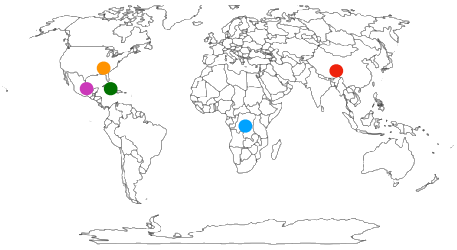
5) Séquençage des génomes de *Lucifuga dentata* et *Lucifuga gibarensis*

Afin d'enrichir ces données, nous avons séquencé deux génomes de poissons cubains : *Lucifuga dentata* qui est une espèce cavernicole et *Lucifuga gibarensis*, espèce semi-cavernicole (ordre des Ophidiiformes)

L'assemblage de *L. dentata* est composé de 52 944 scaffolds pour une taille totale de 634 Mb et avec un scaffold N50[27] d'environ 120 kb. La taille de cet assemblage est cohérente avec la taille du génome d'autres Ophidiiformes. Les tailles de génomes de certaines espèces de ce groupe ont été estimés en mesurant la quantité d'ADN par cellule, et varient de 600 Mb pour l'espèce *Genypterus blacodes* à 800 Mb pour *Ophidion welshi* (Gregory 2021). La production d'un assemblage de bonne qualité pour *L. gibarensis* ayant échoué, les lectures de cette espèce ont été alignées sur le génome de *L. dentata*.

Bien que très fragmenté, l'assemblage du génome de *L. dentata* a une très bonne complétude avec près de 93% des gènes BUSCO trouvés entiers. Cela en fait le génome cavernicole diploïde le plus complet, avec 92% et seulement 57% de gènes BUSCO trouvés chez *Astyanax mexicanus* et *Typhlichthys subterraneus* respectivement (Figure 51).

[27] Le « scaffold N50 » correspond à la longueur médiane des scaffolds dans l'assemblage.



Espèce	Localisation	Yeux	BUSCO
Lucifuga gibarensis		Présents mais petits	Pas d'assemblage
Lucifuga dentata			C:4249 [S:4154, D:95], F:194, M:141, n:4584
Sinocyclocheilus rhinocerous		Présents mais petits	C:4393 [S:1933, D:2460], F:67, M:124, n:4584
Sinocyclocheilus anshuiensis			C:4402 [S:1824, D:2578], F:67, M:115, n:4584
Lamprologus lethops			Pas d'assemblage
Astyanax mexicanus			C:4218 [S:4086, D:132], F:191, M:175, n:4584
Typhlichthys subterraneus			C:2607 [S:2557, D:50], F:973, M:1004, n:4584

[28]

Figure 51 : Localisation, taille des yeux et score BUSCO pour les espèces de téléostéens cavernicoles ou semi-cavernicoles avec un génome disponible, sous forme d'assemblage ou non.

C : Nombre de gènes complets, S : Nombre de gènes complets en une copie (bleu clair), D : Nombre de gènes complets en plusieurs copies (bleu foncé), F : Nombre de gènes incomplets (jaune), M : Nombre de gènes manquants (rouge), n = Nombre de gènes recherchés.

[28] Comme attendu après duplication entière du génome, de nombreux gènes, dont les gènes recherchés par BUSCO, sont retrouvés en deux copies chez les espèces du genre *Sinocyclocheilus.* Cela explique le grand nombre de gènes BUSCO dans la catégorie « D » pour ces espèces (qui signifie complet et en plusieurs copies)

Afin d'avoir une vision plus globale de l'évolution de ces poissons cubains, les gènes de l'étude seront également cherchés chez trois autre espèces d'Ophidiiformes proches du genre *Lucifuga* : *Carapus acus, Brotula barbata* et *Lamprologus exutus* (Figure 50).

C- Construction de trois groupes de gènes

Nous avons défini trois groupes de gènes à partir des connaissances acquises dans la littérature sur l'espèce modèle *Danio rerio*.

1) Gènes de la photoréception

Les gènes de la vision étudiés sont les gènes impliqués dans la cascade de phototransduction ou qui jouent un rôle dans le maintien de la transparence du cristallin (Figure 43). Seuls les gènes dont le domaine d'expression est limité à l'œil ont été sélectionnés. Cette vérification a été effectuée sur la base de donnée ZFIN (Ruzicka et al. 2019) qui regroupe les connaissances actuelles sur la fonction et l'expression des gènes chez le *Danio rerio*. Cette restriction a pour objectif de limiter l'inclusion de gènes pléiotropes, c'est-à-dire des gènes qui ont une fonction autre que celle assurée dans l'œil, et qui de ce fait n'évolueront probablement pas de façon neutre, même chez une espèce vivant dans l'obscurité. Cependant, bien que ces données chez le poisson-zèbre soient actuellement les meilleurs *proxy*s disponibles, il est possible que le territoire d'expression de certains gènes ne soit pas identique chez toutes nos espèces d'intérêt. Le groupe de gène de la photoréception est également composé de toutes les opsines non-visuelles, quel que soit leur domaine d'expression, car leur rôle est en principe uniquement dédié à la réception de lumière.

Au total, ce groupe est composé de 95 gènes.

2) Gènes du cycle circadien et gènes de la pigmentation

Le groupe du cycle circadien est composé de 42 gènes (Figure 44) mais certains sont également impliqués dans la régulation du rythme biologique induits par d'autres *zeitgebers*, potentiellement présents dans les grottes. On s'attend alors à ce que la majorité de ces gènes soit toujours fonctionnelle chez les poissons cavernicoles.

Le groupe de gènes de la pigmentation est composé de 257 gènes, jouant un rôle dans la différenciation des cellules pigmentaires mais également dans la production de mélanine (Lorin et al. 2018). Du fait que la pigmentation est un processus complexe impliquant de très nombreux gènes, il est délicat de définir un groupe de gènes de la pigmentation sans pléiotropie (Baxter et al. 2019). Chez l'humain, les maladies pigmentaires sont souvent associées à d'autres phénotypes et parmi les 128 gènes impliqués dans la pigmentations, 75% sont pléiotropes (Baxter et al. 2019). Ainsi, tout comme les gènes du cycle circadien, il est peu probable de trouver de nombreux pseudogènes chez les espèces cavernicoles dans ce groupe bien qu'elles aient perdu la pigmentation.

D- Méthodes et résultats de la recherche de pseudogènes

1) Nombre de mutations perte de fonction et nombre de pseudogènes

Nous avons utilisé les séquences nucléiques et protéiques du poisson-zèbre pour faire des *blastn* et *tblastn* (Altschul et al. 1990) contre les génomes cavernicoles et de surfaces étudiés, ce qui nous a permis d'identifier les régions où se trouvaient ces gènes chez ces espèces. Ces régions ont ensuite été extraites à l'aide du logiciel *samtools* (Li 2011). Les séquences et structures des gènes ont ensuite été prédites sur ces régions à l'aide d'*Exonerate* (Slater et Birney 2005).

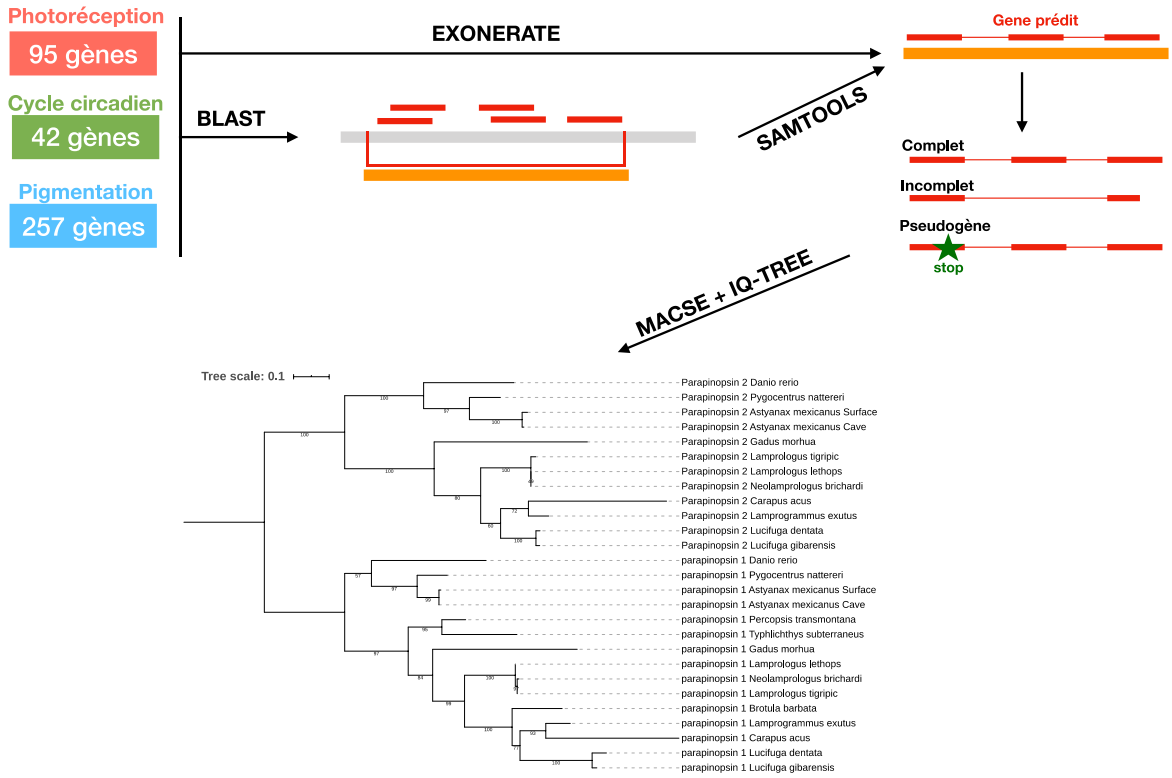


Figure 52 : Méthodologie employée pour la recherche des gènes sur les génomes d'intérêt.

Les gènes prédits ont alors été classés en trois catégories :

- Gènes complets lorsque toute la séquence codante est trouvée intacte ;
- Gènes incomplets lorsque seulement une partie du gène est trouvée ;
- Pseudogènes lorsque le gène, complet ou incomplet, est trouvé avec une ou plusieurs mutations perte de fonction.

Pour les espèces dont le génome n'est pas assemblé mais uniquement sous forme de courtes séquences (également appelées lectures), celles-ci ont été « mappées » avec le logiciel *BWA* (Li et Durbin 2009) contre le génome de l'espèce la plus proche pour trouver la séquence correspondante des gènes d'intérêt.

Une fois les gènes trouvés, les relations d'homologie entre ces derniers ont été confirmées en réalisant des arbres phylogénétiques. Ceci est d'autant plus nécessaire que certains gènes ont des séquences très similaires, et qu'il n'y a pas de discrimination possible lors des étapes de *blast* et d'*exonerate*. Par exemple, chez les espèces étudiées, les gènes *parapinopsin-1* et *parapinopsin-2* (deux opsines non-visuelles) partagent en moyenne 80% de similarité au niveau de leurs séquences protéiques. Lorsqu'on utilise un *blast* avec le gène *parapinopsin-1* contre un génome, un des meilleurs résultats correspondra à la région contenant le gène *parapinopsin-2* et seul un arbre phylogénétique permettra de différencier ces deux gènes (Figure 52). Pour cette étape, les gènes ont d'abord été alignés avec *MACSE* (Ranwez et al. 2011), logiciel qui permet de retirer automatiquement les mutations perte de fonction des séquences, problématiques lors d'un alignement. Les phylogénies ont ensuite été calculées avec *IQ-TREE* (Nguyen et al. 2015) en maximum de vraisemblance avec le meilleur modèle de substitution trouvé par *ModelFinder* (Kalyaanamoorthy et al. 2017).

a) Gènes incomplets

Nous avons retrouvé respectivement 85 et 86 gènes de la photoréception dans les génomes du morphotype cavernicole et du morphotype de surface d'*Astyanax mexicanus*. Parmi les 85 gènes du poisson cavernicole, 19 étaient incomplets avec des exons manquants. Ces parties absentes peuvent avoir pour origine de vraies délétions génomiques mais également une mauvaise qualité de l'assemblage du génome.

Afin d'évaluer la proportion de ces deux facteurs, des amorces de PCR ont été réalisées dans les régions en bordure de ces exons potentiellement absents. Une amplification par PCR a ensuite été réalisée sur des échantillons d'ADN de poissons issus de l'élevage de l'équipe et avec des morphotypes de surface (Texas) ou cavernicoles (Pachón).

Si ces PCR donnent une amplification de taille attendue de l'exon manquant chez les individus cavernicoles, cela signifie que celui-ci est en réalité présent et que son absence dans le génome est le résultat d'une erreur d'assemblage. Au contraire, s'il n'y a pas d'amplification avec l'ADN du morphotype cavernicole alors que l'amplification fonctionne avec celui du morphotype de surface, cela signifie que l'exon manquant est bien le résultat d'une délétion génomique.

Parmi 45 exons manquants répartis dans les 19 gènes incomplets, 36 ont été correctement amplifiés par PCR (amplicons de tailles attendues, et séquences confirmées par séquençage Sanger) chez les individus cavernicoles.

Au contraire, 9 exons amplifiés chez des individus de surface, n'ont pas pu être amplifiés chez les individus cavernicoles (Figure 53).
Ainsi, 80% des exons manquants semblent être des artefacts. En conséquence il a été décidé de ne pas tirer de conclusions quant au nombre de gènes manquants ou incomplets dans les génomes étudiés. Les nombres de pseudogènes qui seront décrits dans les sections suivantes correspondent donc à des sous-estimations des vrais nombres de gènes inactifs.

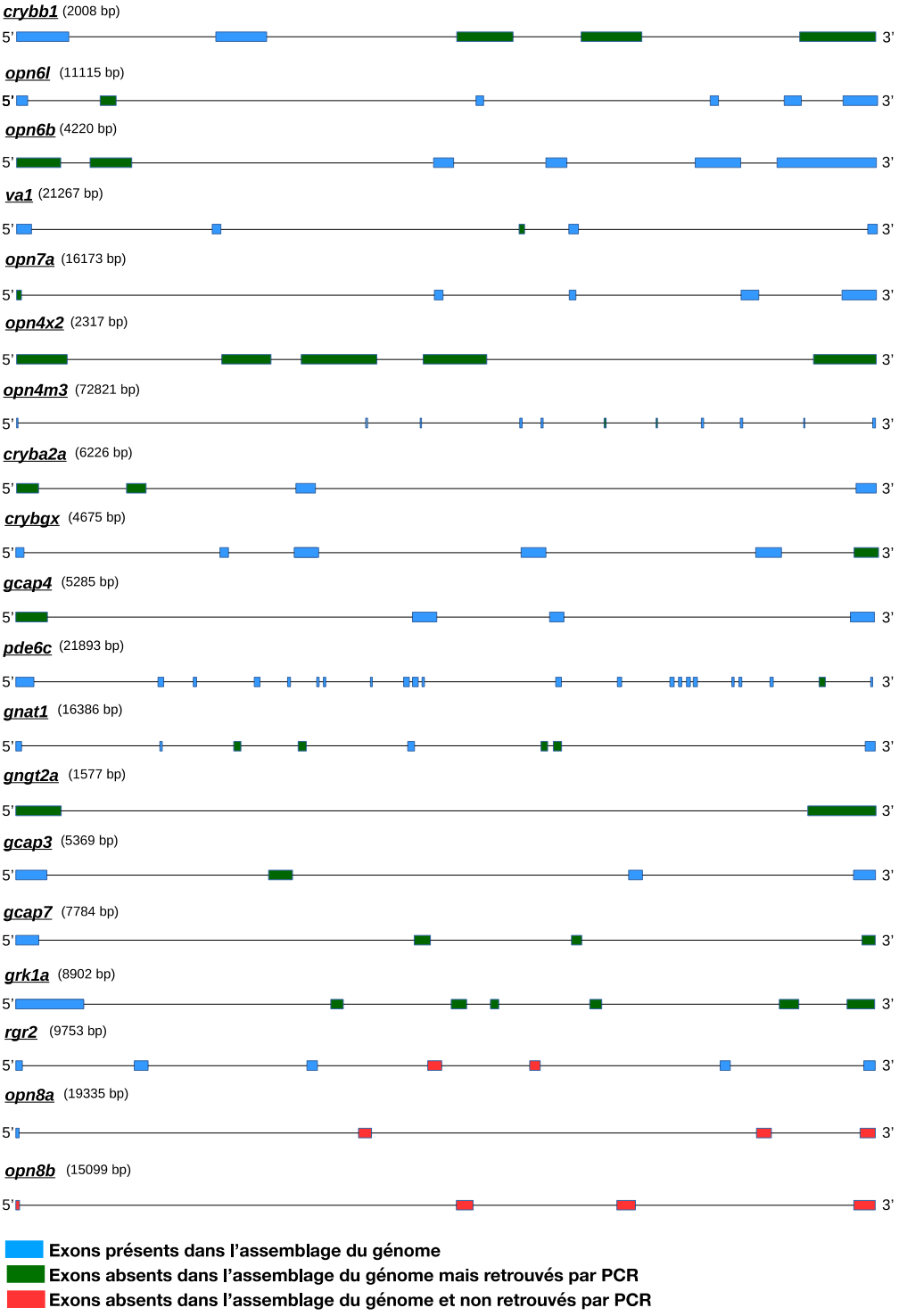


Figure 53 : Exons retrouvés ou non par PCR parmi les gènes incomplets du génome du morphotype cavernicole d'Astyanax mexicanus.

b) Photoréception

- Characiformes

Parmi les 85 gènes de la photoréception trouvés chez le morphotype cavernicole *d'Astyanax mexicanus*, 1 seul gène a été trouvé sous forme de pseudogène, *pde6b*, qui est interrompu par une délétion de 11 paires de base (décalage du cadre de lecture). Aucun pseudogène n'a été trouvé chez le morphotype de surface. *pde6b* n'avait pas été identifié pseudogénisé dans la littérature à cause d'une erreur d'annotation. Effectivement, lorsque l'on regarde ce gène dans la base de données Ensembl (ENSAMXG00000000290, Ensembl 93), le programme d'annotation automatique a inséré un intron d'une paire de base juste avant la délétion, rétablissant le bon cadre de lecture pour la suite du gène (Figure 54).

A

B

19	ENSAMXE00000002965	102,851	102,914	2	0	64	GGCCATGATGATGACAGCTTGTGATCTGTCTGCCATCACCAAGCCATGGGAAGTTCAGAG CAAG
	Intron 19-20	102,915	107,832			4,918	gtgagcttcagcaggctttattatc.....accacggttctgtgatttcctctag
20	ENSAMXE00000002969	107,833	107,907	0	0	75	GTGGCGCTTTCTGTCGCAGCAGAATTTTGGGAGCAGGGTGACTTAGAGAGGACTGTTCTT GAACAGCAACCGATT
	Intron 20-21	107,908	108,760			853	gtgagtatttaatcactataatgtg.....cgtgtttgttggtgtatttctccag
21	ENSAMXE00000002974	108,761	108,799	0	0	39	CCCATGATGGACCGAAACAAAGCAGCAGACCTTCCCAAG
	Intron 21-22	108,800	108,800			1	c
22	ENSAMXE00000002978	108,801	108,833	0	0	33	TTCATAGACTTCGTCTGCACTTTCGTCTACAAG
	Intron 22-23	108,834	108,921			88	gtatttctcagcgtttttctatatt.....gttatattaaactgttcctgtgcag
23	ENSAMXE00000002980	108,922	109,078	0	1	157	GAGTTCTCTCGCTTCCACCCGTCCATCCAGCCCATGTATGACGGCATCCTCAACAACAGG AAGGAGTGGAAGGCGAAGCAGGAGGAGTATGAGGCAAAGCTGAAGGAAGCTGAAGAAGCA GCCAAAGCGAAGCAAGAAGCTGCAGGAGCCAAAAATG
	Intron 23-24	109,079	110,042			964	gtctgaattatacatttttttcatca.....ttctattctttttttttttgtatag
24	ENSAMXE00000002986	110,043	110,101	1	0	59	CCAATGCCAACAGCGTCAACGGAGGTTCAGCCGGATCCAAAACCTGCTCCATCTGCTAG

Figure 54 : Mutation perte de fonction sur pde6b.
A- Exon 22 de pde6b chez Astyanax mexicanus de surface et cavernicole. La région en rouge est délétée chez le cavernicole. B- Erreur d'annotation sur Ensembl.

Des mutations au sein de cette phosphodiestérase sont associées à des maladies chez l'humain, dont l'héméralopie qui cause une difficulté accrue de la vision dans l'obscurité, et la rétinite pigmentaire. Celle-ci réduit également la vision dans l'obscurité ainsi que le champ de vision

(McLaughlin et al. 1993; Gal et al. 1994). Chez la souris, des mutations perte de fonction dans ce gène ont pour conséquence la dégénérescence des bâtonnets pendant le développement et en une absence totale de photorécepteurs à l'âge adulte (Farber et Lolley 1974; B. Chang et al. 2002). L'effet de mutations sur ce gène chez le poisson-zèbre n'est pas encore connu, mais les données collectées sur l'humain et la souris suggèrent que *pde6b* pourrait jouer un rôle dans la dégénérescence de l'œil chez *Astyanax mexicanus*.
Comme pour le gène *Tas1r1* du panda géant cité précédemment, il est difficile de savoir si cette mutation est une cause directe de la dégénérescence des yeux ou si cette mutation n'est apparue qu'ensuite (le gène évoluant de façon neutre une fois les yeux dégénérés), renforçant le processus de dégénérescence. La présence de cette mutation perte de fonction a été confirmée par PCR, chez des individus d'élevage mais également avec des échantillons d'ADN d'individus de populations sauvages.

Chez le piranha, *Pygocentrus nattereri*, un pseudogène a été trouvé parmi 90 gènes de la photoréception. Cette perte n'a probablement aucun impact sur la vision de cette espèce qui dispose de deux yeux fonctionnels, d'autant plus que le gène muté (*lws1*) est une des deux copies issues d'une duplication spécifique à cette espèce. Il est donc probable que la redondance fonctionnelle suite à cette duplication ait permis cette non-fonctionnalisation.

- Ophidiiformes

Chez *Lucifuga dentata* et *Lucifuga gibarensis*, 76 gènes de la photoréception ont été trouvés, en cohérence avec le nombre de gènes trouvés chez les Ophidiiformes de surface : 72 chez *Carapus acus*, 73 chez *Lamprogrammus exutus* et 75 chez *Brotula barbata*. Parmi ces 76 gènes, 19 sont sous forme de pseudogènes chez *L. dentata* (25%) contre seulement 5 chez *L. gibarensis* (Tableau 3). Comme attendu, aucun pseudogène n'a été retrouvé chez *B. barbata* et *C. acus* et seulement un pseudogène est présent chez *L. exutus*.
Parmi les pseudogènes notables chez *Lucifuga dentata*, nous retrouvons la cristalline alpha-A *cryaa*, pour laquelle il a été montré qu'une très forte réduction de son expression pourrait être impliquée dans l'apoptose du cristallin chez *Astyanax mexicanus* (Hinaux et al. 2015). L'inactivation d'une autre cristalline, dans la famille beta (*crybb1*) et avec une mutation dans

un site d'épissage chez *L. dentata*, est responsable de l'opacité du cristallin chez l'humain (Mackay et al. 2002).
Il est également intéressant de noter que le gène *rgr1* partage une mutation perte de fonction chez *L. dentata* et *L. gibarensis* (Tableau 3), indiquant que cette perte remonte à avant leur divergence depuis leur ancêtre commun le plus récent. Chez l'humain, l'apparition de mutations non-synonymes ou perte de fonction dans ce gène sont associées à la rétinite pigmentaire (Morimura et al. 1999), tout comme le gène *gcap2* (Sato et al. 2005), également retrouvé sous forme de pseudogène chez les deux *Lucifuga*, mais dont la perte s'est faite indépendamment (mutations différentes).

Afin de confirmer la non-fonctionnalisation des 19 pseudogènes chez *Lucifuga dentata*, nous avons généré et assemblé un transcriptome de cette espèce. Ce transcriptome est de très bonne qualité, avec 83% des gènes BUSCO trouvés complets, 9% trouvés fragmentés et seulement 8% de gènes manquants. Les transcrits correspondants à ces 19 pseudogènes ont été recherchés dans ce transcriptome et parmi eux, 9 ont été trouvés avec les mêmes mutations perte de fonction que dans l'assemblage du génome, confirmant leur état de pseudogène. Les 10 autres gènes n'ont pas été trouvés dans le transcriptome, probablement parce que ces gènes sont principalement exprimés dans les yeux, structures absentes chez cette espèce. Il est également probable que des mutations se soient accumulées dans les promoteurs de ces gènes, empêchant ou réduisant leur transcription (Tableau 3).

De plus, disposant des lectures génomiques pour les deux espèces de *Lucifuga*, il a été possible de calculer la couverture moyenne des mutations perte de fonction sur les pseudogènes, c'est-à-dire le nombre de fois où la mutation a été séquencée. La grande majorité des mutations ont une très grande couverture (environ 100 pour la plupart) et sont à l'état homozygote. Deux mutations perte de fonction sont à l'état hétérozygote dans le gène *gcap2* de *Lucifuga gibarensis* indiquant que chez cette espèce, le gène existe probablement sous la forme de deux allèles : un fonctionnel et un non fonctionnel (Tableau 3).

Espèce	Gène	Mutation	Couverture
Lucifuga dentata	opn4m3	Stop codon (CGA->TGA)	87
Lucifuga dentata	opn4m3	Frameshift (2bp deletion)	111
Lucifuga dentata	opn7a	Stop codon (CGA->TGA)	104
Lucifuga dentata	opn7b	Stop codon (TCA->TAA)	95
Lucifuga dentata	parapinopsin-1	Stop codon (CAG->TAG)	121
Lucifuga dentata	parietopsin	Frameshift (2bp insertion)	115
Lucifuga dentata	rgr1	Stop codon (CAA->TAA)	101
Lucifuga gibarensis	rgr1	Splice site mutation (Intron2 : GT-AC)	18
Lucifuga dentata *Lucifuga gibarensis*	rgr1	Splice site mutation (Intron4 : GT-TG)	49 33
Lucifuga dentata	tmt1a	Splice site mutation (Intron3 : GC-AC)	95
Lucifuga gibarensis	tmt1a	Splice site mutation (Intron3 : GT-AC)	9
Lucifuga gibarensis	tmt3a	Stop codon (CGA->TGA)	86
Lucifuga dentata	tmt3b	Stop codon (CGA->TGA)	107
Lucifuga dentata	cryaa	Frameshift (4bp deletion)	93
Lucifuga dentata	crybb1	Splice site mutation (Intron2 : TT-AG)	121
Lucifuga dentata	crybgx	Stop codon (CAG->TAG)	92
Lucifuga dentata	crybgx	Splice site mutation (Intron2 : GT-AA)	108
Lucifuga dentata	crygn2	Splice site mutation (Intron3 : GT-GG)	103
Lucifuga dentata	gcap2	Splice site mutation (Intron1 : AT-AG)	116
Lucifuga gibarensis	gcap2	Stop codon (CAG->TAG)	C:45 | T:56
Lucifuga gibarensis	gcap2	Splice site mutation (Intron3 : TG-AG)	TG:65 | GT:58
Lucifuga dentata	grk1b	Frameshift (2bp deletion)	115
Lucifuga dentata	grk7a	Stop codon (CGA->TGA)	117
Lucifuga dentata	pde6ga	Stop codon (AGG->TGA)	92
Lucifuga gibarensis	pde6hb	Frameshift (1bp insertion)	19
Lucifuga dentata	gc2	Frameshift (1bp deletion)	110
Lucifuga dentata	gc3	Stop codon (TGG->TGA)	103
Lucifuga dentata	gnb3b	Stop codon (CGA->TGA)	85

Tableau 3 : Couverture des mutations perte de fonction dans les gènes de la photoréception chez les deux espèces de Lucifuga.
Les mutations en rouge ont été confirmées dans le transcriptome de Lucifuga dentata.

- Percopsiformes

Chez les deux espèces de Percopsiformes, *Typhlichthys subterraneus* et *Percopsis transmontana*, 62 et 69 gènes ont été trouvés dont 27 et 0 pseudogènes respectivement. L'espèce cavernicole *T. subterraneus* a donc près de 44% de ses gènes de phototransduction touchés par des mutations perte de fonction tandis que tous les gènes sont fonctionnels chez l'espèce de surface *P. transmontana.*

De même chez la morue *Gadus morhua*, un Gadiiforme qui nous sert de groupe externe et chez qui parmi 79 gènes trouvés, aucun ne porte de mutations perte de fonction.

Parmi les pseudogènes trouvés chez *T. subterraneus*, on retrouve *cryaa*, comme chez *Lucifuga dentata*. On a donc ici un cas de convergence évolutive où deux espèces cavernicoles ont subi des mutations indépendantes sur le même gène, potentiellement impliqué dans la dégénérescence de l'œil.

- Cichliformes

Les génomes de *Lamprologus lethops* et *Lamprologus tigripictilis* n'étant disponibles que sous forme de lectures, celles-ci ont été « mappées » contre le génome de *Neolamprologus pulcher*, et un même nombre de gènes (79) a été trouvé chez ces trois espèces. Parmi ces 79 gènes, 3, 0 et 2 sous sont forme de pseudogène chez l'espèce de profondeur *Lamprologus lethops*, et les espèces de surface *L. tigripictilis* et *N. pulcher* respectivement.

De façon surprenante, *N. pulcher* a perdu les opsines *lws1* et *sws2*, qui contiennent des décalages de cadre de lecture. L'impact de la perte de *sws2* sur la vision de cette espèce est probablement mineur, ce gène étant présent en 2 copies chez ces espèces de Cichliformes.

De la même façon que pour les *Lucifuga*, il a été possible de quantifier la couverture des mutations perte de fonction chez *L. lethops*. Le pseudogène *tmt1a* est touché par deux mutations, une à l'état homozygote et l'autre à l'état hétérozygote tandis que les autres mutations sont toutes à l'état homozygote (Tableau 4).

Espèce	Gène	Scaffold	Position	DP4	Genotype
Lamprologus lethops	crybgx	NW_006272218.1	642367	1,5,23,26	1/1
Lamprologus lethops	opn4x1	NW_006272015.1	5259668	0,0,26,50	1/1
Lamprologus lethops	tmt1a	NW_006272087.1	1747364	3,0,30,15	1/1
Lamprologus lethops	tmt1a	NW_006272087.1	1747251	5,7,11,16	0/1

Tableau 4 : Couverture des mutation perte de fonction sur les gènes de la photoréception chez Lamprologus lethops.
DP4 : Nombre de lectures portant la mutation perte de fonction dans les deux sens de lecture suivi du nombre de lectures ne portant pas la mutation dans les deux sens de lecture. Un génotype 1/1 indique que la mutation est à l'état homozygote, et 0/1 indique que la mutation est à l'état hétérozygote.

Ces différentes mutations avaient déjà été rapportées dans la littérature (Aardema, Stiassny, et Alter 2020), dans laquelle un quatrième pseudogène, *opn3*, était également trouvé. En examinant de plus près ce gène, il s'est avéré que les auteurs de la publication avaient fait une erreur. La mutation rapportée se situe dans une région non-codante du génome (NW_006272062.1 :3187434), et sur un scaffold différent de celui correspondant vraiment au gène *opn3* (NW_006272060.1).

- Cypriniformes

Les *Sinocyclocheilus*, tétraploïdes, ont environ deux fois plus de gènes que les autres espèces étudiées, avec 173, 171 et 169 gènes de la photoréception chez *Sinocyclocheilus grahami*, *Sinocyclocheilus anshuiensis* et *Sinocyclocheilus rhinocerous* respectivement. Après une duplication entière du génome, les paires de gènes produites sont appelées des ohnologues (d'après le généticien Susumu Ohno). Ici, une fonction sera donc considérée comme perdue seulement si les deux ohnologues sont trouvés sous forme de pseudogènes.
Nous avons trouvé 18 et 32 pseudogènes respectivement chez *S. grahami* et *S. rhinocerous*, mais seulement une paire d'ohnologues est devenue non fonctionnelle chez ces deux espèces. Chez le poisson cavernicole *S. anshuiensis*, le nombre de pseudogènes est plus grand (48) et 7 paires d'ohnologues sont non fonctionnelles.
Par la présence de mutations perte de fonction communes entre ces espèces, nous savons que 2 pseudogènes sont apparus avant la divergence de ces trois espèces depuis leur ancêtre commun le plus récent et 2 pseudogènes sont apparus avant l'ancêtre commun de *S. grahami* et *S. anshuiensi* mais après la séparation de *S. rhinocerous.*

Ainsi, chez tous les organismes étudiés, la même tendance émerge : les poissons cavernicoles ont systématiquement une proportion de pseudogènes plus importante que les espèces de surface les plus proches, et les espèces semi-cavernicoles sont entre les deux. Les espèces cavernicoles qui ont perdu le plus de gènes sont *T. subteranneus* avec 40 % de pseudogènes suivi de *L. dentata* avec 25 % (Figure 55).

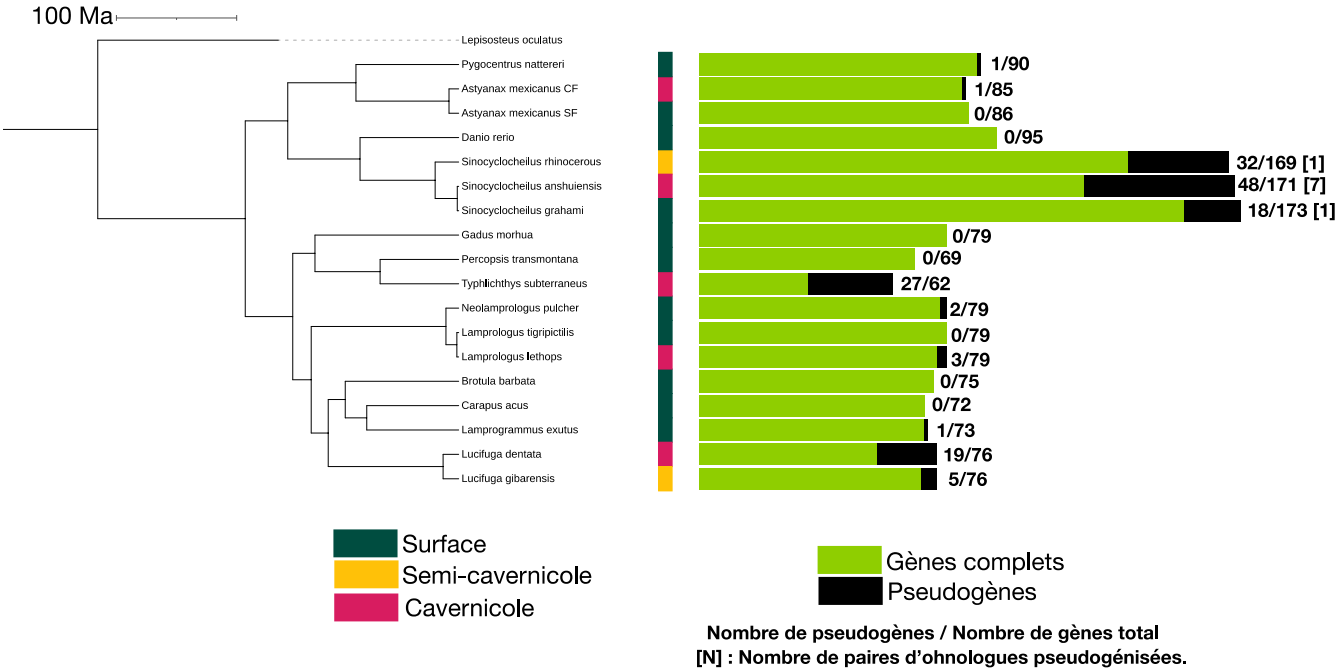


Figure 55 : Phylogénie des espèces étudiées avec les nombres de gènes et de pseudogènes de la photoréception trouvés.

c) Cycle circadien

Parmi les 42 gènes du cycle circadien, un seul pseudogène a été trouvé chez les espèces de surface : *cry-dash* contient un décalage du cadre de lecture chez *L. tigripictilis*, mais cette mutation est à l'état hétérozygote, signifiant qu'une version fonctionnelle du gène existe.

Chez les poissons cavernicoles et semi-cavernicoles, 4 gènes existent sous forme inactivé : *cry1b* et *cry2a*, pour lesquels les deux paires d'ohnologues sont pseudogénisées chez *S. rhinocerous*, *cry-dash* avec les deux gènes ohnologues perdus chez *S. rhinocerous* et *S. anshuiensisi*, et *per2* également perdu chez *S. rhinocerous* et chez le poisson cavernicole *T. subterraneus*. La Figure 56 indique le nombre de gènes et de pseudogènes trouvés chez chaque espèce.

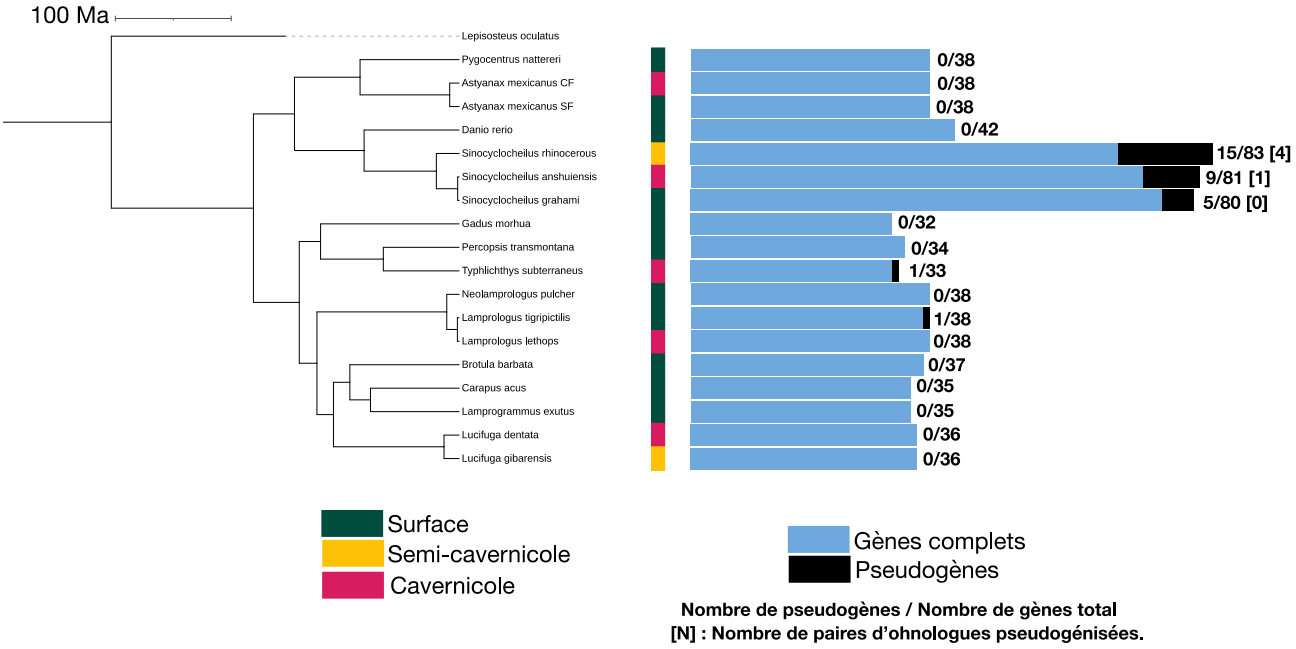


Figure 56: Phylogénie des espèces étudiées avec les nombres de gènes et de pseudogènes du cycle circadien trouvés.

La protéine *cry-dash* est impliquée dans la réparation des dommages sur l'ADN causés par les rayons ultra-violets de la lumière du soleil. Elle est elle-même dépendante de la lumière pour être active. Chez les poissons cavernicoles, l'inactivation de ce gène a été rapporté pour la première fois chez *Phreatichthys andruzzii* (Zhao et al. 2018). Cette espèce a également perdu le gène *per2*, probablement un facteur responsable de la perte du cycle circadien chez cette espèce (Ceinos et al. 2018).

d) Pigmentation

Chez le morphotype cavernicole d'*A. mexicanus*, deux pseudogènes dont un déjà rapporté dans la littérature ont été trouvés parmi 249 gènes de pigmentation : *mc1r* et *tyrp1a* (Figure 57). Bien que non fonctionnel chez *A. mexicanus*, *oca2* n'est pas considéré dans notre étude car ce gène est tronqué par une grande délétion génomique, et non par une des mutations perte de

fonction étudiées. Or, comme mentionné auparavant, la qualité des génomes étant très variable dans cette étude, ces délétions n'ont pas été prises en compte dans nos analyses (Figure 53).
La perte de *oca2* est suffisant à faire perdre la pigmentation chez cette espèce car son inactivation par manipulation génétique chez le morphotype de surface produit des individus albinos (Klaassen et al. 2018). Par ailleurs, cette perte de fonction pourrait avoir un avantage sélectif. Effectivement, ce gène est impliqué dans la voie de synthèse de la mélanine à partir de la molécule L-tyrosine, qui est également une molécule nécessaire pour la production de neurotransmetteurs (catécholamines). En l'absence d'*oca2*, il n'y a plus de production de mélanine et le niveau de L-tyrosine s'accroît, augmentant la production de catécholamines. Celles-ci joueraient alors un rôle dans les changements comportementaux observés chez le morphotype cavernicole, tels que la perte du sommeil et l'augmentation du stress (Bilandžija et al. 2018), qui semblent être avantageux dans les grottes.
Les conséquences des pertes de *mc1r* et *tyrp1a* dans la perte de pigmentation restent floues, étant donné que *mc1r* est dupliqué chez cette espèce avec une copie fonctionnelle qui existe toujours chez le morphotype cavernicole. De plus, la réduction de l'expression de *tyrp1a* chez le poisson-zèbre n'a aucun impact, du moins visible, sur son phénotype (Krauss et al. 2014).

Une très faible proportion de pseudogènes a été trouvée chez les poissons cavernicoles cubains, avec respectivement 8 et 7 pseudogènes chez *L. dentata* et *L. gibarensis* (parmi 237 gènes) (Figure 57). Tout comme le gène de la photoréception *rgr1*, le gène de la pigmentation *adamts20* a subi une mutation perte de fonction avant leur MRCA, tandis que deux autres gènes, *smtla* et *myo7ab* ont été perdus indépendamment chez ces deux espèces. Il est très intéressant de noter la perte des gènes *trpm1a* et *trpm1b*, pseudogénisés chez *L. dentata*. La mutation de l'orthologue de ces gènes (*TRPM1*) chez l'humain provoque une cécité nocturne, également appelée syndrome d'Oguchi (Audo et al. 2009). Cela souligne le caractère pléiotrope du groupe de gènes de la pigmentation, avec ici des gènes impliqués dans la vision et dans la pigmentation évoluant de façon neutre puisque ces deux caractères sont perdus.

Parmi 484, 487 et 490 gènes de la pigmentation trouvés chez *S. grahami*, *S. rhinocerous* et *S. anshuiensis* respectivement, seulement 15, 35 et 28 pseudogènes ont été trouvés (Figure 57). Tandis qu'aucune paire d'ohnologues n'est perdue chez le poisson de surface *S. grahami*, il a été observé que les deux mêmes paires ont été perdues indépendamment chez *S. rhinocerous* et

S. anshuiensis (*gch2* et *pmelb*). Avec seulement 3%, 6% et 7% de pseudogènes chez ces trois espèces, ces gènes sont ceux dont la proportion de rétention est la plus importante après la duplication entière du génome (comparé à 10%, 19% et 28% pour les gènes de la vision et 6%, 18% et 11% pour les gènes du cycle circadien). La même tendance a été observée chez les salmonidés[29], qui sont également tétraploïdes, et chez qui les gènes impliqués dans la pigmentation ont les plus forts taux de rétention (Lorin et al. 2018).

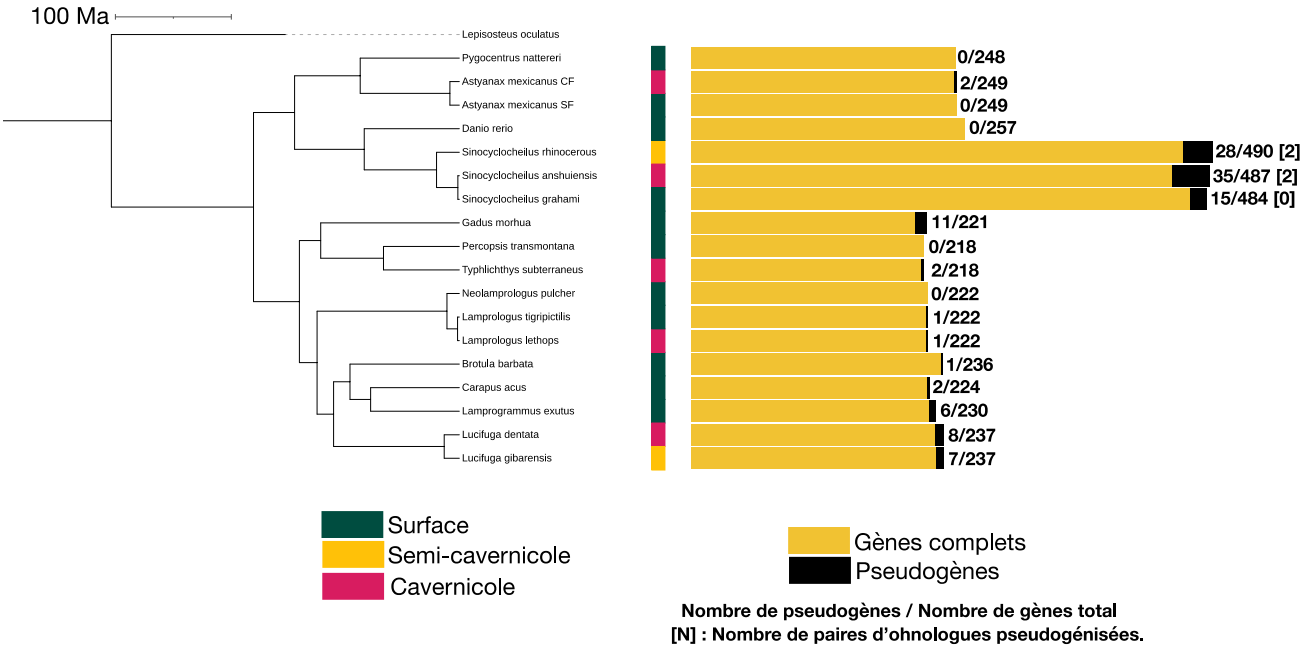


Figure 57 : Phylogénie des espèces étudiées avec les nombres de gènes et de pseudogènes de la pigmentation trouvés.

Enfin, chez les deux poissons troglomorphes *L. lethops* et *T. subterraneus*, un et deux gènes sont impactés respectivement par des mutations perte de fonction (sur un total de 222 et 218 gènes de la pigmentation trouvés, Figure 57). On retrouve, encore une fois, l'apparition de pertes indépendantes chez plusieurs poissons cavernicoles, notamment avec *oca2* qui a perdu

[29] Les gènes de la pigmentation montrent également un plus fort taux de rétention suite à la duplication du génome partagée par les téléostéens (TGD)

son codon start chez *L. lethops* (gène également perdu chez *A. mexicanus*) et *trpm1b* qui est touché par trois décalages du cadre de lecture et un codon stop prématuré chez *T. subterraneus* (gène également perdu chez *L. dentata*). De façon surprenante, l'espèce chez qui le nombre de pseudogènes de la pigmentation est le plus grand est une espèce de surface, la morue de l'Atlantique *Gadus morhua*, avec 11 gènes impactés par des mutations perte de fonction parmi 221 gènes.

2) Nombre de mutations perte de fonction et leur distribution

Chez les espèces étudiées, un total de 134 codons stop prématurés, 199 décalages du cadre de lecture (répartis en 121 délétions et 78 insertions), 6 pertes de codon stop, 16 pertes de codon start et 43 mutations sur des sites d'épissage ont été observés.

a) Distribution de la taille des indels

La grande majorité des décalages du cadre de lecture (frameshifts) est le résultat de délétions ou d'insertions (= indels) d'une ou deux bases, et dans de rares cas, ceux-ci ont des tailles plus grandes. La plus grande délétion trouvée mesure 83 paires de bases dans le gène de la pigmentation *zic2b* chez *S. rhinocerous*, tandis que la plus grande insertion, de 20 paires de bases, est trouvée chez *S. anshuiensis* dans l'opsine non-visuelle *opn7b* (Figure 58). L'observation d'un plus grand nombre de délétions que d'insertions et d'une telle distribution de leurs tailles est cohérente avec les observations rapportées dans la littérature (Fan et al. 2007; Miles et al. 2016).

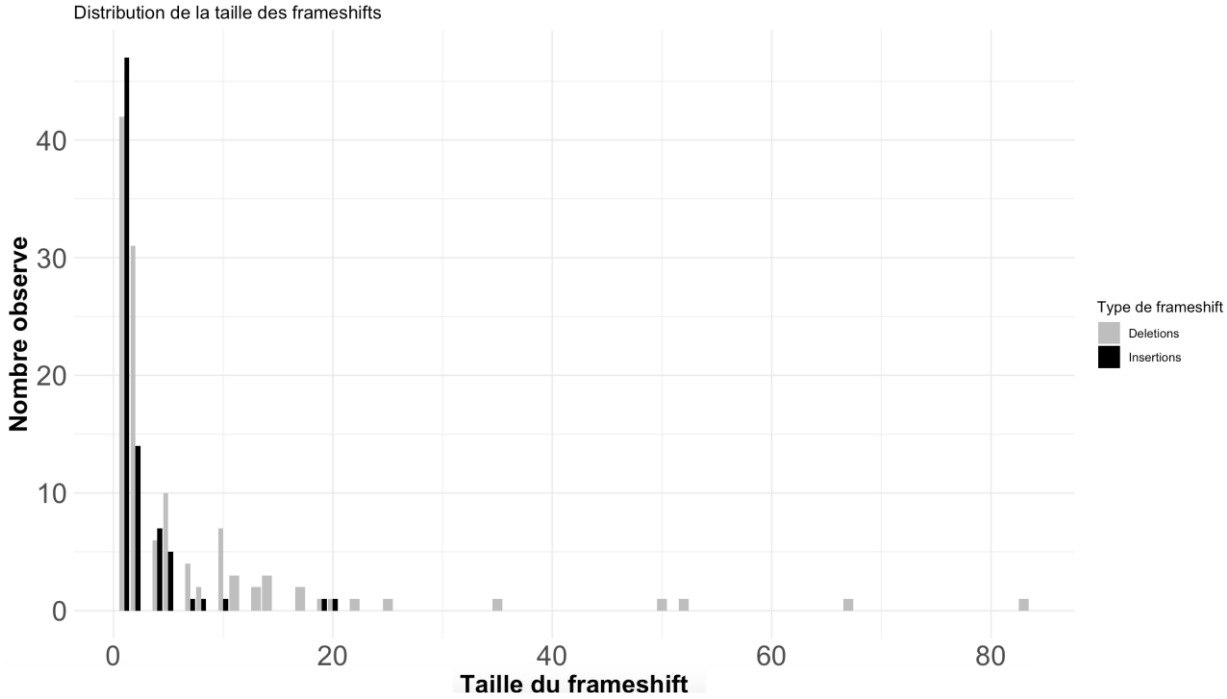


Figure 58 : Distribution de la taille des insertions ou délétions provoquant un décalage du cadre de lecture (=frameshift) chez les espèces étudiées.

b) Fréquence théorique de chaque type de mutation perte de fonction

Il est possible de calculer la fréquence théorique attendue des mutations perte de fonction. Effectivement, le code génétique est composé de 64 codons, 61 codant pour des acides aminés et les trois restants signalant l'arrêt de la traduction (TGA, TAG ou TAA). Sur 549 mutations possibles (61 codons x 9 mutations possible par codon), 23 transforment un codon codant pour un acide aminé en codon stop. Cela représente 4% des mutations, en considérant que chaque codon est présent à la même fréquence et qu'il y a autant de transitions que de transversions[30]. Bien sûr, en réalité, dans les différents gènes et espèces étudiées, ces deux hypothèses ne sont pas vérifiées. Avec un taux de transition/transversion moyen observé chez nos espèces de 2,5[31],

[30] Transition : mutations d'une purine à une autre ou d'une pyrimidine à une autre (C ←→ T ou A ←→ G)
Transversion : autres mutations
[31] Calculé avec le programme PAML (Z. Yang 2007)

et en prenant en compte la fréquence des codons dans les séquences des gènes de la vision, du cycle circadien et de la pigmentation, la fréquence théorique d'apparition d'un codon stop est environ 10% plus faible, soit 3,6% des mutations.

Avec un taux de mutation par site égal à μ, le taux de mutation vers un codon stop est donc $\mu_{STOP} = 0{,}036\mu$.

Il n'existe pas de taux de mutation théorique pour l'apparition d'insertions ou de délétions provoquant un décalage du cadre de lecture chez les vertébrés, car leur probabilité d'apparition semble dépendre fortement du contexte génomique. Une étude de 2008 sur les petites insertions et délétions chez des espèces de drosophiles et chez des primates ont montré que le taux d'apparition de celles-ci pouvait varier d'un facteur 100 en fonction de la séquence (Tanay et Siggia 2008). Ainsi, la fréquence théorique d'apparition de décalages du cadre de lecture chez nos espèces est directement calculée à partir des observations, et liée au nombre de codons stop prématurés. Avec Nf = Nombre de décalages du cadre de lecture observés et Ns = Nombre de codon stop prématurés observés, on a $\mu_{frameshift} = \frac{Nf}{Ns}\mu_{STOP} = \frac{Nf}{Ns}0{,}036\mu = 0{,}05\mu$.

Le taux de perte du codon d'initiation (ATG, également appelé codon start) dépend uniquement du nombre de gènes et de leur longueur cumulée. Effectivement, toutes les mutations sur ce codon provoquent sa disparition, et ce dernier ne représente que trois sites parmi tous les sites du gène. Ainsi, avec N_g = Nombre de gènes et L = longueur cumulée des gènes, ce taux est $\mu_{perte\_start} = \frac{3N_g}{L}\mu = \frac{3 \text{ x } 7431}{12385000}\mu = 0{,}0018\mu$.

Le taux de perte de codon de terminaison (TAG, TAA ou TGA) est un peu plus faible, car une mutation dans un codon stop peut amener un autre codon stop (Par exemple, un TAG peut se transformer en TAA avec une seule mutation). Ainsi, avec 9 mutations possibles par codon, 27 mutations sont possibles sur les trois codons de terminaison. Quatre de ces mutations transforment un codon stop en un autre, résultant en 24 mutations sur 27 (0,89) possibles pour faire disparaître ce dernier. Cela amène l'équation suivante $\mu_{perte\_stop} = \mu_{perte\_start} \text{ x } \frac{24}{27} = 0{,}0016\mu$.

Quatre sites d'épissages étant présents par intron, le taux d'apparition d'une mutation sur un de ces sites dépend du nombre d'introns dans les gènes et de la taille cumulée de ces séquences.

Avec Ni = Nombre cumulé d'introns dans les gènes étudiés, on obtient $\mu_{site\_epissage} = \frac{4Ni}{L}\mu = \frac{4 \text{ x } 68117}{12385000}\mu = 0{,}022\mu$.

Avec ces différents taux de mutation perte de fonction, on s'attend à trouver 32% de codons stop prématurés, 45% de décalages du cadre de lecture, 20% de mutations sur un site d'épissage, 1,6% de pertes de codon start et 1,4% de pertes de codon stop. Cette distribution théorique (et bien que la proportion relative de décalages du cadre de lecture par rapport à celle des codons stop prématurés soit celle observée par construction) est très similaire à la distribution observée (Figure 59).

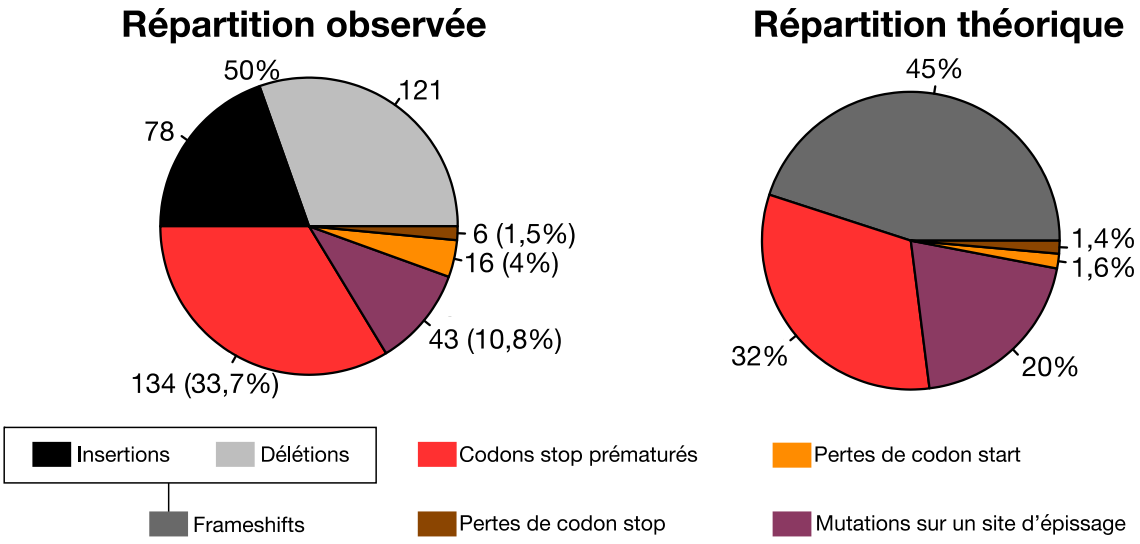


Figure 59 : Fréquences observées et théoriques des mutations perte de fonction étudiées. Pour la répartition théorique, la différence n'est pas faite entre délétions et insertions, ces deux catégories étant additionnées dans le groupe des frameshifts.

c) Position relative des mutations perte de fonction le long des séquences

La répartition des mutations perte de fonction le long des séquences doit être examinée afin de vérifier que celles-ci ne sont pas regroupées à l'extrémité 3' des gènes, où leur impact fonctionnel pourrait être faible. La méthode pour étudier si ces mutations sont bien placées au hasard le long des séquences est inspirée d'une étude sur l'insertion des introns dans les gènes au cours de l'évolution (Lynch et Kewalramani 2003). Celle-ci consiste à calculer la taille effective des segments de gènes séparés par les introns, ou par les mutations dans notre étude. La taille effective Ns de ces segments est : $Ns = \frac{1}{\sum_{i=1}^{n} S_i^2}$ avec n = nombre de segments séparés par n -1 mutations et S_i= taille du segment i. Les gènes étant de différentes tailles, les positions relatives des mutations dans les séquences ont été calculées (position de la mutation / taille du gène). Ainsi la somme de S_i pour chaque gène est égale à un. Si toutes les mutations sont réparties de manière uniforme le long des gènes, tous les segments auront la même taille $\frac{1}{n}$, donnant $Ns = n$ (Figure 60, A). En revanche, si les mutations sont très regroupées à une extrémité du gène, un segment aura une taille proche de 1 tandis que les autres auront des tailles proches de 0, donnant $Ns = 1$ (Figure 60, B).

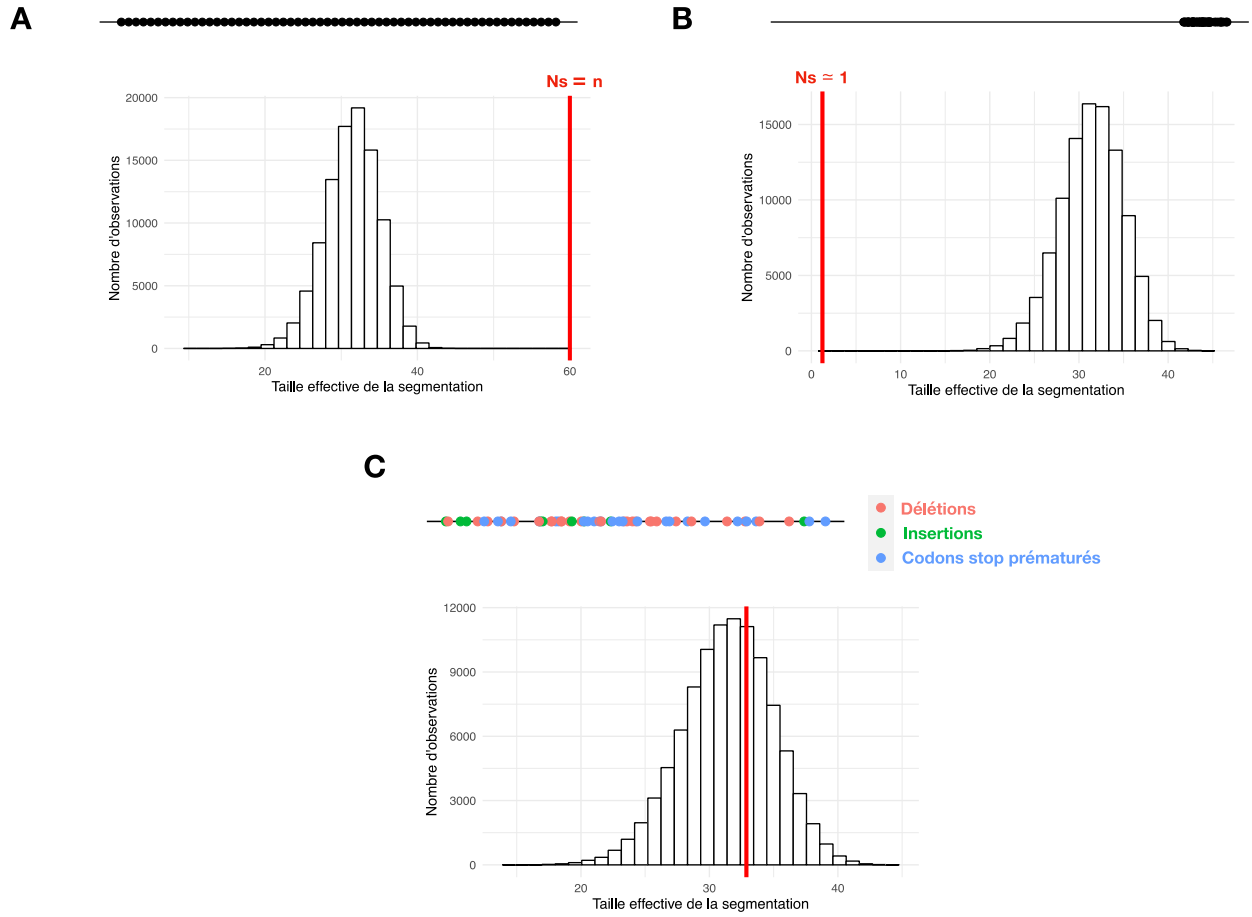


Figure 60 : Distribution théorique des tailles effectives de segmentation issues de 100 000 simulations, où 60 mutations sont réparties au hasard le long des séquences. Les barres rouges représentent la taille effective de segmentation observée dans différents scénarios : A- Répartition uniforme des mutations le long des gènes B- Répartition regroupée à la fin des gènes C- Répartition observée des mutations perte de fonction dans les gènes de la photoréception des espèces cavernicoles diploïdes[32].

Afin d'obtenir la distribution théorique de la taille effective de la segmentation *Ns* lorsque des mutations sont réparties au hasard le long des gènes, 100 000 simulations de répartition au hasard ont été réalisées, avec le même nombre de mutations que le nombre observé. Cela a permis de montrer que chez les espèces et gènes étudiés, les mutations perte de fonction sont

[32] Les mutations des autres espèces de l'étude ne sont pas prises en compte pour une question de lisibilité du graphique. De plus, les pertes de codon start ou pertes de codon stop ne sont pas prises en compte, car localisées aux extrémités des gènes de par leur nature.

bien réparties au hasard dans les séquences, le Ns observé étant au centre de la distribution des Ns simulés (Figure 60, C).

E- Gènes mutés dans les trois groupes

Afin de compléter notre jeu de données, nous avons collecté les résultats d'analyses par gènes candidats dans la littérature sur d'autres espèces cavernicoles. L'orthologue du gène *rh1.1* est inactivé, de façon indépendante, chez plusieurs espèces d'*Amblyopsidae* cavernicoles (ordre des Percopsiformes), notamment chez *Troglichtys rosae*, *Amblyopsis spelaea* et, comme le confirme notre analyse génomique, chez *Typhlichthys subterraneus* (Niemiller et al. 2013).
Chez *Phreatichthys andruzzii*, les opsines non-visuelles *opn4m1*, *opn4m2* et *tmt3a* sont inactivées, ainsi que les gènes du cycle circadien *cry-dash* et *per2* (Cavallari et al. 2011; Kumar 2017; Haiyu Zhao et al. 2018) . Enfin, le gène de la pigmentation *mc1r* est pseudogénisé chez *Oreonectes daqikongensis* (Liu et al. 2019).

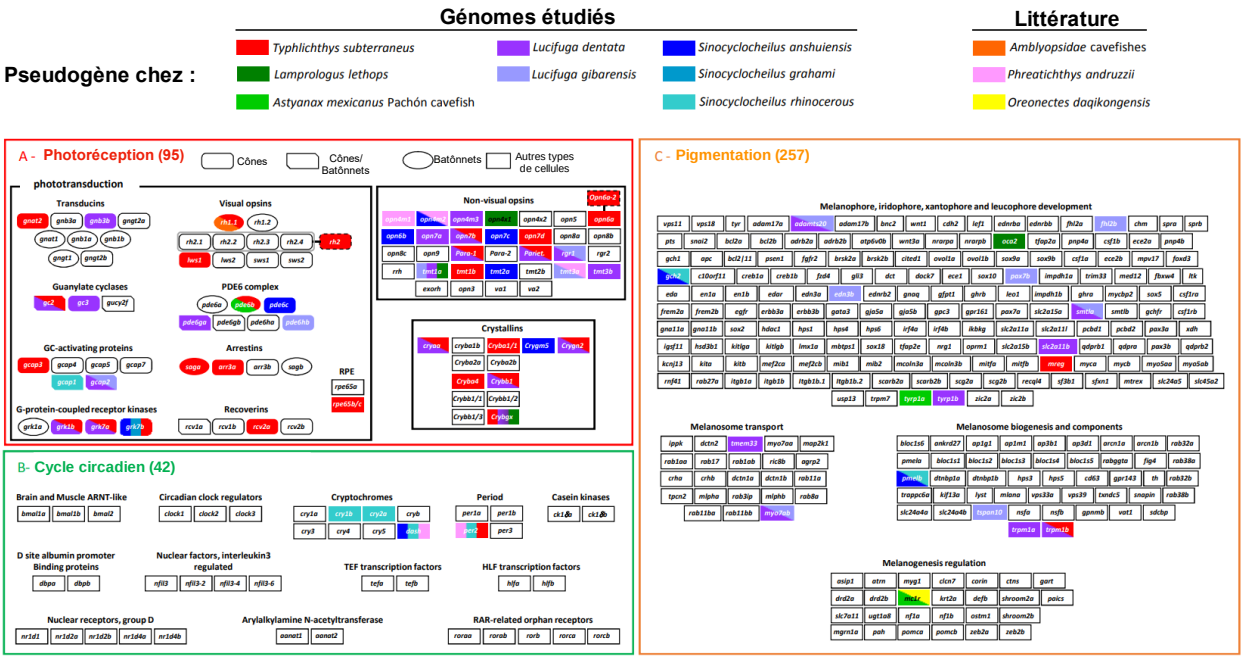


Figure 61 : Groupes de gènes étudiés et pseudogènes observés chez les espèces cavernicoles. (A) Gènes de la photoréception. (B) Gènes du cycle circadien. (C) Gènes de la pigmentation. Le nombre de gènes chez le poisson-zèbre est indiqué entre parenthèse pour les trois groupes. Chaque carré représente un gène, à l'exception de rpe65b/c qui représente une duplication spécifique au poisson-zèbre. De plus, les gènes rh2 sont en 4 copies chez le poisson-zèbre contrairement aux autres espèces chez qui il n'existe qu'une copie, et le gène opn6a est dupliqué chez certains poissons étudiés. Les gènes sont colorés en fonction de l'espèce chez qui ils sont retrouvés sous forme de pseudogène. Un gène non coloré n'est jamais retrouvé inactivé chez ces espèces.

Parmi les 257 gènes de la pigmentation recherchés, seulement 18 sont touchés par des mutations perte de fonction chez les différents organismes cavernicoles (Figure 61), soit 7%. De plus, une perte récurrente des mêmes gènes est observée, avec 6 de ces gènes perdus indépendamment[33] chez différents poissons cavernicoles. Ces deux constatations semblent confirmer l'hypothèse que la grande majorité des gènes de ce groupe sont maintenus par sélection, même chez des organismes dépigmentés, dû à leur pléiotropie. Ainsi, seule une sous-partie de ces gènes peut être perdue, ce qui explique que les mêmes pseudogènes sont observés de façon récurrente.

[33] *adamts20* n'est pas compté dans les six gènes car uniquement perdu dans l'ancêtre commun de *L. dentata* et *L. gibarensis.*

Dans le groupe des gènes du cycle circadien, uniquement 4 pseudogènes, répartis dans deux familles ont été trouvés (familles des cryptochromes et period, Figure 61), et ces deux familles ont une expression induite par la lumière (Figure 44). De la même manière que pour le groupe de la pigmentation, il semble que seulement une sous partie de ces gènes ait la possibilité d'être perdue, avec la perte indépendante de *cry-dash* et *per2* chez trois espèces cavernicoles. De façon surprenante, il semble ne pas y avoir de corrélation entre le niveau de régression de la vision et la perte des gènes du cycle circadien, étant donné que *S. rhinocerous*, qui possède toujours des yeux fonctionnels, est l'espèce ayant perdu le plus de gènes (4), indiquant une perte probable de son cycle circadien.
Il faut néanmoins souligner que les opsines non-visuelles, dans le groupe des gènes de la photoréception, jouent également un rôle majeur dans le cycle circadien. Leurs pertes chez les différents organismes peuvent donc également avoir un impact sur ce cycle. Cela est notamment connu pour le gène *tmt3a*, impliqué dans la perte du cycle circadien chez *P. andruzzii* et qui est également perdu chez *L. gibarensis* (qui est, comme *S. rhinocerous*, équipé de petits yeux).

Au contraire des groupes de gènes précédents, dans le groupe des gènes de la photoréception, il y a un nombre important de gènes qui ont été perdus chez au moins une espèce : 46 gènes sur 95, ce qui représente un peu plus de 48% (Figure 61). Une grande hétérogénéité du nombre de gènes inactivés existe au sein des différents poissons cavernicoles étudiés. Il ne semble pas y avoir de corrélation directe entre le nombre de pseudogènes et la régression de l'œil. Par exemple, chez le poisson semi-cavernicole *L. gibarensis*, 5 pseudogènes ont été trouvés tandis que le poisson sans yeux et cavernicole *A. mexicanus* possède uniquement 1 pseudogène. Il faut souligner que de nombreux gènes, incluant des gènes de développement qui ne sont pas présents dans les trois groupes étudiés, peuvent potentiellement être responsables de la dégénérescence de l'œil.
Ainsi, une fois les grottes colonisées par une espèce, la vitesse à laquelle celle-ci va perdre ses yeux va dépendre d'un processus aléatoire : une espèce pourra accumuler plusieurs pseudogènes de la vision sans que ces dernières ne provoquent une perte de structure, tandis que les premières mutations chez une autre espèce pourront provoquer cette perte.
De plus, bien que nous fassions l'hypothèse de neutralité, il est possible que les premières mutations impliquées dans la perte de l'œil se fixent dans les populations cavernicoles par sélection naturelle. Effectivement, la présence de l'œil est très coûteuse en énergie (Moran,

Softley, et Warrant 2015). A l'exception de quelques gènes dans le groupe de photoréception, l'inactivation de ces derniers ne devrait cependant pas impliquer la dégénérescence des yeux, leur rôle étant surtout dans la phototransduction ou le relais du signal et non dans la mise en place de structures ou leur maintien.

La pseudogénisation de presque la moitié des gènes de ce groupe soulève la question de savoir si les gènes non mutés sont des gènes avec des effets pléiotropes non-connus, les rendant essentiels même chez les poissons cavernicoles, ou si l'absence de mutations perte de fonction dans ces derniers est seulement la conséquence d'un manque de temps d'évolution, leur apparition et fixation étant un processus stochastique. La première hypothèse est peu probable car ce groupe de gènes a été établi de façon à éviter la pléiotropie, et de nombreux gènes qui ne sont pas retrouvés sous forme de pseudogènes sont sous-exprimés voire pas exprimés du tout chez certains poissons cavernicoles. Cela peut être une marque d'inactivation ou tout du moins d'un début d'inactivation de ces gènes, sans mutations dans les séquences codantes mais probablement dans des régions régulatrices (Tableau 5).

Gène	Espèces cavernicoles	Références
lws1	*P.mexicana, S. tileihornes, A.mexicanus*	(Tobler et al. 2010), (Meng, Zhao, et al. 2013), (Simon et al. 2019)
sws1	*P.mexicana, S. tileihornes*	(Tobler et al. 2010), (Meng, Zhao, et al. 2013)
sws2	*A.mexicanus*	(Simon et al. 2019)
rh1.1	*A.mexicanus, S.tileihornes, S. anophtalmus*	(Simon et al. 2019), (Meng, Zhao, et al. 2013), (Meng, Braasch, et al. 2013)
rh2	*P.mexicana*	(Tobler et al. 2010)
rgr1	*A.mexicanus*	(Simon et al. 2019)
rgr2	*A.mexicanus*	(Simon et al. 2019)
exorh	*A.mexicanus*	(Simon et al. 2019)
gnat1	*A.mexicanus, S. anophtalmus*	(Gross et al. 2013), (Meng, Braasch, et al. 2013)
gnat2	*A.mexicanus, S. anophtalmus*	(Gross et al. 2013), (Meng, Braasch, et al. 2013)
arr3a	*A.mexicanus*	(Gross et al. 2013)
cryba2a	*A.mexicanus*	(Gross et al. 2013)
crybgx	*A.mexicanus*	(Hinaux et al. 2015)
cryaa	*A.mexicanus*	(Hinaux et al. 2015)
crygm5	*A.mexicanus*	(Hinaux et al. 2015)

Tableau 5 : Gènes sous-exprimés chez une ou plusieurs espèces de poissons cavernicoles.

Afin de vérifier la seconde hypothèse, à savoir que la totalité ou la quasi-totalité des gènes du groupe de la photoréception évoluent de façon neutre mais n'ont pas eu le temps d'accumuler des mutations perte de fonction, nous avons étudié plus en détail le mode d'accumulation de celles-ci chez les poissons cavernicoles diploïdes.

F- Estimation du nombre de gènes neutres à partir du nombre de mutations perte de fonction et du nombre de pseudogènes

1) Répartition des mutations perte de fonction

La première méthode utilisée afin d'estimer le nombre de gènes de photoréception évoluant de façon neutre est basée sur le nombre et la répartition des mutations perte de fonction observés. Soit N un nombre de gènes évoluant de façon neutre au sein d'un groupe de gènes G, N gènes pourront accumuler des mutations perte de fonction tandis que G-N gènes n'auront jamais ce type de mutation (G-N gènes sous sélection purificatrice). En faisant l'hypothèse que chaque gène a la même longueur et que toutes les séquences neutres ont un même taux de mutation, une mutation perte de fonction aura une chance 1/N d'apparaitre sur chacun d'eux. Ainsi, la probabilité qu'un gène contienne X mutations perte de fonction est calculée comme suit, avec m = nombre total de mutations perte de fonction.

$$P(X = i) = \frac{N}{G}\left(1 - \frac{1}{N}\right)^{m} + \frac{G - N}{G} \qquad si\ i = 0$$

$$P(X = i) = \frac{N}{G}\,\frac{m!}{i!\,(m-i)!}\left(\frac{1}{N}\right)^{i}\left(1 - \frac{1}{N}\right)^{m-i} \qquad si\ i \neq 0$$

Cependant, en réalité, les gènes ont des tailles et des nombres d'introns différents. Les plus petits gènes sont *gngt2a* et *gngt2b* qui ne font que 210 paires de bases, tandis que le plus grand gène est *gc3* avec 3531 paires de bases, soit presque 17 fois plus. De même, certains gènes

n'ont qu'un seul exon comme *opn4m2* ou *rh1*, alors que *pde6b* et *pde6c* en ont 22. Ceux-ci ont donc des probabilités d'être touchés par des mutations très différentes. Pour prendre en compte cette hétérogénéité, des simulations ont été réalisées. Celles-ci consistent simplement à distribuer de façon aléatoire *m* mutations perte de fonction sur N gènes pris au hasard parmi G gènes, avec une probabilité de mutation qui tient compte de la taille du gène et du nombre d'introns. En réalisant un grand nombre de simulations, il est alors possible d'avoir une distribution théorique de la répartition des mutations sur les gènes que l'on pourra comparer à la répartition observée. Le but sera alors de faire varier N et d'estimer sa valeur la plus probable en comparaison avec les observations. Chez une espèce cavernicole, N se situe entre le nombre de pseudogènes (qui est le nombre minimal de séquences qui peuvent accumuler des mutations perte de fonction) et G.

Étant donné que les espèces du genre *Sinocyclocheilus* sont tétraploïdes et ont donc accumulé des mutations depuis la duplication du génome et non uniquement depuis la vie cavernicole, elles ne seront pas utilisées. Par ailleurs, le très petit nombre de mutations chez le morphotype cavernicole d'*Astyanax mexicanus* et chez l'espèce *Lamprologus lethops* ne permet pas non plus l'application de cette méthode chez ces organismes.

Chez les poissons cavernicoles *L. dentata* et *T. subterraneus*, 22 et 43 mutations perte de fonction sont réparties au sein de 19 et 27 pseudogènes respectivement (Tableau 6). Chez le poisson semi-cavernicole *L. gibarensis*, 7 mutations sont réparties au sein de 5 pseudogènes (Tableau 6).

		Nombre de mutations perte de fonction						
		0	1	2	3	4	5	Total
Espèce	*L. dentata*	57	16	3	0	0	0	76
	L. gibarensis	71	3	2	0	0	0	76
	T. subterraneus	35	17	7	1	1	1	62

Tableau 6 : Répartition des mutations perte de fonction chez les poissons cavernicoles L. dentata et T. subterraneus et chez le poisson semi-cavernicole L. gibarensis.
Le tableau indique le nombre de gènes sans ou avec 1 à 5 mutations.

Pour chaque espèce, 10 000 simulations ont été réalisées pour chaque valeur de N, dans une gamme allant du nombre de pseudogènes observés jusqu'à G (nombre total de gènes).

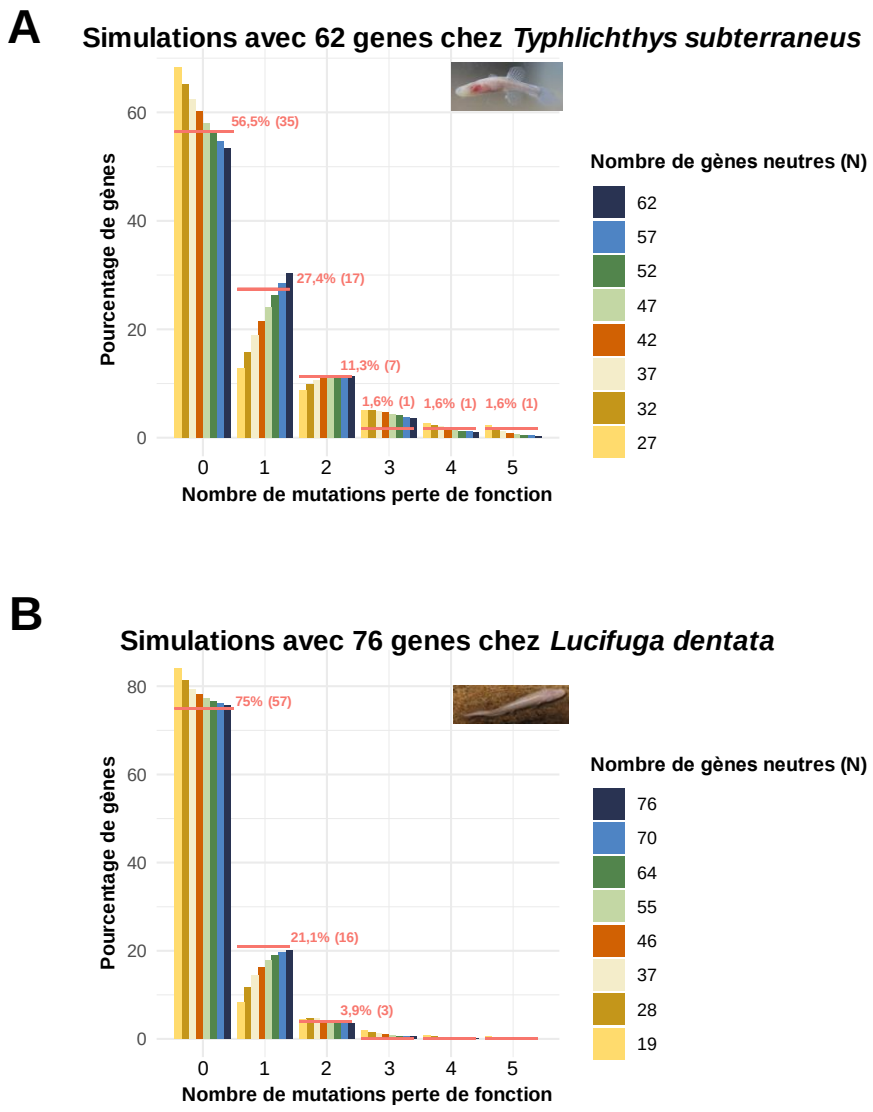


Figure 62 : Distributions observées et simulées du nombre de mutations perte de fonction par gènes chez les poissons cavernicoles.
A- T. subterraneus et B- L. dentata. Les lignes rouges correspondent aux valeurs observées. Les histogrammes correspondent aux simulations, avec N variant dans une gamme de 27 à 62 pour T. subterraneus ou de 19 à 76 pour L. dentata.

Chez les deux poissons cavernicoles, la même tendance est observée : les simulations qui se rapprochent le plus de la distribution observée sont celles où N est proche du nombre total de gènes (Figure 62), c'est-à-dire proche de 62 pour *T. subterraneus* et proche de 76 pour *L. dentata*.

Au contraire, chez *L. gibarensis*, on observe une tendance inverse, où la distribution observée est plus proche des simulations lorsque N reste relativement petit, entre 5 et 16 (Figure 63). Dans ce cas, le nombre de mutations étant petit, cette conclusion doit être considérée avec prudence. Il apparait tout de même que la plupart, sinon tous, les gènes de la photoréception seraient dispensables chez un poisson aveugle, mais seulement une petite fraction chez un poisson avec de petits yeux fonctionnels.

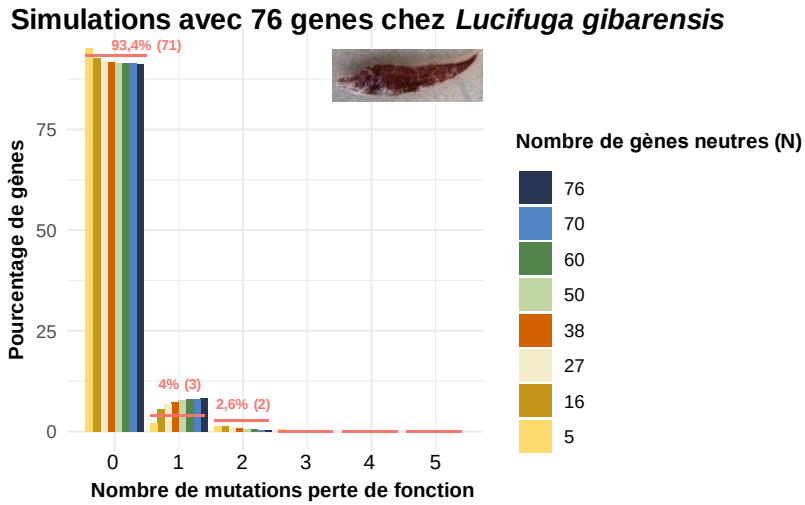


Figure 63 : Distribution observée et simulée du nombre de mutations perte de fonction par gènes chez le poisson semi-cavernicole L. gibarensis.
Les lignes rouges correspondent aux valeurs observées. Les histogrammes correspondent aux simulations, avec N variant dans une gamme allant de 5 à 76.

2) Comparaison des pseudogènes de *L. dentata* et *T. subterraneus*.

Si tous les gènes de la photoréception évoluent de façon neutre chez ces espèces aveugles, comme suggéré par les résultats ci-dessus, et qu'on ne prend pas en compte les différences en termes de tailles de ces gènes, il est possible de calculer le nombre attendu de pseudogènes en commun entre *L. dentata* et *T. subterraneus* par le simple fait du hasard. Soit E le nombre de pseudogènes en commun, P_{Ts} la proportion de pseudogènes chez *T. subterraneus*, P_{Ld} la proportion de pseudogènes chez *L. dentata* et I le nombre de gènes en commun retrouvé chez *L. dentata* et *T. subterraneus (= 58,* Figure 64).

$$E = P_{Ts} \times P_{Ld} \times I$$

$$E = \frac{27}{62} \times \frac{19}{76} \times 58 = 6{,}3$$

D'après ce calcul élémentaire, on observe plus de pseudogènes en commun entre les deux espèces (10) que ce qui est attendu (6) si tous les gènes évoluent de façon neutre (Figure 64). Afin d'avoir un modèle plus réaliste, 10 000 simulations ont été réalisées dans lesquelles des mutations sont produite au hasard dans les gènes des deux espèces cavernicoles. Dans celles-ci, les tailles des gènes et leur nombre d'introns peuvent être pris en compte. Cela est suivi d'un comptage du nombre de pseudogènes partagés (E). Il est ensuite possible de calculer la probabilité d'obtenir chaque valeur de E entre 0 et 14, 14 étant le nombre de pseudogènes chez *L. dentata* parmi les 58 gènes retrouvés en commun chez les deux espèces (Figure 64).

Dans le cas où ni la taille des gènes, ni leur nombre d'intron est pris en compte dans les simulations, la plus grande probabilité est observée pour 6 pseudogènes en commun, en cohérence avec le calcul ci-dessus (Figure 64). En prenant en compte uniquement les tailles de gènes ou les tailles de gènes et le nombre d'introns, ce maximum est de 7. Avec ce dernier

modèle, plus réaliste, la probabilité d'observer 10 pseudogènes en commun est assez faible (0,06) mais se situe au-dessus du seuil de 5% (Figure 64).

Les résultats des différentes simulations, que ce soit pour la répartition des mutations perte de fonction ou pour le nombre de pseudogènes en commun, convergent vers une même tendance : il semble probable que la majorité des gènes de la photoréception, si ce n'est pas tous, peuvent être inactivés chez les poissons cavernicoles aveugles.

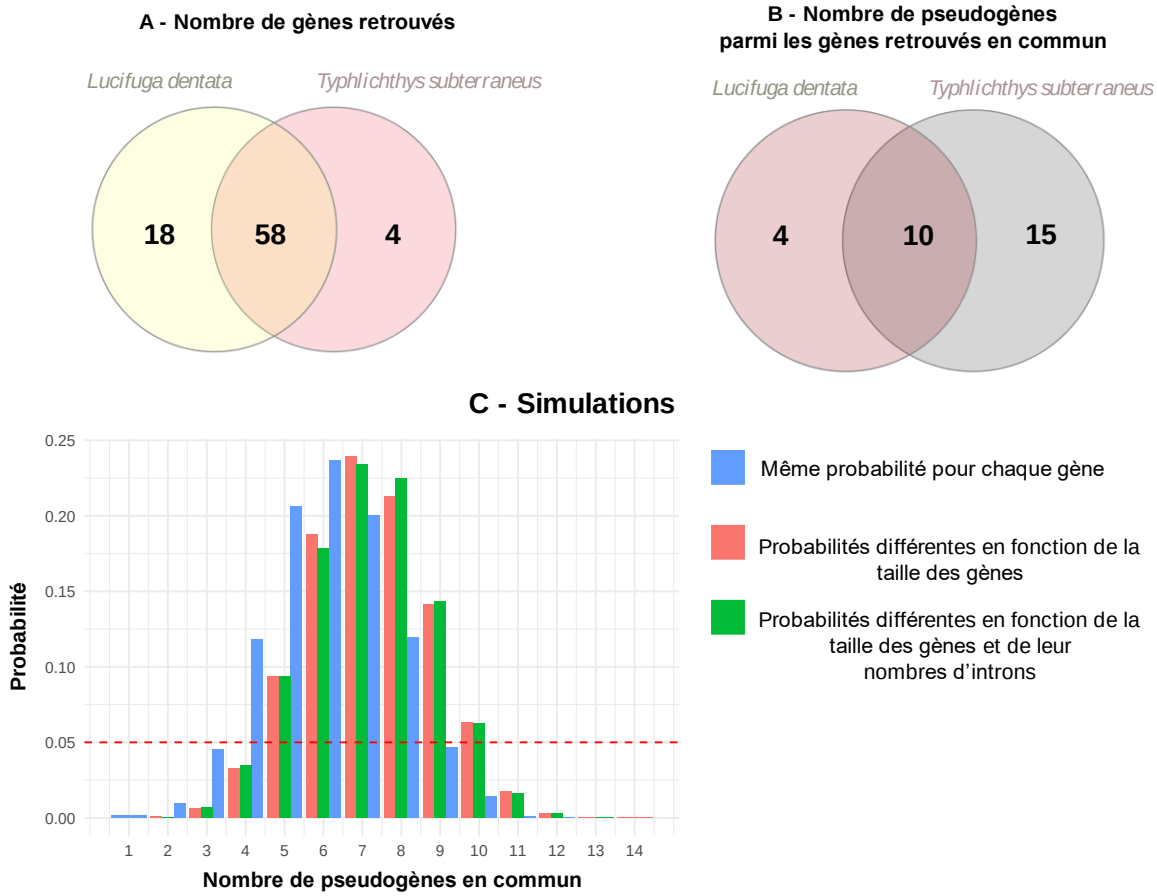


Figure 64 : Nombre de gènes et pseudogènes en commun entre L. dentata et T. subterraneus. A- Nombre de gènes retrouvés uniquement chez L. dentata et T. subterraneus ou retrouvés en commun. B- Nombre de pseudogènes partagés ou non chez L. dentata et T. subterraneus parmi les 58 gènes en commun C- Résultats de 10 000 simulations de pseudogénisation au hasard, avec une même probabilité de mutation sur chaque gène ou avec une probabilité différente en fonction de la taille en paires de bases et du nombre d'introns.

G- Estimation de la neutralité à l'aide du dN/dS

Les mutations perte de fonction ne représentant qu'une petite partie de toutes les mutations, nous nous sommes aussi intéressés aux mutations synonymes et non-synonymes, en essayant tout d'abord d'estimer des valeurs de dN/dS à l'aide du programme PAML (Yang 2007).

Le dN/dS

Également appelée Ka/Ks, le dN/dS permet d'estimer la pression de sélection agissant sur un gène. Si la différence de probabilité entre transition et transversion n'est pas prise en compte, le code génétique implique que 25 % des mutations sont synonymes. Celles-ci ne provoquent pas de changement d'acide aminé, et sont donc des mutations neutres. Les autres mutations sont des mutations non-synonymes, sur lesquelles la sélection va agir afin de les faire disparaître si elles sont désavantageuses (sélection négative), ou de les fixer dans la population si elles sont avantageuses (sélection positive).

Le dN/dS est le rapport entre le nombre de substitutions non-synonymes par sites non-synonymes et le nombre de substitutions synonymes par sites synonymes. Si le gène évolue sous sélection purificatrice, il y aura peu de substitutions non-synonymes et le dN/dS sera inférieur à 1. Au contraire, si le gène évolue sous sélection positive, il y aura un excès de substitutions non-synonymes et le dN/dS sera supérieur à 1. Dans le cas où un gène évolue de façon neutre, les mutations non-synonymes auront la même chance de se fixer ou de disparaître que les mutations synonymes, donnant un dN/dS égal à 1.

Chez un organisme donné, si les mutations synonymes sont réellement neutres, le dS devrait être proportionnel au taux de mutation, et sa valeur devrait être la même pour tous les gènes. Plusieurs limitations à l'utilisation du dN/dS existent : par exemple, l'apparition de mutations adjacentes sur le même codon va potentiellement impliquer une pression de sélection sur une mutation synonyme par liaison avec une mutation non-synonyme, engendrant une corrélation entre le dN/dS et le dS (Stoletzki et Eyre-Walker 2011).

Le taux de divergence nucléotidique des séquences codantes entre les espèces cavernicoles et les espèces de surface les plus proches étant très faible (0,9 % entre les deux espèces de *Lucifuga* ; 0,2 % entre les deux morphotypes d'*Astyanax mexicanus* et les deux espèces de *Lamprologus*), les alignements de gènes ont été concaténés pour les trois groupes. Nous avons donc généré un alignement des gènes de la photoréception, un alignement des gènes du cycle circadien et un alignement des gènes de la pigmentation. Les espèces du genre *Sinocyclocheilus* ont été traitées indépendamment, en concaténant les alignements contenant un seul gène par paire d'ohnologues (donc chaque espèce correspond à deux séquences dans l'alignement final) pour les trois groupes de gènes.

Trois modèles ont été testés : un modèle où toutes les branches évoluent sous la même pression de sélection, avec un dN/dS équivalent (= modèle un ratio) ; un modèle où les branches terminales des espèces cavernicoles ont un dN/dS différent des branches menant aux espèces de surface (= modèle deux ratios) ; un modèle où toutes les branches ont un dN/dS différent (= modèle ratios libres).

	Modèle	lnL	ω (dN/dS)	Modèles testés	2ΔlnL	d.f.	LRT p-value
Photoréception	Un ratio	-127568	0.08				
	Deux-ratios	-127483	ω_{SF} = 0.08 ω_{CF} = 0.17	Deux-ratios vs un ratio	170	1	<0.001
	Ratios libres	**-127369**	**Voir figure ratios-libres**	**Ratios libres vs Deux-ratios**	228	**26**	**<0.001**
Cycle circadien	Un ratio	-59932	0.05				
	Deux-ratios	-59909	ω_{SF} = 0.05 ω_{CF} = 0.10	Deux-ratios vs un ratio	46	1	<0.001
	Ratios libres	**-59841**	**Voir figure ratios-libres**	**Ratios libres vs Deux-ratios**	136	**26**	**<0.001**
Pigmentation	Un ratio	-586350	0.06				
	Deux-ratios	-586255	ω_{SF} = 0.06 ω_{CF} = 0.13	Deux-ratios vs un ratio	190	1	<0.001
	Ratios libres	**-585899**	**Voir figure ratios-libres**	**Ratios libres vs Deux-ratios**	712	**26**	**<0.001**

Figure 65 : Modèles testés pour les trois groupes de gènes.
Le lnL (log likelihood) le plus élevé correspond au meilleur modèle. Pour le modèle deux-ratios, le dN/dS des branches cavernicoles (L. dentata, A. mexicanus, L. lethops et T. subterraneus) est noté ω_{CF} tandis que le dN/dS des branches de surface est noté ω_{SF}.

Les résultats montrent que le modèle deux ratios est meilleur que le modèle un ratio pour les trois groupes de gènes, avec un dN/dS systématiquement plus élevé chez les espèces cavernicoles que chez les espèces de surface (Figure 65). Cependant le modèle ratios libres reste le meilleur pour les trois groupes de gènes.

A- Valeurs de dN, dS et dN/dS pour les gènes de la photoréception

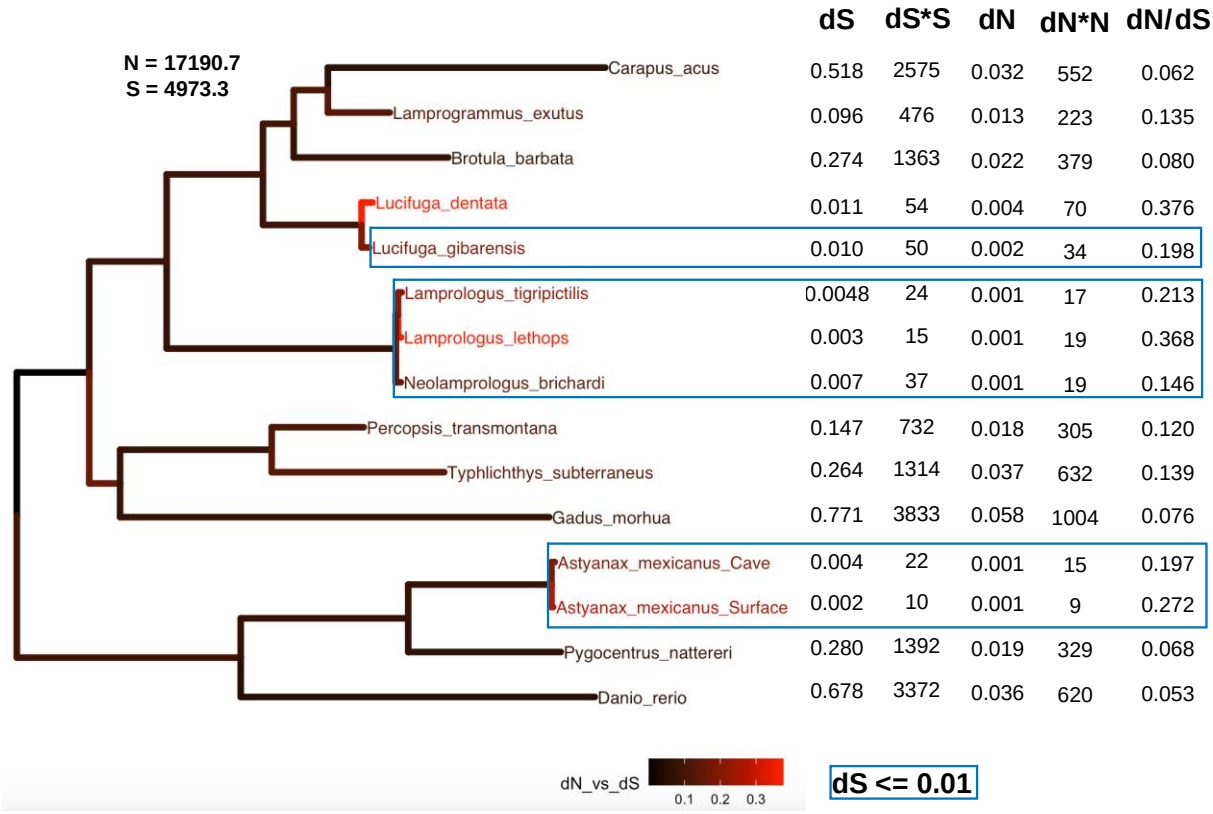


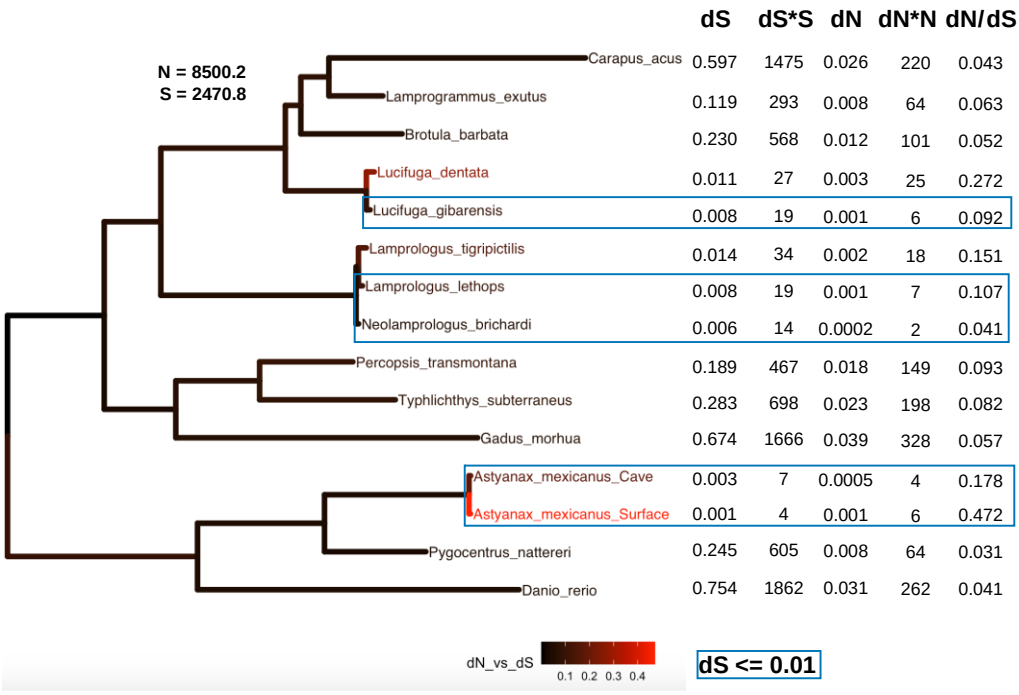
Figure 66 : Valeurs du dS, dN et dN/dS pour les espèces étudiées sur les gènes de la photoréception sous le modèle ratios libres.
*Les branches avec un dS inférieur ou égal à 0.01 sont encadrées. La valeur de dS*S correspond au nombre de substitutions synonymes et la valeur du dN*N correspond au nombre de substitutions non-synonymes. Les valeurs de dN/dS pour les trois espèces de Sinocyclocheilus, non présentes dans la figure sont 0,36, 0,32 et 0,24 pour S. anshuiensis, S. rhinocerous et S. grahami respectivement.*

La Figure 66 indique les valeurs de dN/dS sous le modèle ratios libres chez toutes les espèces, pour les gènes de la photoréception. Les espèces cavernicoles (*L. dentata*, *T. subterraneus*, *L. lethops et S. anshuiensis*) ont des valeurs de dN/dS plus élevées que les espèces de surface ou semi-cavernicoles les plus proches *(L. gibarensis*, *P. transmontana*, *L. tigripictilis et S. grahami*). Cependant, même en concaténant tous les gènes, deux limitations majeures existent quant à l'utilisation de cette méthode dans notre cas d'étude. La première résulte du fait que le temps de divergence entre espèces n'est pas équivalent au temps de colonisation des grottes. Cela implique que les branches cavernicoles représentent un mélange entre un temps d'évolution sous sélection et un temps d'évolution neutre, après colonisation. Il en résulte que

même si les gènes évoluent maintenant de façon neutre, il ne sera pas possible d'observer des dN/dS proches de 1. La deuxième limitation est le nombre de substitutions très faible, notamment entre les deux morphotypes d'*Astyanax mexicanus*. Cela rend les estimations des dN/dS très imprécises, pouvant expliquer la valeur plus élevée observée chez le morphotype de surface par rapport au morphotype cavernicole.

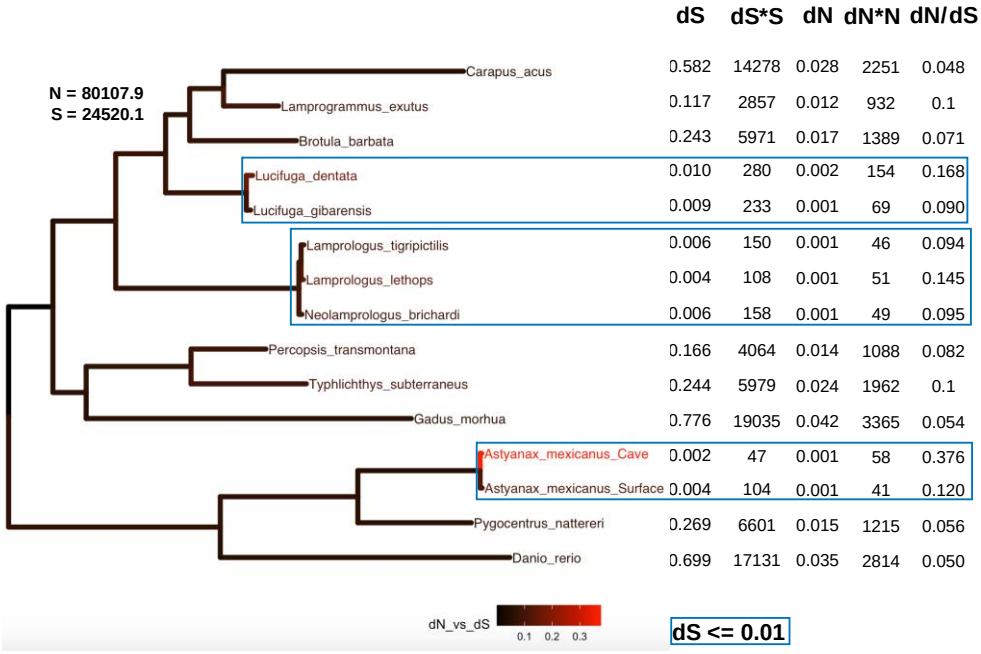
Les résultats pour les gènes du cycle circadien et de la pigmentation, bien que difficilement interprétables en tenant en compte de ces deux limitations, se trouvent dans la Figure 67. On retrouve des valeurs de dN/dS plus élevées chez les espèces cavernicoles que pour les espèces de surface proches, avec quelques exceptions, par exemple pour *T. subterraneus* et *P. transmontana* qui ont des valeurs équivalentes pour les deux groupes de gènes. Encore une fois, le dN/dS du morphotype de surface d'*Astyanax mexicanus* demeure très imprécis, notamment pour le cycle circadien, avec seulement 4 substitutions synonymes et 6 substitutions non-synonymes sur la branche menant à ce dernier (Figure 67).

B- Valeurs de dN, dS et dN/dS pour les gènes du cycle circadien



	dS	dS*S	dN	dN*N	dN/dS
Carapus_acus	0.597	1475	0.026	220	0.043
Lamprogrammus_exutus	0.119	293	0.008	64	0.063
Brotula_barbata	0.230	568	0.012	101	0.052
Lucifuga_dentata	0.011	27	0.003	25	0.272
Lucifuga_gibarensis	0.008	19	0.001	6	0.092
Lamprologus_tigripictilis	0.014	34	0.002	18	0.151
Lamprologus_lethops	0.008	19	0.001	7	0.107
Neolamprologus_brichardi	0.006	14	0.0002	2	0.041
Percopsis_transmontana	0.189	467	0.018	149	0.093
Typhlichthys_subterraneus	0.283	698	0.023	198	0.082
Gadus_morhua	0.674	1666	0.039	328	0.057
Astyanax_mexicanus_Cave	0.003	7	0.0005	4	0.178
Astyanax_mexicanus_Surface	0.001	4	0.001	6	0.472
Pygocentrus_nattereri	0.245	605	0.008	64	0.031
Danio_rerio	0.754	1862	0.031	262	0.041

C- Valeurs de dN, dS et dN/dS pour les gènes de la pigmentation



	dS	dS*S	dN	dN*N	dN/dS
Carapus_acus	0.582	14278	0.028	2251	0.048
Lamprogrammus_exutus	0.117	2857	0.012	932	0.1
Brotula_barbata	0.243	5971	0.017	1389	0.071
Lucifuga_dentata	0.010	280	0.002	154	0.168
Lucifuga_gibarensis	0.009	233	0.001	69	0.090
Lamprologus_tigripictilis	0.006	150	0.001	46	0.094
Lamprologus_lethops	0.004	108	0.001	51	0.145
Neolamprologus_brichardi	0.006	158	0.001	49	0.095
Percopsis_transmontana	0.166	4064	0.014	1088	0.082
Typhlichthys_subterraneus	0.244	5979	0.024	1962	0.1
Gadus_morhua	0.776	19035	0.042	3365	0.054
Astyanax_mexicanus_Cave	0.002	47	0.001	58	0.376
Astyanax_mexicanus_Surface	0.004	104	0.001	41	0.120
Pygocentrus_nattereri	0.269	6601	0.015	1215	0.056
Danio_rerio	0.699	17131	0.035	2814	0.050

Figure 67 : Valeurs du dS, dN et dN/dS pour les espèces étudiées sur les gènes du cycle circadien et de la pigmentation, sous le modèle ratios libres.
*Les branches avec un dS inférieur ou égal à 0,01 sont encadrées. La valeur de dS*S correspond au nombre de substitutions synonymes et la valeur du dN*N correspond au nombre de substitutions non-synonymes. Valeurs de dN/dS pour S. anshuiensis, S. rhinocerous et S. grahami respectivement : pour le cycle circadien, 0,36, 0,37 et 0,25 ; pour la pigmentation, 0,29, 0,33 et 0,25.*

H- Estimation du régime d'évolution des groupes de gènes à partir de l'effet délétère des mutations non synonymes

Face à ces faibles taux de divergences, nous avons développé une méthode prenant en compte l'impact des substitutions non-synonymes plutôt que leur nombre. Plusieurs programmes, tels que CADD ou SIFT (Zeng et Bromberg 2019) existent pour évaluer l'effet délétère des substitutions non-synonymes sur un gène, se basant sur de l'apprentissage automatique (machine learning). Ici, le programme Mutpred2, qui est le plus récent et performant à ce jour (Pejaver et al. 2020) a été utilisé pour attribuer un score entre 0 et 1 à toutes les substitutions. Plus le score est proche de 1, plus l'effet de la substitution est délétère pour le gène. Mutpred2 a été entraîné à l'aide d'un jeu de données composé de 53 180 mutations pathogènes et 206 946 mutations bégnines chez l'humain.

En plus des substitutions observées, un score a été attribué à des substituions produites par des simulations d'évolution neutre des gènes des trois groupes (en utilisant les séquences du poisson-zèbre). Dans ces simulations, le taux de transition/transversion a peu d'impact sur la distribution des scores des substitutions, qui varie cependant en fonction du groupe de gènes (Figure 68).

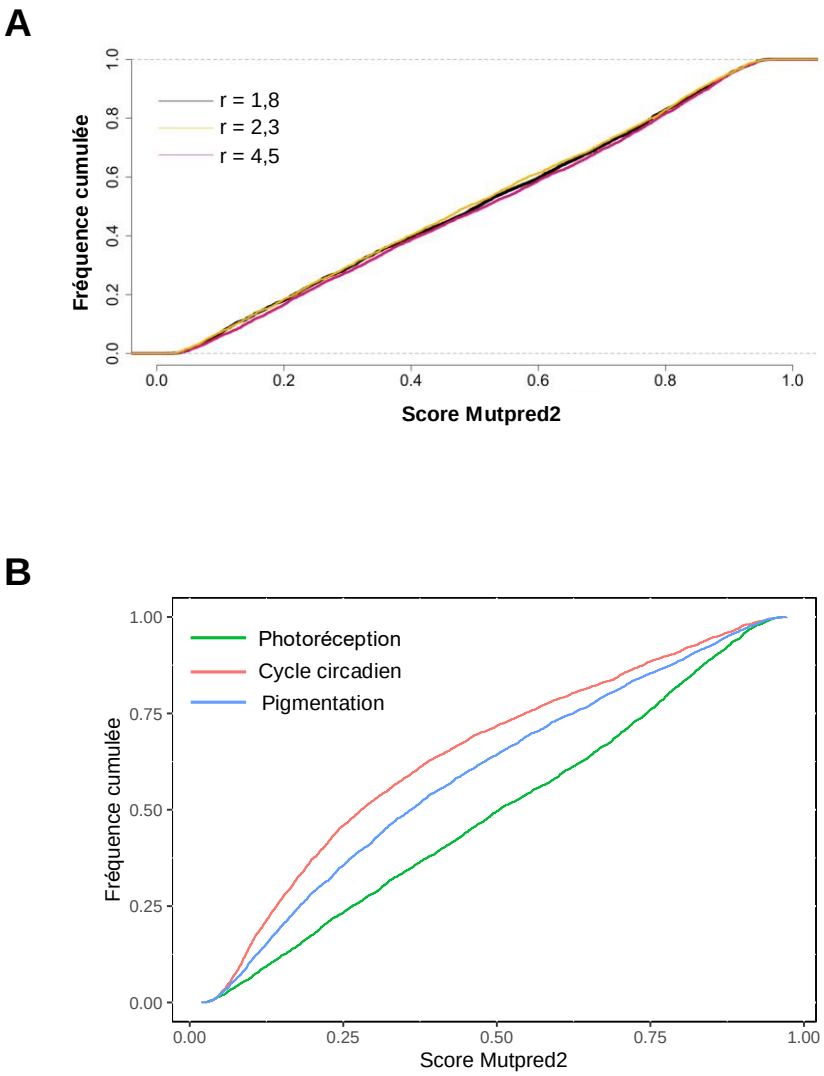


Figure 68 : Distribution cumulée des scores Mutpred2 en fonction du ratio transition/transversion et du groupe de gènes.
A- 3000 mutations avec un taux de transition/transversion (r) correspondant à celui observé chez le Danio rerio (1,8), chez les Characiformes (2,3) ou chez les Ophiidiformes (4,5) sur les gènes de la photoréception. B- 3000 mutations sur les gènes de la photoréception, du cycle circadien ou de la pigmentation avec le même taux de transtition/transversion (1,8).

Ainsi, pour chaque groupe de gènes, les distributions de scores observées chez chaque espèce ont été comparées à une distribution théorique sous évolution neutre, représentant la distribution

moyenne de 100 simulations avec un nombre de substitutions égal au plus petit nombre de substitutions observé chez un poisson cavernicole (51 substitutions pour les gènes de la photoréception, 33 pour les gènes du cycle circadien et 221 pour les gènes de la pigmentation).

1) Gènes de la photoréception

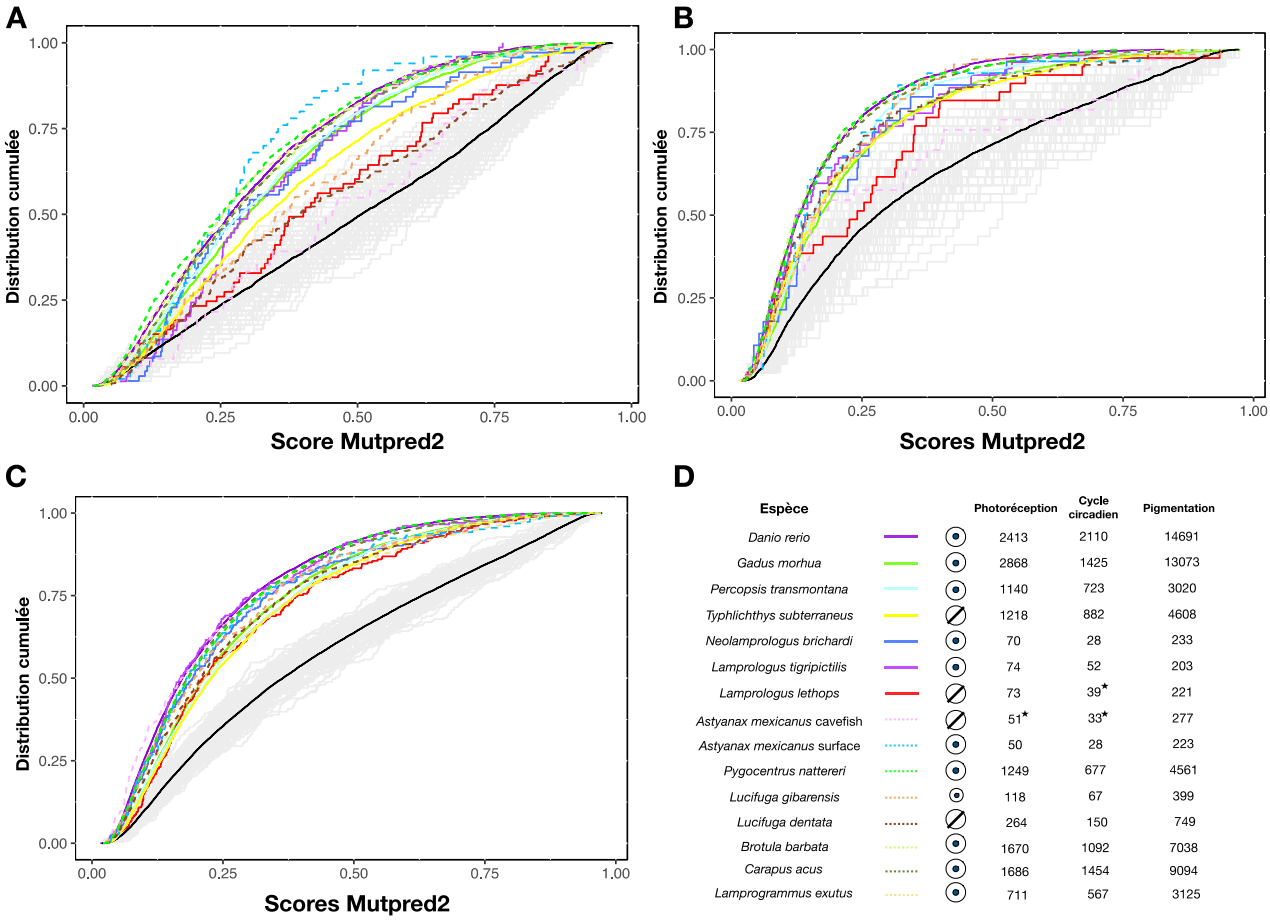


Figure 69 : Distributions cumulées des scores Mutpred2 pour chaque espèce de l'étude et chaque groupe de gènes.
Le score varie entre 0 (mutation non délétère) et 1 (mutation très délétère). Les courbes grises représentent 100 simulations d'évolution neutre et la courbe noire représente la moyenne de ces simulations. A- Gènes de la photoréception B- Gènes du cycle circadien C- Gènes de la pigmentation. D- Pour chaque espèce, la couleur de la courbe sur les distributions cumulées ainsi que la présence/absence d'œil est représentée. De plus, pour chaque groupe de gènes, le nombre de mutations non-synonymes scorées est indiqué et les étoiles représentent les distributions non significativement différentes de la distribution sous évolution neutre.

Les distributions des scores pour les gènes de la photoréception et pour les espèces de surface suivent toutes une distribution attendue sous sélection purificatrice, avec la grande majorité des substitutions ayant un effet délétère faible et un petit nombre de substitutions estimées avec un fort effet délétère (Figure 69). Toutes ces distributions sont significativement différentes (tests de Kolmogorov-Smirnov) de celle attendue sous évolution neutre (simulations). Au contraire, et de façon cohérente avec le fait que les branches terminales menant aux espèces cavernicoles sont un mélange d'évolution sous sélection et d'évolution neutre, les distributions de ces espèces se situent entre les distributions des poissons de surface et la distribution neutre.

Des mélanges entre substitutions issues du poisson-zèbre (substitutions apparues sous sélection) et de substitutions issues des simulations (substitutions apparues sans sélection) ont été réalisées afin d'estimer le mélange le plus probable observé dans ces branches. Le mélange le plus probable étant celui avec la p-value la plus grande quand comparée à la distribution observée avec un test de Kolmogorov-Smirnov. Ainsi pour les espèces de surface, les mélanges les plus probables sont ceux où la grande majorité des substitutions est issue du poisson-zèbre (Figure 70).

La distribution observée pour le morphotype cavernicole d'*Astyanax mexicanu*s est la seule qui n'est pas significativement différente de la distribution neutre (p-value = 0,2) (Figure 69). Le mélange le plus probable pour cette espèce se compose de 82% de substitutions neutres et 18% de substitutions sous sélection. Cela indique que plus de 4 substitutions sur 5 observées dans cette branche sont probablement apparues après colonisation des grottes (Figure 70).

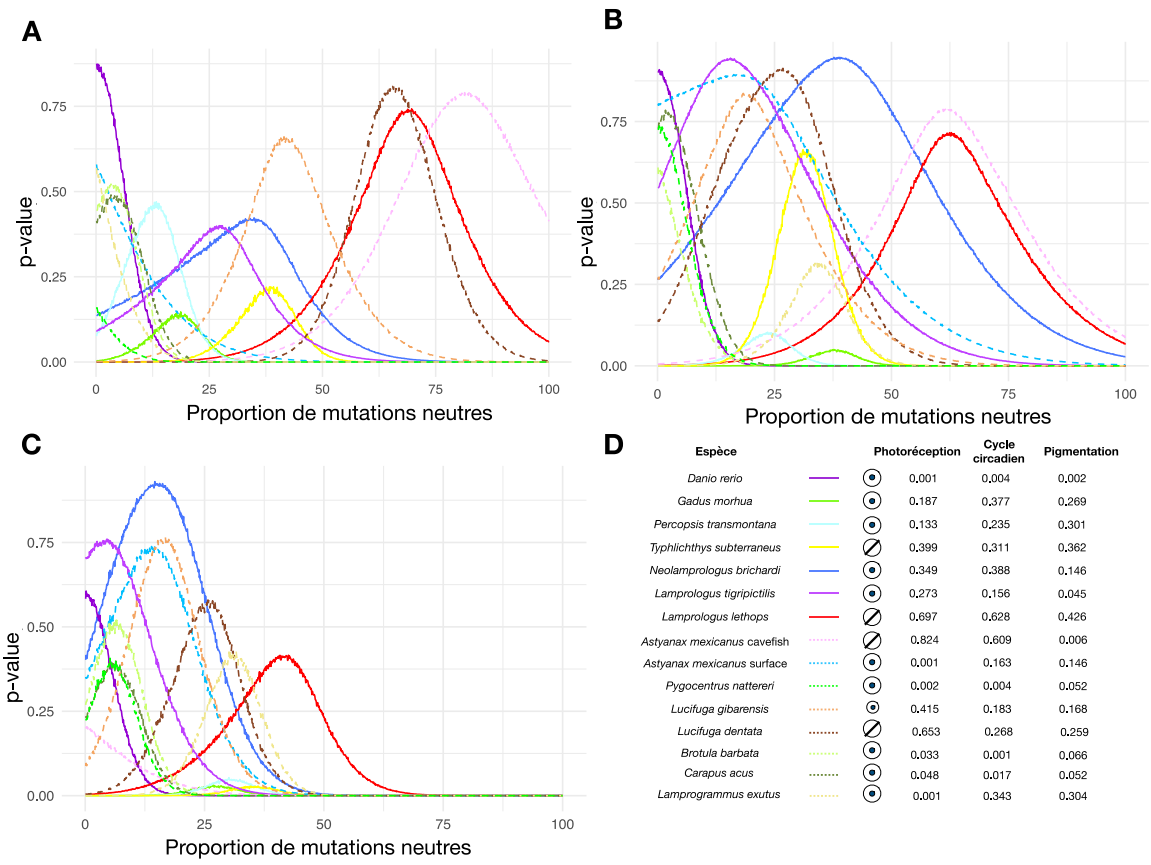


Figure 70 : p-values du test de Kolmogorov-Smirnov entre les distributions observées pour chaque espèce et des mélanges de la distribution sous évolution neutre et de la distribution sous sélection en différentes proportions.
A- Gènes de la photoréception B- Gènes du cycle circadien C- Gènes de la pigmentation D- Pour chaque espèce, la couleur de la courbe sur les graphiques A, B et C ainsi que la présence/absence d'œil est représentée. De plus, pour chaque groupe de gènes, la proportion de mutations neutres qui donne le mélange le plus similaire à la distribution observée est indiquée (p-value la plus élevée).

La distribution pour le poisson cavernicole *Lucifuga dentata* est plus proche de celle attendue sous évolution neutre que la distribution pour le poisson semi-cavernicole *Lucifuga gibarensis*. Les mélanges les plus probables pour ces espèces comprennent 65% et 42% de substitutions issues de la distribution neutre respectivement. Pour les espèces cavernicoles *Lamprologus lethops* et *Typhlichthys subterraneus*, les mélanges les plus probables se composent de 70% et 40% de substitutions apparues sans sélection respectivement (Figure 69, Figure 70).

Pour les espèces du genre *Sinocyclocheilus*, les distributions sont toutes différentes de celles attendues sous sélection (Figure 71). Cela peut s'expliquer par une baisse de la pression de sélection sur la moitié des gènes (sur une copie de chaque paire d'ohnologues) après la duplication entière du génome. De plus, les distributions pour l'espèce cavernicole *S. anshuiensis* et pour l'espèce semi-cavernicole *S. rhinocerous* se rapprochent davantage de la distribution neutre que pour l'espèce de surface S. grahami, ce qui est probablement dû à une perte secondaire de sélection sur ces gènes après colonisation des grottes. Les mélanges les plus probables pour ces espèces se composent de 29%, 60% et 57% de substitutions sous évolution neutre pour *S. grahami*, *S. rhinocerous* et *S. anshuiensis* respectivement (Figure 71).

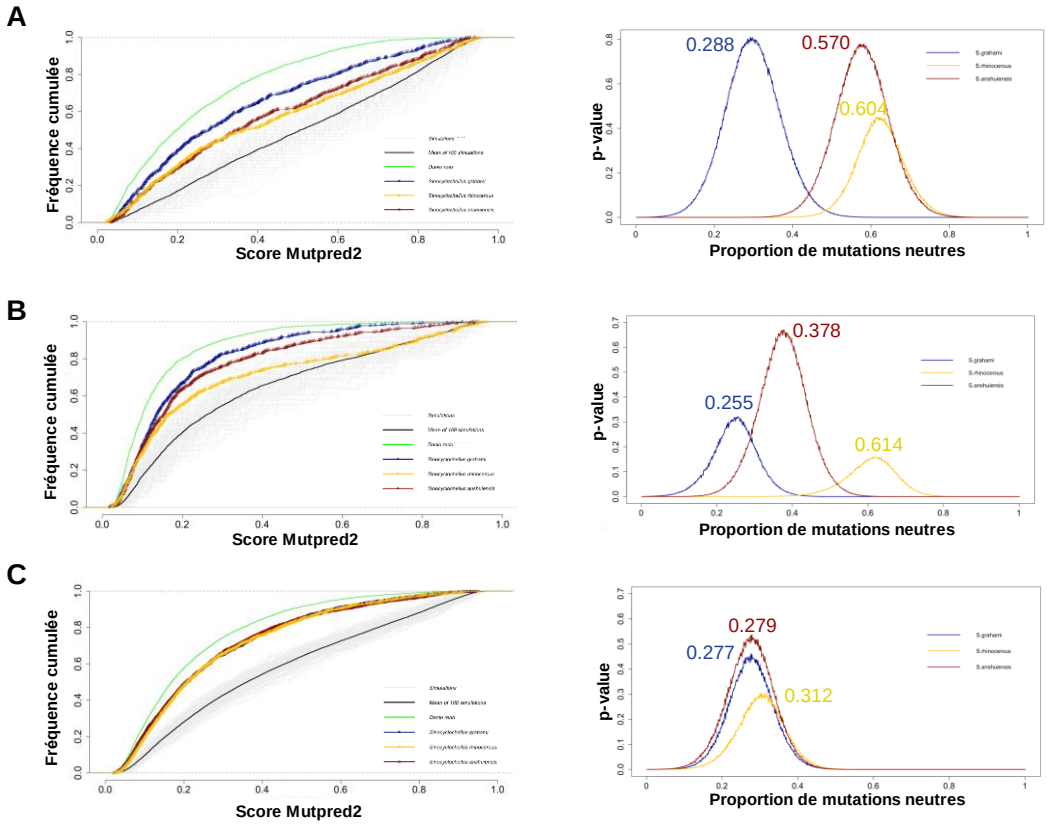


Figure 71 : Distribution cumulée des scores Mutpred2 pour les espèces du genre Sinocyclocheilus et p-value du test de Kolmogorov-Smirnov entre les distributions observées et des mélanges de la distribution neutre et de la distribution sous sélection. La proportion de mutations sous évolution neutre qui donne le mélange le plus similaire aux distributions observées est indiquée au-dessus de chaque courbe. A- Gènes de la photoréception B- Gènes du cycle circadien C- Gènes de la pigmentation.Courbes bleues : S. grahami. Courbes rouges : S. anshuiensis. Courbes jaunes : S. rhinocerous.

2) Gènes du cycle circadien et de la pigmentation

Pour les gènes du cycle circadien, les distributions pour la majorité des espèces cavernicoles sont semblables à celles des poissons de surface et suivent la distribution attendue sous sélection. Trois espèces sortent cependant du lot et semblent s'écarter un peu plus de cette distribution que les autres organismes. Le morphotype cavernicole d'*A. mexicanus* et *L. lethops*, pour qui les distributions observées ne sont pas significativement différentes de la distribution neutre et chez qui le mélange le plus probable est constitué de 60 % et 63 % de substitutions sous évolution neutre respectivement (Figure 69, Figure 70). De plus, et de façon cohérente avec le grand nombre de pseudogènes observés chez cette espèce, la distribution de *S. rhinocerous* est plus proche de la distribution neutre que pour les autres espèces tétraploïdes, avec plus de 61% des substitutions apparues sans sélection (Figure 71).
Le très faible nombre de pseudogènes observés chez les espèces cavernicoles pour les gènes de la pigmentation, indiquant une pression de sélection toujours forte sur ces gènes même dans les grottes, semble se confirmer avec les mutations non-synonymes. Effectivement, toutes les distributions, que ce soit pour les espèces cavernicoles ou pour les espèces de surface, suivent la distribution attendue sous sélection (Figure 69). Il en va de même pour les espèces du genre *Sinocyclocheilus*, qui ont toutes des distributions similaires, représentatives d'une baisse de sélection initiale après duplication du génome mais sans perte de sélection suite à la vie cavernicole (Figure 71).

Ces résultats obtenus avec Mutpred2 vont dans le même sens que les observations de mutations perte de fonction : il semble y avoir une évolution neutre des gènes de la photoréception chez les espèces cavernicoles, qui s'est déclenchée plus ou moins longtemps après la divergence de l'espèce de surface séquencée la plus proche. Pour la majorité des espèces, les gènes du cycle circadien et de la pigmentation semblent toujours sous sélection purificatrice. La méthode que nous avons développée permet une approche complémentaire au calcul du dN/dS pour évaluer la pression de sélection, notamment utile pour les espèces ayant accumulé un très faible nombre de mutations telles que *Astyanax mexicanus* ou *Lamprologus lethops*.

I- Datations de la colonisation des grottes par différentes méthodes

Plusieurs méthodes de datations ont été utilisées afin d'estimer le temps depuis lequel les espèces cavernicoles ont colonisé les grottes.

1) Datation à partir du nombre de pseudogènes

Lors d'une perte de pression de sélection sur un gène, la probabilité qu'une mutation perte de fonction se fixe sur celui-ci dans une population après un temps t est :

$$p(t) = 1 - e^{-\mu_P t} \qquad si\ Ne \ll 1/\mu_P$$

Avec μ_P = taux de mutation perte de fonction par gène par génération et Ne = taille effective de la population (Li et Nei 1977).

Si l'on considère maintenant un groupe de T gènes évoluant de façon neutre, avec la même probabilité de mutation pour chaque gène (même longueur et même nombre d'introns), la probabilité que D gènes aient fixé une mutation perte de fonction après un temps t est :

$$p(X = D) = \frac{T!}{D!\,(T-D)!}\,(1 - e^{-\mu_P t})^D \times (e^{-\mu_P t})^{T-D}$$

La dérivée de cette fonction selon t permet de trouver la valeur t pour laquelle la probabilité de $p(X = D)$ est la plus forte :

$$t = \frac{1}{\mu_P} \times \ln\left(\frac{T}{T-D}\right)$$

L'estimation du taux de mutation perte de fonction μ_P a été réalisée indépendamment pour les gènes de la photoréception chez *L. dentata* (0,071μ) et chez *T. subterraneus* (0,128μ), espèces pour lesquelles un grand nombre de mutations ont été retrouvées, en suivant les calculs présentés précédemment. Cependant, comme seulement 1 mutation a été retrouvée chez *A. mexicanus* et 4 chez *L. lethops*, le taux de mutation perte de fonction chez ces espèces sera estimé comme étant égal au taux calculé chez *T. subterraneus*.

Le taux de mutation μ est encore mal connu chez les vertébrés. Chez les humains, celui-ci est estimé à 10^{-8} par site par génération (Roach et al. 2010). Pour les téléostéens, deux estimations indépendantes ont été faites chez dans la famille des Cichlidés : $3{,}5 \times 10^{-9}$ (Malinsky et al. 2018) ou $6{,}6 \times 10^{-8}$ (Recknagel, Elmer, et Meyer 2013). Ainsi, dans les analyses qui suivent, la valeur du taux de mutation sera de 10^{-8} par site par génération. Avec une longueur moyenne des gènes de la vision égale à 1091, 790, 1045 et 1081 chez *L. dentata*, *T. subterraneus*, *L. lethops* et *A. mexicanus* respectivement, le taux de mutation par gène par génération est égal à $0{,}78 \times 10^{-6}$, 1×10^{-6} , $1{,}3 \times 10^{-6}$ et $1{,}4 \times 10^{-6}$.

Pour chaque espèce cavernicole[34], il est alors possible de trouver t :

$$t(L.dentata) = \frac{1}{0{,}78 \times 10^{-6}} \times \ln\left(\frac{76}{76-19}\right) \approx 368\,000\ générations$$

$$t(T.subterraneus) = \frac{1}{1 \times 10^{-6}} \times \ln\left(\frac{62}{62-27}\right) \approx 572\,000\ générations$$

$$t(L.lethops) = \frac{1}{1{,}3 \times 10^{-6}} \times \ln\left(\frac{79}{79-3}\right) \approx 30\,000\ générations$$

[34] Le raccourci est fait pour *L. lethops*. La datation présentée correspond à la colonisation des profondeurs de la rivière du Congo et non d'une grotte.

$$t(A.mexicanus) = \frac{1}{1{,}4 \times 10^{-6}} \times \ln\left(\frac{85}{85-1}\right) \approx 8\,500\ générations$$

L'histoire de la relaxation de sélection sur les gènes est plus complexe chez *S. anshuiensis.* Elle ne dépend pas seulement de la colonisation des grottes, mais aussi de la duplication du génome. Cette méthode de datation, et celles qui seront présentées par la suite, n'ont donc pas été appliquées à cette espèce.

Afin de prendre en compte la différence de taille des différents gènes, des simulations ont également été réalisées. Celles-ci permettent de faire varier la taille effective de la population Ne bien que cette valeur ait très peu d'impact sur les résultats. Ces simulations consistent à faire évoluer de façon neutre le groupe de gènes de la photoréception et d'estimer après combien de temps D gènes ont accumulé au moins une mutation perte de fonction. En réalisant 10 000 simulations pour chaque espèce, il est possible d'obtenir le temps t le plus probable mais également d'obtenir la distribution de la probabilité d'avoir D pseudogènes en fonction du temps. On pourra ainsi définir un intervalle de confiance.

N'étant pas certain de la neutralité de tous les gènes de la photoréception (qui ne sont pas tous trouvés sous forme de pseudogènes chez au moins une espèce cavernicole), une deuxième série de simulations a été réalisée pour chaque organisme cavernicole. Dans celles-ci, T et D seront calculés par rapport aux gènes trouvés avec une mutation perte de fonction chez au moins une autre espèce troglomorphe et seront appelés T_s et D_s. Par exemple, pour *T. subterraneus*, 27 pseudogènes (D) sont retrouvés parmi 62 gènes de la photoréception (T). En ne prenant en compte que les 33 pseudogènes trouvés chez les autres espèces cavernicoles, 21 gènes de la photoréception (T_s) sont trouvés dont 13 pseudogènes (D_s). Dans ce cas, on s'assure que les gènes étudiés peuvent être pseudogénisés chez les poissons cavernicoles.
La probabilité maximale de t avec T et D chez cette espèce est 569 570 générations (probabilité >5% avec t entre 447 390 et 712 830 générations). Avec T_s et D_s, la probabilité maximale de t est égale à 910 940 générations (probabilité >5% avec t entre 557 730 et 1 401 750 générations) (Figure 72, courbes rouges).

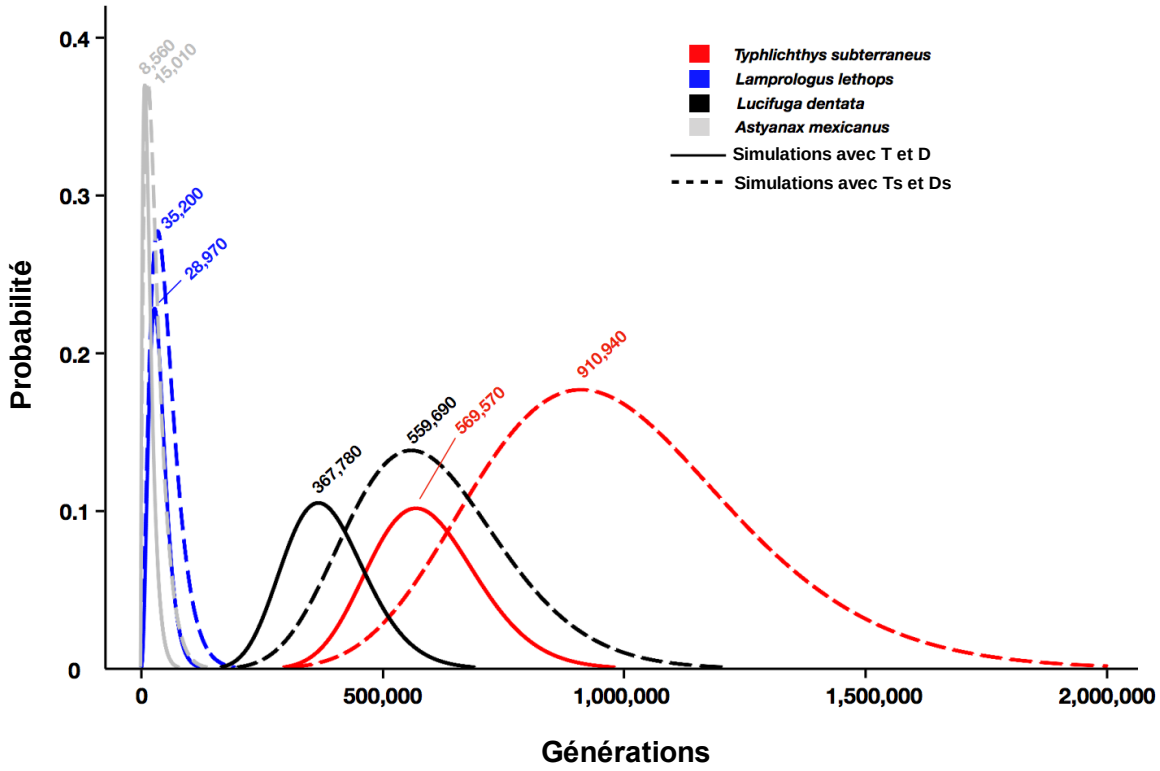


Figure 72 : Datations de la relaxation de sélection sur les gènes de la photoréception chez les espèces cavernicoles.
Les courbes pleines représentent les simulations en faisant l'hypothèse que tous les gènes de la photoréception évoluent de façon neutre. Les courbes pointillées représentent les simulations dans lesquelles seulement les gènes trouvés pseudogénisés chez au moins une autre espèce cavernicole sont pris en compte. La probabilité maximale est indiquée au-dessus de chaque courbe.

Pour *Lucifuga dentata*, avec $T = 76$ et $D = 19$, la probabilité maximale de t est égale à 367 780 générations (probabilité >5% avec t entre 273 990 et 480 980 générations). Avec $T_s = 25$ et $D_s = 13$, la probabilité maximale de t est égale à 559 690 générations (probabilité >5% avec t entre 365 220 et 814 710 générations) (Figure 72, courbes noires).

Pour *Lamprologus lethops*, avec $T = 79$ et $D = 3$, la probabilité maximale de t est égale à 28 970 générations (probabilité >5% avec t entre 8 670 et 68 620 générations). Avec $T_s = 41$ et $D_s = 2$, la probabilité maximale de t est égale à 35 200 générations (probabilité >5% avec t entre 6 640 et 103 250 générations) (Figure 72, courbes bleues).

Pour *Astyanax mexicanus*, avec $T = 85$ et $D = 1$, la probabilité maximale de t est égale à 8 560 générations (probabilité >5% avec t entre 450 et 38 580 générations). Avec $T_s = 41$ et $D_s = 1$, la probabilité maximale de t est égale à 15 010 générations (probabilité >5% avec t entre 790 et 67 740 générations) (Figure 72, courbes grises).

J'ai ainsi réalisé des simulations permettant de prendre en compte des facteurs négligés dans la méthode basés sur (Li et Nei 1977), comme la taille efficace de la population ou les variations du taux de mutation des gènes, mais les résultats sont très similaires.

Dû au très faible nombre de mutations et à la grande imprécision des estimations des dN/dS sur les gènes de la photoréception, les méthodes qui seront présentées dans les parties suivantes ne seront pas appliquées pour la datation de *L. lethops* et *A. mexicanus.*

2) Datation à partir du nombre de substitutions

En 1981, Wen-Hsiung Li et ses collaborateurs datent la perte de sélection d'un pseudogène de la famille des globines chez la souris, *ψα3* (Li, Gojobori, et Nei 1981). Pour cela, ils comparent la vitesse d'évolution d'une globine fonctionnelle avec celle de ce pseudogène. Pour la séquence d'un gène fonctionnel, le code génétique implique que les troisièmes positions des codons sont souvent neutres car leur mutation ne change pas l'acide aminé codé, contrairement aux premières et deuxièmes positions. Le taux de substitution sur les troisièmes positions des codons sera donc plus élevé que sur les deux premières. Au contraire, les pseudogènes évoluant de façon neutre par définition, toutes les positions des codons ont une même probabilité de muter, et évoluent donc à la même vitesse. La méthode repose alors sur la mesure de cette

accélération de l'évolution des pseudogènes, plus particulièrement sur les deux premières positions de chaque codon.

Afin d'utiliser cette méthode pour dater le temps depuis lequel les gènes de la photoréception évoluent de façon neutre, il est tout d'abord nécessaire d'avoir deux espèces de surface proches des espèces cavernicoles. Pour *L. dentata*, les espèces *B. barbata* et *L. gibarensis* seront utilisés. Bien que *L. gibarensis* soit une espèce semi-cavernicole, le nombre de pseudogènes semble indiquer que la majorité des gènes de la photoréception est toujours sous sélection chez cet organisme.
Pour *T. subterraneus*, les espèces de surface *P. transmontana* et *G. morhua* seront utilisées. Les temps T_p , T_d et T sont définis dans la Figure 73.

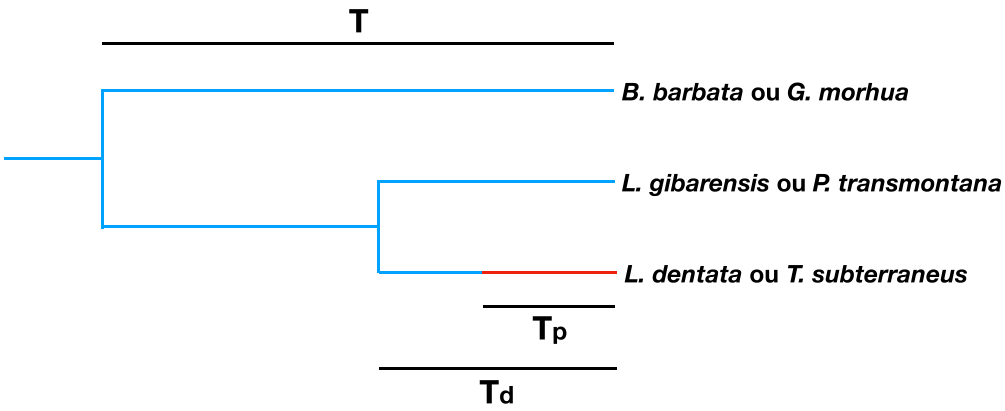


Figure 73 : Arbres phylogénétiques utilisés pour la datation de L. dentata et T. subterraneus. Tp = Temps depuis lequel les gènes de la photoréception évoluent de façon neutre, équivalent au temps de colonisation des grottes. Td = Temps de divergence entre l'espèce cavernicole et l'espèce de surface la plus proche. T = Temps de divergence entre le clade regroupant l'espèce cavernicole plus l'espèce de surface la plus proche et une deuxième espèce de surface.

Les temps de divergence T entre *L. dentata* et *B. barbata* et entre *T. subterraneus* et *G. morhua* ont été extraits du site TimeTree.org et sont égaux à 80 et 132 millions d'années respectivement.

Cette méthode permet d'obtenir les temps T_d et T_p comme suit :

$$T_d = \frac{(d_{AB1} + d_{AB2} + d_{AB3}) - (y_1 + y_2 + y_3)}{2 \times (a_1 + a_2 + a_3)}$$

$$T_p = \frac{y_{12} - y_3}{a_3 - a_{12}}$$

Avec :

$A = Espèce\ cavernicole\ (L.dentata\ ou\ T.subterraneus)$
$B = Espèce\ de\ surface\ la\ plus\ proche\ (L.gibarensis\ ou\ P.transmontana)$
$C = Espèce\ de\ surface\ plus\ eloignée\ (B.barbata\ ou\ G.morhua)$

$d_{ABi} = Nombre\ de\ substitutions$ à $la\ position\ i\ de\ chaque\ codon\ entre\ l'espèce\ A\ et\ l'espèce\ B.Cette\ valeur\ est\ estimée$ à $partir\ du\ nombre\ de\ substituions\ observées.$

$$y_i = d_{ACi} - d_{BCi} \quad \text{et} \quad y_{ij} = \frac{y_i + y_j}{2}$$

$$a_i = \frac{d_{BCi}}{2T} \quad \text{et} \quad a_{ij} = \frac{a_i + a_j}{2}$$

Avec ces équations, pour *L. dentata*, T_p est égal à 927 165 ans et T_d est égal à 4,5 millions d'années. Pour *T. subterraneus*, T_p est égal à 1,74 millions d'années et T_d est égal à 57,6 millions d'années. Cette valeur de T_d est cohérente avec le temps de divergence moyen entre *P. transmontana* et *T. subterraneus* dans la littérature, compris entre 59 et 68 millions d'années (Kumar et al. 2017, timetree.org).

3) Datation à partir des estimations de dN/dS

Une troisième méthode de datation, basée sur les valeurs de dN/dS, a été utilisée. Elle repose sur l'hypothèse que les branches menant aux espèces cavernicoles sont un mélange d'un temps d'évolution sous sélection des gènes de la photoréception (noté T_f) et d'un temps d'évolution neutre (noté T_p). Cette méthode a été introduite dans une publication pour dater le temps depuis lequel le gène *ENAM*, codant la protéine énaméline, évolue de façon neutre chez certains groupes de mammifères placentaires dépourvus de dents ou qui possèdent des dents sans émail (Meredith et al. 2009).

En faisant l'hypothèse que le taux de mutation est égal chez les espèces cavernicoles et les espèces de surface, on a :

$$T_d \times \left(\frac{dN}{dS}\right)_m = \left(T_f \times \left(\frac{dN}{dS}\right)_f\right) + (T_d - T_f) \times \left(\frac{dN}{dS}\right)_p$$

avec T_d = longueur de la branche cavernicole = temps de divergence avec l'espèce de surface la plus proche ;

$\left(\frac{dN}{dS}\right)_m$ = valeur de dN/dS sur la branche cavernicole ;

$\left(\frac{dN}{dS}\right)_f$ = valeur de dN/dS sous sélection = valeur de dN/dS chez l'espèce de surface la plus proche ;

$\left(\frac{dN}{dS}\right)_p$ = valeur de dN/dS sur le gène après relaxation de la sélection = 1.

Il est alors possible d'obtenir les valeurs de T_f et T_p :

$$T_f = \frac{\left(T_d \times \left(\left(\frac{dN}{dS}\right)_m - \left(\frac{dN}{dS}\right)_p\right)\right)}{\left(\frac{dN}{dS}\right)_f - \left(\frac{dN}{dS}\right)_p} = \frac{\left(T_d \times \left(\left(\frac{dN}{dS}\right)_m - 1\right)\right)}{\left(\frac{dN}{dS}\right)_f - 1}$$

$$T_p = T_d - T_f$$

Les valeurs de dN/dS utilisées peuvent être retrouvées dans la Figure 66. Les temps de divergences T_d entre *L. dentata* et *L. gibarensis* ou entre *T. subterraneus* et *P. transmonata* ont été obtenues avec la méthode précédente (méthode de Li and Nei, 1981).

Il est ainsi possible de trouver la valeur T_p pour les deux espèces cavernicoles :

$$T_f(L.dentata) = \frac{4{,}5 \times (0{,}376 - 1)}{0{,}198 - 1} = 3{,}5\ millions\ d'années$$

$$T_p(L.dentata) = 4{,}5 - 3{,}5 = 1\ million\ d'années$$

$$T_f(T.subterraneus) = \frac{58 \times (0{,}139 - 1)}{0{,}120 - 1} = 56{,}4\ millions\ d'années$$

$$T_p(T.subterraneus) = 57{,}6 - 56{,}4 = 1{,}2\ million\ d'années$$

4) Datation à partir des effets délétères des mutations non-synonymes

La dernière approche utilise la proportion de substitutions apparue sous sélection, noté p_f , et sous évolution neutre, noté p_p , estimées à partir des distributions de scores Mutpred2 (Figure 70). Tout comme la méthode tirée de la publication de Meredith et al. 2009, l'hypothèse qui sera faite est que les branches cavernicoles débutent par un temps sous sélection (T_f) avec un dN/dS égal à $\left(\frac{dN}{dS}\right)_f$ suivi par un temps d'évolution neutre (T_p) avec un dN/dS égal à $\left(\frac{dN}{dS}\right)_p = 1$. On a alors :

$$\frac{p_f}{p_p} = \frac{T_f \times \left(\frac{dN}{dS}\right)_f}{T_p \times \left(\frac{dN}{dS}\right)_p}$$

$$T_p = \frac{\left(\frac{dN}{dS}\right)_f}{\left(\frac{dN}{dS}\right)_p} \times \frac{p_p}{p_f} \times T_f$$

Ainsi pour *L. dentata* :

$$T_p = \frac{0{,}198}{1} \times \frac{0{,}653}{1 - 0{,}653} \times T_f = 0{,}37T_f$$

Étant donné que $T_d = temps\ de\ divergence\ entre\ L.dentata\ et\ L.gibarensis = 4{,}5\ Ma$:

$$T_d = 1{,}37T_f$$

$$T_f = \frac{4{,}5}{1{,}37} = 3{,}28\ Ma$$

$$T_p = T_d - T_f = 1{,}22\ millions\ d'années$$

Pour *T. subterraneus* :

$$T_p = \frac{0{,}120}{1} \times \frac{0{,}399}{1 - 0{,}399} \times T_f = 0{,}08T_f$$

$T_d = temps\ de\ divergence\ entre\ T.subterraneus\ et\ P.transmontana = 57{,}6\ Ma$:

$$T_d = 1{,}08T_f$$

$$T_f = \frac{57{,}6}{1{,}08} = 53\ Ma$$

$$T_p = T_d - T_f = 4{,}6\ millions\ d'années$$

5) Comparaison des différentes méthodes de datation

	Li et Nei 1977	Li et al. 1981	Meredith et al. 2009	Mutpred2
L. dentata	368 000 g	0,93 Ma	1 Ma	1,22 Ma
T. subterraneus	570 000 g	1,74 Ma	1,2 Ma	4,6 Ma
L. lethops	29 000 g			
A. mexicanus	8 600 g			

Tableau 7 : Résumé des résultats de datations de colonisations des grottes, ou des profondeurs du Congo pour L. lethops, avec les différentes méthodes présentées. Seule la méthode de Li et Nei 1977 a été utilisée pour L. lethops et A. mexicanus face à leur très faible taux de divergence avec l'espèce de surface la plus proche. g = générations, Ma = Millions d'années.

Bien que le temps de génération soit inconnu pour l'espèce *L. dentata*, si l'on fait l'hypothèse que celui-ci est d'environ trois ans, les différentes méthodes de datation donnent des résultats convergents (Tableau 7). Cette espèce pourrait avoir colonisé les grottes au milieu du Pléistocène (période qui s'étend entre 2,6 Ma et 12 000 ans avant le temps présent).

Pour *T. subterraneus*, le temps de génération a été estimé et se situe entre deux et trois ans[35]. Ainsi, la méthode de Li et Nei (1977) est très cohérente avec les résultats des méthodes de Li et al. (1981) et Meredith et al. (2009) et correspondrait, comme *L. dentata*, à une colonisation des grottes au milieu du Pléistocène. Cependant, la datation à l'aide de l'effet délétère des substitutions non-synonymes semble s'écarter de ces estimations avec un temps trois fois plus grand. Cela est dû à une surestimation du nombre de substitutions apparues sous évolution neutre (Tableau 7).

Les colonisations des profondeurs de rivière par *L. lethops* et des grottes par *A. mexicanus* semblent être beaucoup plus récentes, il y a environ 30 000 générations et 9 000 générations respectivement. L'estimation pour *A.* mexicanus, dont le temps de génération est estimé à 3,5

[35] https://www.fishbase.se/summary/Typhlichthys-subterraneus

ans (Simon et al. 2017), est cohérente avec une datation récente de 20 000 ans (Fumey et al. 2018), qui correspondrait à la fin de Pléistocène. Le temps de génération n'est pas connu pour *L. lethops*, mais si l'on estime celui-ci à environ deux ans comme chez d'autres espèces de Cichlidés (Won et al. 2005), un temps de colonisation il y a 30 000 générations correspondrait à 60 000 ans, également à la fin du Pléistocène.

6) Comparaison avec les mammifères souterrains

En 2014, une étude s'est intéressée à la perte des gènes de la vision chez les mammifères souterrains (Emerling et Springer 2014), notamment chez la taupe dorée du cap (*Chrysochloris asiatica*), le rat-taupe nu (*Heterocephalus glaber*) et le condylure étoilé (*Condylura cristata*). Ces espèces semblent avoir colonisé les environnements souterrains il y a longtemps, entre le début et la fin de l'Oligocène[36].

Ces animaux diffèrent des poissons cavernicoles car ils possèdent toujours des yeux fonctionnels et vont parfois à la surface. *C. cristata* peut par exemple régulièrement quitter son terrier et ainsi être exposée à la lumière (Chapman et Mark 2007), ce qui peut expliquer la présence de seulement six pseudogènes parmi 65 gènes de la vision chez cette espèce. En comparaison, *H. glaber* et *C. asiatica* possèdent 12 et 17 pseudogènes respectivement.

Les distributions des mutations perte de fonction sur les pseudogènes de la photoréception ont été comparées entre les poissons aveugles et ces mammifères. *A. mexicanus* et *L. lethops* ne sont pas inclus dans l'analyse en raison du très petit nombre de mutations chez ces espèces. De plus, seulement les gènes inactivés en commun dans les deux groupes d'espèces ont été inclus.

[36] Période qui s'étend entre 34 et 23 millions d'années avant le temps présent

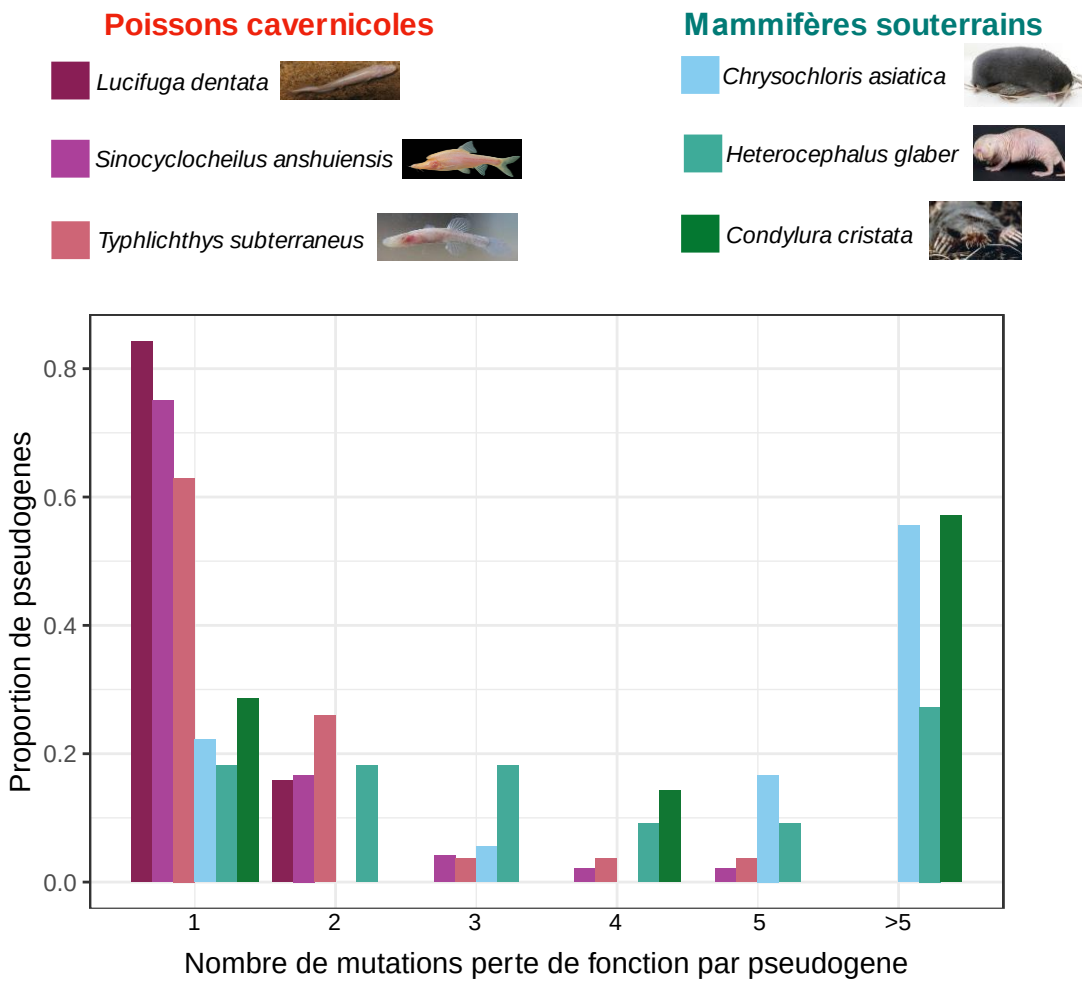


Figure 74 : Distributions du nombre de mutations perte de fonction par pseudogène de la vision chez les poissons cavernicoles et les mammifères souterrains.

Il y a une différence notable entre les distributions observées chez les mammifères souterrains et les poissons cavernicoles. Les poissons ont un grand nombre de pseudogènes mais avec peu de mutations sur chacun d'eux, indiquant une relaxation de sélection récente et au même moment sur une grande proportion des gènes de la photoréception.

Au contraire, chez les mammifères, une faible proportion de pseudogènes est retrouvée mais avec un grand nombre de mutations sur chacun d'eux. Chez la taupe dorée, jusqu'à 28 mutations

perte de fonction ont été trouvées sur un seul gène, tandis que chez le condylure étoilé, le gène *CRB1* contient 54 mutations perte de fonction. Cependant, il existe également des pseudogènes avec seulement une ou deux mutations chez ces espèces. Cela indique une relaxation de la sélection ancienne sur un sous-groupe des gènes de la vision suivie par des relaxations secondaires et plus récentes pour d'autres gènes. Contrairement aux poissons cavernicoles, chez ces espèces qui ont toujours des yeux, les gènes de la vision semblent en majorité toujours sous sélection purificatrice.

J- Discussion

Nous avons vu que la grande majorité des gènes de la pigmentation semble toujours évoluer sous sélection chez les poissons cavernicoles. Un très faible nombre de pseudogènes a été observé dans ce groupe de gènes, et de nombreuses pertes récurrentes des mêmes gènes se sont produites au cours de l'évolution, notamment chez *S. rhinocerous* et *S. anshuiensis* qui ont perdu indépendamment les mêmes deux paires de gènes ohnologues. Cette hypothèse est également supportée par les distributions des effets délétères des mutations non-synonymes, qui sont similaires entre les espèces de surface et les espèces troglomorphes. En cohérence avec les analyses fonctionnelles des gènes de ce groupe, qui montrent que la plupart ont des effets pléiotropiques, la perte d'un grand nombre de ces gènes semble impossible, même en l'absence de pigmentation.

De la même façon, très peu de pseudogènes ont été retrouvés parmi les gènes du cycle circadien. Seulement deux classes de gènes, pour lesquelles la transcription est activée par la lumière, sont perdues : certains gènes de la famille des *cryptochromes* et le gène *per2*, dans la famille des gènes *period*. Ce dernier est perdu chez *P. andruzzii* (Ceinos et al. 2018), espèce ayant perdu le cycle circadien, et est également inactivé chez *S. rhinocerous* et *T. subterraneus*, ce qui pourrait indiquer une perte de ce cycle chez ces espèces. Cette hypothèse est soutenue par l'accumulation de mutations non-synonymes délétères chez *S. rhinocerous*, pour lesquelles la distribution des scores Mutpred2 se rapproche d'une distribution sous évolution neutre.

L'analyse du groupe de la photoréception indique que la majorité de ces gènes semble évoluer de façon neutre chez les poissons aveugles. Le faible nombre de mutations perte de fonction indiquerait donc une colonisation récente des grottes par certaines espèces, entre le milieu et la fin du Pléistocène. Cette estimation est cohérente avec différentes méthodes qui se basent sur les nombres de mutations synonymes et non-synonymes. De plus, une nouvelle méthode développée à l'aide du programme Mutpred2 permet d'avoir une estimation du nombre de mutations délétères apparues lorsque la pression de sélection s'est relâchée sur ces gènes. Il faut cependant noter que pour l'espèce *Astyanax mexicanus*, la présence d'un seul pseudogène pourrait être due à des migrations récurrentes de poissons de surface de la même espèce (Bradic et al. 2012), apportant des allèles fonctionnels dans les populations cavernicoles et remplaçant les mutations perte de fonction en cours de fixation.

Notre étude va dans le sens de l'hypothèse selon laquelle la colonisation des grottes est un cul-de-sac évolutif pour les poissons, les espèces cavernicoles ne pouvant pas persister dans ces environnements pendant plus de quelques millions d'années. L'observation d'un grand nombre de téléostéens cavernicoles s'expliquerait alors par une certaine facilité d'adaptation à ces environnements. Cette hypothèse est notamment supportée par une publication récente, où des morphotype de surface d'*Astyanax mexicanus* qui ont été élevés dans l'obscurité pendant deux ans montrent déjà des signes d'adaptation, avec des changements d'activité de gènes régulant le métabolisme et le comportement (Bilandžija et al. 2020). Il se peut alors que les espèces observées actuellement dans les grottes soient celles dotées de la meilleure capacité d'adaptation à ce milieu. Les rivières où vivent les poissons de surface d'*Astyanax mexicanus* sont par exemple également peuplées d'autre espèces de téléostéens notamment d'autres Characiformes ou des Cichliformes (observations personnelles). Pourtant, dans les grottes où vivent les morphotypes cavernicoles d'*A. mexicanus*, aucune autre espèce de poissons n'est présente (Simon, Hyacinthe, et Rétaux 2019). Ceci indique que celles-ci n'ont pas réussi à survivre ni à se reproduire dans ces environnements. Il serait intéressant de savoir si l'absence de cohabitation par différents poissons est vraiment due à un différentiel de capacité d'adaptation ou si, après la colonisation d'une première espèce, les autres ne peuvent plus s'y reproduire par désavantage compétitif.

Seule l'observation d'un poisson troglomorphe ayant accumulé un très grand nombre de pseudogènes, chacun contenant un grand nombre de mutations perte de fonction, permettra de réfuter l'hypothèse du cul-de-sac évolutif. De plus, le séquençage d'autres espèces cavernicoles pourra indiquer de façon plus précise si les gènes de la photoréception sont bien tous sous un régime d'évolution neutre. Si ce n'est pas le cas, cela permettra de savoir quels gènes ont de potentielles fonctions autres que la réception et l'intégration des signaux lumineux.

Ces colonisations relativement récentes signifient que les adaptations à cet environnement obscur peuvent être très rapides, impliquant probablement une pression de sélection sur certains gènes du développement mais également sur des gènes liés à des modalités sensorielles utiles, tels que l'olfaction et la mécanoréception. Cependant, les parts relatives de sélection et de dérive génétique, particulièrement forte chez ces petites populations, dans l'évolution des caractères troglomorphes restent encore mal connus (Casane et Rétaux 2016).

Nos résultats montrent qu'il n'est pas possible d'utiliser les observations des phénotypes pour estimer l'âge des populations cavernicoles (méthode par exemple utilisée dans Wilkens, Strecker, et Yager 1989) : différentes espèces cavernicoles de notre étude montrent une absence d'yeux fonctionnels et de pigmentation. Elles ont pourtant colonisé les grottes à des périodes très différentes, il y a quelques milliers d'années pour *A. mexicanus* et il y a plusieurs millions d'années pour *T. subterraneus*. Le degré de troglomorphisme n'est donc pas informatif pour discuter le tempo de colonisation du milieu souterrain.

Il semble aussi ne pas y avoir de corrélation entre les différentes adaptations phénotypiques, comme l'atteste une perte du cycle circadien probable chez *S. rhinocerous* sans perte des yeux. Cette espèce semi-cavernicole, tout comme *L. gibarensis*, n'est pas nécessairement une espèce en cours de transition vers une vie complètement cavernicole. Le phénotype intermédiaire qui les caractérise pourrait correspondre à une adaptation à des paramètres écologiques différents, comme par exemple une vie principalement dans l'obscurité mais toujours avec une nécessité de détecter la présence de lumière pour se nourrir ou se reproduire dans certaines circonstances. *Lucifuga gibarensis* a notamment été observé dans la mer, proche des grottes. Cette hypothèse

est supportée par le fait que la majorité des gènes de la photoréception semble toujours sous sélection chez ces deux espèces[37] , la vision étant donc probablement toujours sous sélection.

VI) Étude et échantillonnage de populations sauvages d'Astyanax mexicanus

Afin d'étudier plus en détail les deux modalités sensorielles décrites dans cette thèse, la photoréception et l'olfaction, chez *Astyanax mexicanus*, des missions de terrains sont effectuées au Mexique depuis quelques années, la dernière, à laquelle j'ai eu la chance de participer, ayant eu lieu en 2019.

A- Étude des capacités olfactives

En condition de laboratoire, le seuil de détection de l'alanine est 100 000 fois plus bas chez les morphotypes cavernicoles que chez les morphotypes de surface d'*A. mexicanus*. Les poissons cavernicoles peuvent détecter les acides aminés, notamment l'alanine, à une concentration de 10^{-10} M, mais les poissons de surface ne la détectent plus en dessous de 10^{-5} M (Hinaux et al. 2016). Les expériences ayant conduit à ces observations n'avaient été réalisées que sur des poissons élevés en laboratoire, et provenant tous de la grotte Pachón.
Il était donc important d'évaluer si ce phénotype est restreint aux individus élevés en conditions de laboratoire, qui sont évidemment différentes des conditions naturelles, et si ce seuil de détection est également trouvé chez des poissons issus d'autres populations cavernicoles.

Dans cette optique, des expériences comportementales ont été réalisées directement dans les grottes de Pachón, Sabinos, Tinaja Chica et Subterraneo, durant les expéditions de 2016, 2017 et 2019. Des piscines gonflables, facilement transportables, ont été apportées dans les

[37] Résultats des analyses des mutations perte de fonction et non-synonymes.

différentes grottes, puis gonflées et remplies d'eau de ces mêmes grottes. Huit poissons cavernicoles ont été placés dans celles-ci, et trois molécules odorantes ont été perfusées dans l'eau successivement, chacune étant précédée et suivie d'une perfusion d'eau. Pour éviter un comportement d'attraction simplement dû à un flux de liquide, chaque perfusion d'odeur était accompagnée d'une perfusion d'eau dans le coin opposé (Figure 75). De plus, afin éviter un biais de position, les trois molécules odorantes ont été perfusées dans des coins différents. Les expériences ont été réalisées sous un éclairage infra-rouge auquel ne sont pas sensibles ces poissons, et leur comportement a été filmé avec une caméra infra-rouge (Blin et al. 2020).

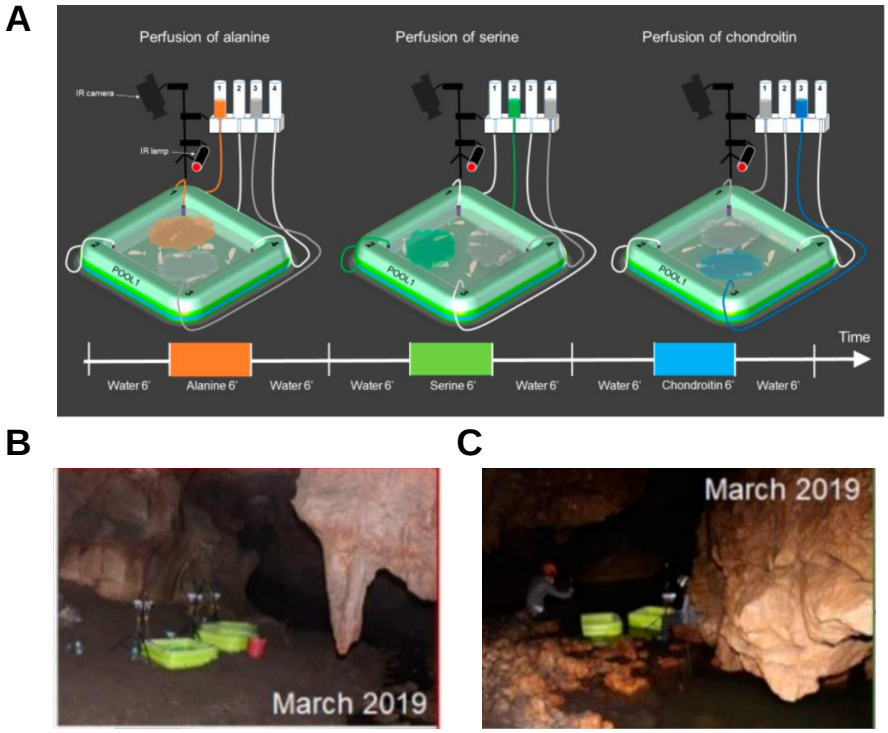


Figure 75 : Expérience comportementale dans les grottes.
A- Configuration de l'expérience comportementale pour évaluer les capacités olfactives des poissons sauvages. Les trois odeurs perfusées sont deux acides aminés, l'alanine et la sérine, ainsi qu'une molécule faisant partie de la substance d'alarme : la chondroïtine. B- Piscines gonflables placées dans la grotte Pachón. B- Piscines gonflables placées dans la grotte Subterraneo. Modifié d'après Blin et al. (2020).

De façon cohérente avec les observations en laboratoire, les individus de Pachón détectent des concentrations plus faibles que les poissons de surface, mais quelques différences

comportementales ont été observées entre ces deux conditions. Par exemple, les poissons sauvages ont un délai de réponse face à la perfusion de sérine, alors qu'en laboratoire, l'attraction est immédiate.

Ces expériences ont également montré de grandes différences de réponses comportementales entre les différentes populations, avec deux populations qui semblent bien détecter les odeurs perfusées (Pachón et Subterraneo), une population qui semble répondre de façon faible (Sabinos) et deux populations qui ne répondent pas du tout (Tinaja et Chica). De nouvelles études devront être réalisées afin de comprendre cette hétérogénéité des comportements, qui peut être due à des différences environnementales ou génétiques.

En outre, les expériences sur les populations de laboratoire et les populations sauvages sont réalisées en groupe, ce qui ne permet pas d'évaluer la variabilité entre individus d'une même population.

B- Échantillonnage d'ADN

Ces différentes expéditions ont permis d'obtenir des échantillons d'ADN d'individus issus des différentes populations. Ces banques d'ADN nous ont notamment permis de génotyper les délétions génomiques observées dans le génome d'*A. mexicanus* cavernicole (correspondant à la population originaire de Pachón), dans les opsines non-visuelles *rgr2*, *opn8a* et *opn8b* (Figure 53).

Afin d'évaluer la présence ou l'absence de ces délétions dans les différentes populations, des amplifications par PCR ont été réalisées d'un exon contrôle (non délété) et d'un exon dans les régions délétées de ces gènes. Si l'amplification de ce dernier échoue, cela signifie que le poisson est homozygote délété. S'il y a une amplification correcte, le poisson est soit homozygote non délété, soit hétérozygote (un chromosome avec la délétion, un chromosome sans) (Figure 76). L'exon contrôle nous permet de vérifier qu'une éventuelle absence d'amplification ne résulte pas d'un problème expérimental.

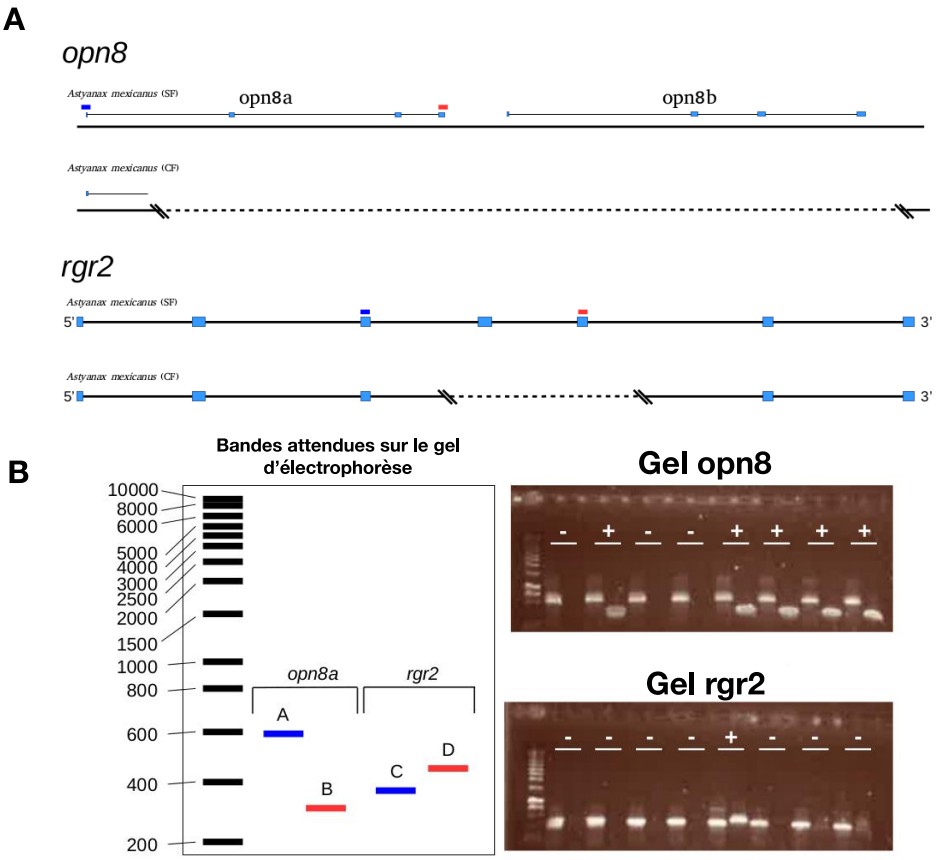


Figure 76 : Génotypage de deux grandes délétions chez A. mexicanus.
A- Deux délétions impactent des opsines non-visuelles chez le morphotype cavernicole d'A. mexicanus : une délétion impactant les trois derniers exons d'opn8a et le gène opn8b en entier, une délétion sur deux exons de rgr2. Pour la première délétion (opn8), une PCR contrôle est réalisée sur le premier exon d'opn8a et une PCR est réalisée sur le dernier exon d'opn8a afin d'évaluer si la délétion est présente ou non. Pour la deuxième délétion, la PCR contrôle est réalisée sur le troisième exon de rgr2 et la deuxième PCR est réalisée sur le cinquième exon. B- Bandes attendues et observées sur le gel d'électrophorèse pour les quatre PCR. On peut voir que les PCR contrôles résultent toutes en une amplification tandis qu'en fonction de l'individu, la PCR sur les exons délétés résulte en une amplification ou non. Les individus homozygotes délétés sont indiqués par un « - » tandis que les homozygotes non délétés ou les hétérozygotes sont indiqués par un « + ».

Ces expériences nous ont permis de démontrer que la délétion de *rgr2* était partagée par toutes les populations cavernicoles, à l'exception de la population Subterraneo. La délétion sur les gènes *opn8* est quant à elle restreinte aux populations de la région d'El Abra. De plus, à l'aide de génomes sous forme de lecture, disponibles pour des poissons issus des grottes Tinaja, Molino et Pachón (Herman et al. 2018), nous avons pu identifier de nombreuses mutations perte de fonction sur des gènes de la vision (*rh1.2*, *pde6b*, *cryba2a*, *impg1b*, *prph2b*, *gja8a* et *lim2.1*), du cycle circadien (*per2*) et de la pigmentation (*tyrp1a*). Des amplifications PCR suivis de séquençages Sanger ont été réalisées afin d'évaluer la fréquence de ces différentes mutations dans les populations (Figure 77).

Population	rgr2	opn8	tyrp1a	rh1.2	pde6b	cryba2a	impg1b	per2(M)	prph2b	gja8a	per2(T)	lim2.1
Jineo (n=6)	1	0	0	0	0	0	0	1	0	0	0	1
Molino (n=9)	1	0	0	0	0	0	0	1	0	0	0	1
Escondido (n=7)	1	0	0	0	0	0	0	1	0	0	0	1
Caballo Moro (n=2)	1	0	0	0	0	0	0	0	0	0	0	0
Pachon (n=55)	0,95	0,66	0,69	0,036	0,76	0	0,24	0	0,15	0	0	0
Sabinos (n=33)	0,83	1	1	0,24	0	0,5	0,53	0	1	0,88	0	0
Tinaja (n=16)	1	0,87	0,88	0,44	0	0,16	0,56	0	0,94	0,69	0,125	0
Curva (n=23)	1	1	1	1	0	0	1	0	1	1	0	0
Toro (n=30)	0,75	0,73	0,05	0	0	0,65	0,18	0	0,75	0,65	0	0
Chica (n=33)	0,7	0,52	0,47	0	0	0,53	0	0	0	0	0	0
Subterraneo (n=26)	0	0	0	0	0	0	0	0	0	0	0	0
Surface (n=234)	0	0	0	0	0	0	0	0	0	0	0	0

Figure 77 : Fréquence des délétions génomiques et des mutations perte de fonction sur les gènes de la vision (rouge), du cycle circadien (vert) ou de la pigmentation (bleu) dans les différentes populations.
Le gène per2 est impacté par deux mutations perte de fonction : une spécifique aux populations de Guatemala appelée per2(M) et une spécifique à la population de Tinaja (appelée per2(T)).

Quelques mutations perte de fonction sont spécifiques à une seule population : une mutation dans *pde6b* (Pachón) et une mutation dans *per2* (Tinaja). Cependant, la plupart des mutations sont partagées par les populations au sein des deux régions : Guatemala (Jineo, Molino,

Escondido[38]) et El Abra (Pachón, Sabinos, Tinaja, Curva, Toro et Chica). La délétion de *rgr2* est retrouvée dans toutes les grottes échantillonnées, sauf Subterraneo.

Cette répartition géographique des mutations semble indiquer que la colonisation des grottes suit un modèle « stepping-stone », c'est-à-dire une colonisation unique suivie par une dispersion dans les différentes grottes des régions Guatemala et El Abra respectivement, dans lesquelles les populations ont ensuite évolué de façon indépendante. Une colonisation indépendante de la région Micos est envisageable sur la base de nos données.
Cela remet notamment en question l'hypothèse des cinq colonisations indépendantes (Bradic et al. 2012). Il est improbable que les mêmes mutations perte de fonction soient apparues dans les différentes populations, ou que celles-ci aient été apportées par des migrations de poissons de surface, chez qui leur présence serait contre sélectionnée [39] (Figure 77).

[38] Caballo Moro fait aussi parti de la région Guatemala mais nous disposons de l'ADN d'uniquement deux individus.
[39] Corroboré par le fait que ces mutations n'ont jamais été retrouvées en surface, avec plus de 200 individus génotypés.

VII) Conclusion générale

Cette thèse m'a permis d'étudier deux modalités sensorielles, que sont l'olfaction et la vision, et à deux échelles : au niveau de tous les génomes de poissons à nageoires rayonnées disponibles et à celle d'un petit groupe d'espèces en relation avec l'adaptation à la vie cavernicole.

D'une part, j'ai montré que chez les poissons à nageoires rayonnées, il existait une corrélation entre la morphologie de l'épithélium olfactif et le nombre de récepteurs olfactifs dans le génome. Cependant, je me suis rendu compte qu'il y avait un grand manque de données écologiques, morphologiques ou comportementales dans la littérature par rapport au nombre de séquences génomiques disponibles. Sur 185 espèces étudiées, nous connaissons le nombre de lamelles dans l'épithélium pour moins de la moitié (72) et il n'existe presque aucune donnée sur le bulbe olfactif. Il sera intéressant dans le futur d'étudier la dynamique de ces récepteurs olfactifs dans d'autres groupes, tels que les requins, les oiseaux ou les amphibiens, chez qui peu d'études comparatives ont été menées. De plus, je me penche actuellement sur une automatisation de la recherche des gènes *TAAR*, *ORA* et *OlfC* qui possèdent plusieurs exons, et qui impliquaient jusqu'ici une grosse part de travail manuel.

D'autre part, notre étude de gènes chez les poissons cavernicoles était limitée par le très petit nombre de mutations présentes chez ces espèces par rapport aux espèces de surface. Nous avons développé de nouvelles méthodes pour estimer si un groupe de gènes évolue de façon neutre pour ensuite estimer le temps depuis lequel la pression de sélection s'est relâchée. En appliquant ces dernières sur les génomes de poissons cavernicoles disponibles, nous avons montré que les colonisations des grottes par ces espèces étaient relativement récentes. Nos résultats confirment une étude précédente de l'équipe (Fumey et al. 2018) sur *Astyanax mexicanus*, qui semble avoir colonisé les grottes il y a quelques milliers d'années seulement. De plus, nos données de génotypage de pseudogènes semblent indiquer que cette espèce a colonisé les grottes seulement une ou deux fois indépendamment. Avec 300 téléostéens (Proudlove 2021) cavernicoles connus, nos méthodes pourront être élargies à une grand nombre d'espèces afin de confirmer les résultats obtenus et avoir une vision plus globale des modalités temporelles de la colonisation des grottes par les poissons.

Table des figures

Figure 1 : Phylogénie des vertébrés. 15

Figure 2 : Phylogénie des ordres d'actinoptérygiens avec quelques espèces représentatives de la diversité existante. 17

Figure 3 : Phylogénie avec cinq espèces représentatives des Nototheniidae. 19

Figure 4 : Photographique du cabillaud arctique *A. glacialis* et carte de sa répartition 21

Figure 5 : Exemples de mutations perte de fonction et leur conséquence sur la séquence protéique. 27

Figure 6 : Phylogénie des Ursidés raciné avec Canis lupus comme groupe externe. 29

Figure 7 : A- Gains et pertes secondaires de l'électroperception chez les vertébrés. B- Ictalurus punctatus (Siluriforme) capable d'électroperception passive C- Electrophorus electricus (Gymnotiforme), capable d'électroperception active. ... 31

Figure 8 : Routes migratoires des saumons de l'atlantique (Salmo salar). 33

Figure 9 : Sonagrammes de 6 sons produits par l'espèce Astyanax mexicanus. 35

Figure 10 : Nombre de bourgeons gustatifs chez trois téléostéens (Danio rerio, Phoxinus phoxinus et Ictalurus punctatus) et trois tétrapodes (Cyanistes caeruleus, Homo sapiens et Sus scrofa) 37

Figure 11 : Système olfactif des actinoptérygiens. 39

Figure 12 : Schéma du système olfactif chez les téléostéens. 40

Figure 13 : Arbre phylogénétique des quatre familles de récepteurs olfactifs, extraites du poisson-zèbre (Danio rerio) et du Fugu (Takifugu rubripes). 43

Figure 14 : Exemples d'espèces ayant subi une expansion ou une contraction du répertoire de gènes OR. 45

Figure 15 : Filtrage des génomes selon leur taille d'assemblage. 49

Figure 16 : Nombre de génomes en fonction d'un seuil de pourcentage de gènes BUSCO retrouvés. 50

Figure 17 : Comparaison phylogénétique des gènes OlfC chez Takifugu rubripes. 55

Figure 18 : Phylogénie des 185 espèces d'actinoptérygiens étudiés avec leur nombre de récepteurs olfactifs fonctionnels dans chaque famille. 57

Figure 19 : Corrélations du nombre de pseudogènes et du nombre de gènes tronqués ou gènes de bordures par famille. 58

Figure 20 : Phylogénie des 185 espèces d'actinoptérygiens étudiés avec leur nombre de récepteurs olfactifs fonctionnels dans la famille OR. 61

Figure 21 : Phylogénie en maximum de vraisemblance des gènes OR issus de 44 actinoptérygiens (représentant 44 ordres). 63

Figure 22 : Phylogénie des 185 espèces d'actinoptérygiens étudiés avec leur nombre de récepteurs olfactifs fonctionnels dans la famille TAAR. 65

Figure 23 : Phylogénie des 185 espèces d'actinoptérygiens étudiées avec leur nombre de récepteurs olfactifs fonctionnels dans la famille OlfC. 66

Figure 24 : Phylogénie en maximum de vraisemblance des gènes TAAR issus des 185 actinoptérygiens de l'étude. 67

Figure 25 : Phylogénie en maximum de vraisemblance des gènes OlfC issus des 185 actinoptérygiens de l'étude..... 68

Figure 26 : Phylogénie en maximum de vraisemblance des gènes ORA issus des 185 actinoptérygiens de l'étude..... 70

Figure 27 : Phylogénie des 185 espèces d'actinoptérygiens étudiés avec leur nombre de récepteurs olfactifs fonctionnels dans la famille ORA..... 71

Figure 28 : Régressions linéaires phylogénétiques entre le nombre de gènes fonctionnels et la proportion de pseudogènes chez les 185 espèces d'actinoptérygiens. 72

Figure 29 : Résultats des simulations d'évolution neutre d'un gène OR pendant N générations. 76

Figure 30 : Variation du coefficient de corrélation de Pearson, entre le nombre de gènes perdus estimé par la méthode de réconciliation de l'arbre de gènes et l'arbre des espèces, et le nombre de gènes inactifs observé. 77

Figure 31 : Taux de naissance et de perte des récepteurs olfactifs estimées sur toutes les branches de l'arbre des espèces de l'étude (368 branches). 79

Figure 32 : Corrélations de Pearson entre le nombre de gains ou nombre de pertes de gènes des familles OR, TAAR et OlfC sur tout l'arbre phylogénétique. 80

Figure 33 : Corrélation entre le nombre total de récepteurs olfactifs et la vie en cours d'eau. 82

Figure 34 : Catégories utilisées pour définir les épithéliums olfactifs et exemples du nombre de lamelles chez quelques téléostéens. 84

Figure 35 : Reconstruction de l'état ancestral de l'épithélium olfactif. 86

Figure 36 : Épithélium à plusieurs lamelles sans forme de rosette. 88

Figure 37 : Arbre phylogénétique des 72 espèces pour lesquelles des données sont disponible pour la morphologie de l'épithélium olfactif et pour le nombre de récepteurs olfactifs 89

Figure 38 : Régressions linéaires phylogénétiques entre le nombre de lamelles et A- le nombre de gènes OR fonctionnels B- le nombre de récepteurs olfactifs fonctionnels C- le nombre de gènes TAAR fonctionnels D- le nombre total de récepteurs olfactifs fonctionnels. 90

Figure 39 : Schéma de l'organe olfactif et de l'œil de Polypterus senegalus. 94

Figure 40 : Phylogénie des tétraodontiformes avec Lophius piscatorius en outgroup. 95

Figure 41 : Morphologie de l'œil chez les vertébrés. 98

Figure 42 : Exemples de stratégies de camouflage utilisées par les téléostéens. 100

Figure 43 : Schéma récapitulant les différents gènes impliqués dans la phototransduction et dans le maintien du cristallin. 102

Figure 44 : Schéma représentant la boucle rétroactive du cycle circadien. 105

Figure 45 : Photographies d'organismes cavernicoles issus, de gauche à droite, d'une grotte marine, d'une grotte karstique et d'un tunnel de lave. 107

Figure 46 : Évolution phénotypique chez les organismes cavernicoles. 109

Figure 47 : Différence entre temps de divergence et temps de colonisation des grottes 112

Figure 48 : Principe d'une analyse QTL chez Astyanax mexicanus. 115

Figure 49 : Carte de répartition des populations cavernicoles d'Astyanax mexicanus 118

Figure 50 : Phylogénie des espèces chez lesquelles les différents gènes sont recherchés dans notre étude. 120

Figure 51 : Localisation, taille des yeux et score BUSCO pour les espèces de téléostéens cavernicoles avec un génome disponible, sous forme d'assemblage ou non. 122

Figure 52 : Méthodologie employée pour la recherche des gènes sur les génomes d'intérêts. 125

Figure 53 : Exons retrouvés ou non par PCR parmi les gènes incomplets du génome du morphotype cavernicole d'Astyanax mexicanus 128

Figure 54 : Mutation perte de fonction sur pde6b. 129

Figure 55 : Phylogénie des espèces étudiées avec les nombres de gènes et pseudogènes de la photoréception retrouvés. 135

Figure 56: Phylogénie des espèces étudiées avec les nombres de gènes et pseudogènes du cycle circadien retrouvés 136

Figure 57 : Phylogénie des espèces étudiées avec les nombres de gènes et pseudogènes de la pigmentation retrouvés 138

Figure 58 : Distribution de la taille des insertions ou délétions provoquant un décalage du cadre de lecture (=frameshift) chez les espèces étudiées 140

Figure 59 : Fréquences observées et théoriques des mutations perte de fonction étudiées. .. 142

Figure 60 : Distribution théorique des tailles effectives de segmentation issues de 100 000 simulations, où 60 mutations sont réparties au hasard le long des séquences. 144

Figure 61 : Groupes de gènes étudiés et pseudogènes observés chez les espèces cavernicoles. 146

Figure 62 : Distributions observées et simulées du nombre de mutations perte de fonction par gènes chez les poissons cavernicoles. 152

Figure 63 : Distribution observée et simulée du nombre de mutations perte de fonction par gènes chez le poisson semi-cavernicole *L. gibarensis*. 153

Figure 64 : Nombre de gènes et pseudogènes en commun entre *L. dentata* et *T. subterraneus*. 155

Figure 65 : Modèles testés pour les trois groupes de gènes. 159

Figure 66 : Valeurs du dS, dN et dN/dS pour les espèces étudiées sur les gènes de la photoréception sous le modèle ratios libres. 160

Figure 67 : Valeurs du dS, dN et dN/dS pour les espèces étudiées sur les gènes du cycle circadien et de la pigmentation, sous le modèle ratios libres. 162

Figure 68 : Distribution cumulée des scores mutpred2 en fonction du ratio transition/transversion et du groupe de gènes. 164

Figure 69 : Distributions cumulées des scores Mutpred2 pour chaque espèce de l'étude et chaque groupe de gènes. 165

Figure 70 : p-values du test de Kolmogorov-Smirnov entre les distributions observées pour chaque espèce et des mélanges de la distribution sous évolution neutre et de la distribution sous sélection en différentes proportions. 167

Figure 71 : Distribution cumulée des scores Mutpred2 pour les espèces du genre Sinocyclocheilus et p-value du test de Kolmogorov-Smirnov entre les distributions observées et des mélanges de la distribution neutre et de la distribution sous sélection. 168

Figure 72 : Datations de la relaxation de sélection sur les gènes de la photoréception chez les espèces cavernicoles. 173

Figure 73 : Arbres phylogénétiques utilisés pour la datation de *L. dentata* et *T. subterraneus*. 175

Figure 74 : Distributions du nombre de mutations perte de fonction par pseudogènes de la vision chez les poissons cavernicoles et les mammifères souterrains. 182

Figure 75 : Expérience comportementale dans les grottes. 187

Figure 76 : Génotypage de deux grandes délétions chez A. mexicanus. 189

Figure 77 : Fréquence des délétions génomiques et des mutations perte de fonction sur les gènes de la vision (rouge), du cycle circadien (vert) ou de la pigmentation (bleu) dans les différentes populations. 190

Tableau 1 : Comparaison du nombre de gènes fonctionnels (F), incomplets (I) ou pseudogènes (P) entre notre étude et les études précédentes sur quatre espèces de téléostéens. 54

Tableau 2 : Longueurs d'ondes absorbées par les différentes familles d'opsines présentes chez les vertébrés. 103

Tableau 3 : Couverture des mutations perte de fonction sur les gènes de la photoréception chez les deux espèces de *Lucifuga*. 132

Tableau 4 : Couverture des mutation perte de fonction sur les gènes de la photoréception chez Lamprologus lethops. 133

Tableau 5 : Gènes sous-exprimés chez une ou plusieurs espèces de poissons cavernicoles.. 149

Tableau 6 : Répartition des mutations perte de fonction chez les poissons cavernicoles L. dentata et T. subterraneus et chez le poisson semi-cavernicole L. gibarensis..... 151

Tableau 7 : Résumé des résultats de datations de colonisations des grottes, ou des profondeurs du Congo pour *L. lethops*, avec les différentes méthodes présentées. 180

Crédits photographiques et graphiques

Figure 2 : JJ Harrison (*Boleophthalmus boddarti*), Per-Ola Norman (*Mola mola*), Bae Jiwon (*Hippocampus haema*), Karel Jakubec (*Symphysodon aequifasciatus*), James St. John (*Carassius auratus*), Ezra Weller (*Macropinna microstoma*) and the Smithsonian Environmental Research Center (*Aphredoderus sayanus*)

Figure 4 : Claude Nozères (*Arctogadus glacialis*)

Figure 7 : Eric Engbretson (*Ictalurus punctatus*) et Steven G. Johnson (*Electrophorus electricus*)

Figure 10 : Maximilian Dorsch (*Cyanistes caeruleus*), Azul (*Danio rerio*), Cliff (*Ictalurus punctatus*), Emőke Dénes (*Phoxinus phoxinus*) et Dave Pape (*Sus scrofa*)

Figure 14 : Charles J. Sharp (*Loxodonta africana*), Holger Krisp (*Sus scrofa*), Aditya Joshi (*Megaderma lyra*) et la NASA (*Tursiops truncatus*).

Figure 20 : Michał Zalewski (*Erpetoichthys calabaricus)*, Per-Ola Norman (*Mola mola*)

Figure 22 : Michał Zalewski (*Erpetoichthys calabaricus)*, TinoStrauss *(Callionymus lyra*)

Figure 23 : Michał Zalewski (*Erpetoichthys calabaricus)*, Per-Ola Norman (*Mola mola*)

Figure 34 : Figure réalisée par Didier Casane.

Figure 41 : James St. John (*Anableps anableps*) et Pavel Riha (*Diretmus argenteus*)

Figure 42 : Stefano Guerrieri (*Aphia minuta*) ; Alessandro Duci (*Sardina pilchardus*) ; (Grimsditch, Basheer, et Bryant 2017) (*Antennarius maculatus*) ; (Johnsen 2001) (*Bathophilus* sp.) ; (Sazima et al. 2006) (*Steatogenys duidae*)

Figure 45 : Thomas M. Iliffe et Renée E. Bishop (*Pelagomacellicephala iliffei*), U.S. Fish and Wildlife Service (*Amblyopsis rosae*), Jut Wynne (*Tiphia andersoni*)

Figure 46 : Arne Hodalič (*Proteus anguinus*) et Jorge Martínez Huelves (*Euproctus asper*)

Figure 49 : Carte réalisée par Laurent Legendre.

Figure 51 : Shu et al. 2013 (*Sinocyclocheilus rhinocerous*), H. Zell (*Astyanax mexicanus*) , Melanie Stiassny (*Lamprologus lethops*) , Jorg Freyhof (*Sinocyclocheilus asnshuiensis*) , Niemiller et al. 2016 (*Typhlichthys subterraneus*)

Figure 74 : JonRichfield (*Chrysochloris asiatica*), TxanTxunai (*Heterocephalus glaber*) , US National Parks Service (*Condylura cristata*)

Boites

Sélection naturelle …….. 24
Dérive génétique ……...…….. 25
BUSCO ……...…… 48
Signal phylogénétique ……....……. 60
Réconciliation arbres des gènes – arbre des espèces ……74
Le dN/dS ……....…….. 157

Références

Aardema, Matthew L, Melanie L J Stiassny, et S Elizabeth Alter. 2020. « Genomic Analysis of the Only Blind Cichlid Reveals Extensive Inactivation in Eye and Pigment Formation Genes ». Genome Biology and Evolution 12 (8): 1392-1406. https://doi.org/10.1093/gbe/evaa144.

Ahuja, Gaurav, et Sigrun Korsching. 2014. « Zebrafish olfactory receptor ORA1 recognizes a putative reproductive pheromone ». Communicative & Integrative Biology 7 (5): e970501. https://doi.org/10.4161/19420889.2014.970501.

Aljančič, Gregor. 2019. « History of research on Proteus anguinus Laurenti 1768 in Slovenia / Zgodovina raziskovanja človeške ribice (Proteus anguinus Laurenti 1768) v Sloveniji ». *Folia biologica et geologica* 60 . https://doi.org/10.3986/fbg0050.

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, et David J. Lipman. 1990. « Basic Local Alignment Search Tool ». *Journal of Molecular Biology* 215 (3): 403-10. https://doi.org/10.1016/S0022-2836(05)80360-2.

Alunni, Alessandro, Arnaud Menuet, Eva Candal, Jean-Baptiste Pénigault, William R. Jeffery, et Sylvie Rétaux. 2007. « Developmental mechanisms for retinal degeneration in the blind cavefish Astyanax mexicanus ». *Journal of Comparative Neurology* 505 (2): 221-33. https://doi.org/10.1002/cne.21488.

Andersson, Dan I., Jon Jerlström-Hultqvist, et Joakim Näsvall. 2015. « Evolution of New Functions De Novo and from Preexisting Genes ». *Cold Spring Harbor Perspectives in Biology* 7 (6). https://doi.org/10.1101/cshperspect.a017996.

Andley, Usha P. 2009. « Effects of Alpha-Crystallin on Lens Cell Function and Cataract Pathology ». *Current Molecular Medicine* 9 (7): 887-92. https://doi.org/10.2174/156652409789105598.

AquaMaps (2019, October). Computer generated distribution maps for Arctogadus glacialis (Arctic cod), with modelled year 2050 native range map based on IPCC RCP8.5 emissions scenario. Retrieved from https://www.aquamaps.org.

Aspiras, Ariel C., Nicolas Rohner, Brian Martineau, Richard L. Borowsky, et Clifford J. Tabin. 2015. « Melanocortin 4 Receptor Mutations Contribute to the Adaptation of Cavefish to Nutrient-Poor Conditions ». *Proceedings of the National Academy of Sciences* 112 (31): 9668-73. https://doi.org/10.1073/pnas.1510802112.

Atema, Jelle, Richard R. Fay, Arthur N. Popper, et William N. Tavolga, éd. 1988. *Sensory Biology of Aquatic Animals*. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4612-3714-3.

Audo, Isabelle, Susanne Kohl, Bart P. Leroy, Francis L. Munier, Xavier Guillonneau, Saddek Mohand-Saïd, Kinga Bujakowska, et al. 2009. « TRPM1 Is Mutated in Patients with Autosomal-Recessive Complete Congenital Stationary Night Blindness ». *American Journal of Human Genetics* 85 (5): 720-29. https://doi.org/10.1016/j.ajhg.2009.10.013.

Auvinet, J., P. Graça, L. Belkadi, L. Petit, E. Bonnivard, A. Dettaï, W. H. Detrich, C. Ozouf-Costaz, et D. Higuet. 2018. « Mobilization of retrotransposons as a cause of chromosomal diversification and rapid speciation: the case for the Antarctic teleost genus Trematomus ». *BMC Genomics* 19 (1): 339. https://doi.org/10.1186/s12864-018-4714-x.

Azzouzi, Naoual, Frederique Barloy-Hubler, et Francis Galibert. 2015. « Identification and characterization of cichlid TAAR genes and comparison with other teleost TAAR repertoires ». *BMC Genomics* 16 (1): 335. https://doi.org/10.1186/s12864-015-1478-4.

Barlow, George. 2008. *The Cichlid Fishes: Nature's Grand Experiment In Evolution*. Hachette UK.

Barreiro-Iglesias, Antón, Ramón Anadón, et María Celina Rodicio. 2010. « The Gustatory System of Lampreys ». *Brain, Behavior and Evolution* 75 (4): 241-50. https://doi.org/10.1159/000315151.

Baxter, Laura L., Dawn E. Watkins-Chow, William J. Pavan, et Stacie K. Loftus. 2019. « A Curated Gene List for Expanding the Horizons of Pigmentation Biology ». *Pigment Cell & Melanoma Research* 32 (3): 348-58. https://doi.org/10.1111/pcmr.12743.

Beale, Andrew, Christophe Guibal, T. Katherine Tamai, Linda Klotz, Sophie Cowen, Elodie Peyric, Víctor H. Reynoso, Yoshiyuki Yamamoto, et David Whitmore. 2013. « Circadian Rhythms in Mexican Blind Cavefish Astyanax Mexicanus in the Lab and in the Field ». *Nature Communications* 4 (1): 2769. https://doi.org/10.1038/ncomms3769.

Benjamini, Yoav, et Yosef Hochberg. 1995. « Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing ». *Journal of the Royal Statistical Society. Series B (Methodological)* 57 (1): 289-300.

Berti, Roberto, Jacques Pierre Durand, Samuele Becchi, Rossana Brizzi, Nicole Keller, et Gisèle Ruffat. 2001. « Eye degeneration in the blind cave-dwelling fish *Phreatichthys andruzzii* ». *Canadian Journal of Zoology* 79 (7): 1278-85. https://doi.org/10.1139/cjz-79-7-1278.

Betts, Holly C., Mark N. Puttick, James W. Clark, Tom A. Williams, Philip C. J. Donoghue, et Davide Pisani. 2018. « Integrated Genomic and Fossil Evidence Illuminates Life's Early Evolution and Eukaryote Origin ». *Nature Ecology & Evolution* 2 (10): 1556-62. https://doi.org/10.1038/s41559-018-0644-x.

Bi, Xupeng, Kun Wang, Liandong Yang, Hailin Pan, Haifeng Jiang, Qiwei Wei, Miaoquan Fang, et al. 2021. « Tracing the Genetic Footprints of Vertebrate Landing in Non-Teleost Ray-Finned Fishes ». *Cell* 184 (5): 1377-1391.e14. https://doi.org/10.1016/j.cell.2021.01.046.

Bilandžija, Helena, Lindsey Abraham, Li Ma, Kenneth J. Renner, et William R. Jeffery. 2018. « Behavioural changes controlled by catecholaminergic systems explain recurrent loss of pigmentation in cavefish ». *Proceedings of the Royal Society B: Biological Sciences* 285 (1878): 20180243. https://doi.org/10.1098/rspb.2018.0243.

Bilandžija, Helena, Breanna Hollifield, Mireille Steck, Guanliang Meng, Mandy Ng, Andrew D Koch, Romana Gračan, et al. 2020. « Phenotypic plasticity as a mechanism of cave colonization and adaptation ». *eLife* 9: e51830. https://doi.org/10.7554/eLife.51830.

Bleckmann, H. 2008. « Peripheral and Central Processing of Lateral Line Information ». *Journal of Comparative Physiology A* 194 (2): 145-58. https://doi.org/10.1007/s00359-007-0282-2.

Blin, Maryline, Julien Fumey, Camille Lejeune, Maxime Policarpo, Julien Leclercq, Stéphane Père, Jorge Torres-Paz, Constance Pierre, Boudjema Imarazene, et Sylvie Rétaux. 2020. « Diversity of Olfactory Responses and Skills in Astyanax Mexicanus Cavefish Populations Inhabiting Different Caves ». *Diversity* 12 (10): 395. https://doi.org/10.3390/d12100395.

Blin, Maryline, Eugène Tine, Lydvina Meister, Yannick Elipot, Jonathan Bibliowicz, Luis Espinasa, et Sylvie Rétaux. 2018. « Developmental Evolution and Developmental Plasticity of the Olfactory Epithelium and Olfactory Skills in Mexican Cavefish ». *Developmental Biology*, Cavefish Development, 441 (2): 242-51. https://doi.org/10.1016/j.ydbio.2018.04.019.

Bodznick, D., et R. G. Northcutt. 1981. « Electroreception in Lampreys: Evidence That the Earliest Vertebrates Were Electroreceptive ». *Science* 212 (4493): 465-67. https://doi.org/10.1126/science.7209544.

Bonferroni, Carlo E. 1936. *Teoria statistica delle classi e calcolo delle probabilità*. Libreria internazionale Seeber.

Borowsky, R., et H. Wilkens. 2002. « Mapping a Cave Fish Genome: Polygenic Systems and Regressive Evolution ». *The Journal of Heredity* 93 (1): 19-21. https://doi.org/10.1093/jhered/93.1.19.

Bradic, Martina, Peter Beerli, Francisco J. García-de León, Sarai Esquivel-Bobadilla, et Richard L. Borowsky. 2012. « Gene flow and population structure in the Mexican blind cavefish complex (Astyanax mexicanus) ». *BMC Evolutionary Biology* 12 (1): 9. https://doi.org/10.1186/1471-2148-12-9.

Burne, RH. 1909. « The anatomy of the olfactory organ of teleostean fishes ». *Proc Zool Soc Lond.* 2: 610-63.

Burr, HS. 1928. « The central nervous system of Orthagoriscus mola ». *J Comp Neurol.* 45(1): 33-128.

Calderoni, L., O. Rota-Stabelli, E. Frigato, A. Panziera, S. Kirchner, N. S. Foulkes, L. Kruckenhauser, C. Bertolucci, et S. Fuselli. 2016. « Relaxed Selective Constraints Drove Functional Modifications in Peripheral Photoreception of the Cavefish P . Andruzzii and Provide Insight into the Time of Cave Colonization ». *Heredity* 117 (5): 383-92. https://doi.org/10.1038/hdy.2016.59.

Caro, Tim, Yvette Argueta, Emmanuelle Sophie Briolat, Joren Bruggink, Maurice Kasprowsky, Jai Lake, Matthew J. Mitchell, Sarah Richardson, et Martin How. 2019. « Benefits of Zebra Stripes: Behaviour of Tabanid Flies around Zebras and Horses ». *PLOS ONE* 14 (2): e0210831. https://doi.org/10.1371/journal.pone.0210831.

Casane, D., et S. Rétaux. 2016. « Evolutionary Genetics of the Cavefish Astyanax Mexicanus ». *Advances in Genetics* 95: 117-59. https://doi.org/10.1016/bs.adgen.2016.03.001.

Cavallari, Nicola, Elena Frigato, Daniela Vallone, Nadine Fröhlich, Jose Fernando Lopez-Olmeda, Augusto Foà, Roberto Berti, Francisco Javier Sánchez-Vázquez, Cristiano Bertolucci, et Nicholas S. Foulkes. 2011. « A Blind Circadian Clock in Cavefish Reveals That Opsins Mediate Peripheral Clock Photoreception ». *PLOS Biology* 9 (9): e1001142. https://doi.org/10.1371/journal.pbio.1001142.

Cavanagh, Rachel D., et Claudine Gibson. 2007. *Overview of the Conservation Status of Cartilaginous Fishes (Chondrichthyans) in the Mediterranean Sea*. IUCN. https://doi.org/10.2305/IUCN.CH.2007.MRA.3.en.

Ceinos, Rosa Maria, Elena Frigato, Cristina Pagano, Nadine Fröhlich, Pietro Negrini, Nicola Cavallari, Daniela Vallone, Silvia Fuselli, Cristiano Bertolucci, et Nicholas S. Foulkes. 2018. « Mutations in blind cavefish target the light-regulated circadian clock gene, period 2 ». *Scientific Reports* 8: 8754. https://doi.org/10.1038/s41598-018-27080-2.

Chalopin, Domitille, Magali Naville, Floriane Plard, Delphine Galiana, et Jean-Nicolas Volff. 2015. « Comparative Analysis of Transposable Elements Highlights Mobilome Diversity and Evolution in Vertebrates ». *Genome Biology and Evolution* 7 (2): 567-80. https://doi.org/10.1093/gbe/evv005.

Chang, B., N. L. Hawes, R. E. Hurd, M. T. Davisson, S. Nusinowitz, et J. R. Heckenlively. 2002. « Retinal Degeneration Mutants in the Mouse ». *Vision Research* 42 (4): 517-25. https://doi.org/10.1016/S0042-6989(01)00146-8.

Chang, Yu-Lin K., Eric Feunteun, Yasumasa Miyazawa, et Katsumi Tsukamoto. 2020. « New Clues on the Atlantic Eels Spawning Behavior and Area: The Mid-Atlantic Ridge Hypothesis ». *Scientific Reports* 10 (1): 15981. https://doi.org/10.1038/s41598-020-72916-5.

Chen, K., D. Durand, et M. Farach-Colton. 2000. « NOTUNG: A Program for Dating Gene Duplications and Optimizing Gene Family Trees ». Journal of Computational Biology: A Journal of Computational Molecular Cell Biology 7 (3-4): 429-47. https://doi.org/10.1089/106652700750050871.

Collin, Shaun, et Darryl Whitehead. 2004. « The functional roles of passive electroreception in non-electric fishes ». *Animal Biology* 54. https://doi.org/10.1163/157075604323010024.

Comte, Nicolas, Benoit Morel, Damir Hasić, Laurent Guéguen, Bastien Boussau, Vincent Daubin, Simon Penel, et al. 2020. « Treerecs: an integrated phylogenetic tool, from sequences to reconciliations ». Bioinformatics 36 (18): 4822-24. https://doi.org/10.1093/bioinformatics/btaa615.

Coombs, S, C B Braun, et B Donovan. 2001. « The Orienting Response of Lake Michigan Mottled Sculpin Is Mediated by Canal Neuromasts ». *The Journal of Experimental Biology* 204: 337-48.

Crampton, William G. R. 2019. « Electroreception, Electrogenesis and Electric Signal Evolution ». *Journal of Fish Biology* 95 (1): 92-134. https://doi.org/10.1111/jfb.13922.

Darwin, Charles. 1859. *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. London.

Davesne, Donald, Matt Friedman, Armin D. Schmitt, Vincent Fernandez, Giorgio Carnevale, Per E. Ahlberg, Sophie Sanchez, et Roger B. J. Benson. 2021. « Fossilized Cell Structures Identify an Ancient Origin for the Teleost Whole-Genome Duplication ». *Proceedings of the National Academy of Sciences* 118 (30). https://doi.org/10.1073/pnas.2101780118.

Davies, Wayne I.L., T. Katherine Tamai, Lei Zheng, Josephine K. Fu, Jason Rihel, Russell G. Foster, David Whitmore, et Mark W. Hankins. 2015. « An extended family of novel vertebrate photopigments is widely expressed and displays a diversity of function ». *Genome Research* 25 (11): 1666-79. https://doi.org/10.1101/gr.189886.115.

Davies, W.E., et I.M. Morgan. 1991. « Geology of Caves ». U.S. Department of the Interior / U.S. Geological Survey. https://pubs.usgs.gov/gip/7000072/report.pdf.

Devos, Lucie, François Agnès, Joanne Edouard, Victor Simon, Laurent Legendre, Naima El Khallouki, Sosthène Barbachou, Frédéric Sohm, et Sylvie Rétaux. 2021. « Eye Morphogenesis in the Blind Mexican Cavefish ». https://doi.org/10.1101/698035.

Durif, Caroline M. F., Howard I. Browman, John B. Phillips, Anne Berit Skiftesvik, L. Asbjørn Vøllestad, et Hans H. Stockhausen. 2013. « Magnetic Compass Orientation in the European Eel ». *PLOS ONE* 8 (3): e59212. https://doi.org/10.1371/journal.pone.0059212.

Dymek, J., M. Kuciel, et K. Żuwała. 2021. « Structural Diversity of Olfactory Organs in Osteoglossiformes ». *Journal of Zoology* 314 (1): 43-57. https://doi.org/10.1111/jzo.12854.

Dymek, Jakub, Gunilla Rosenqwist, Michał Kuciel, Eugenia Rita Lauriano, Gioele Capillo, Giacomo Zaccone, et Krystyna Żuwała. 2021. « Micro- and Macro-morphology of the Olfactory Organ of *Syngnathus Typhle* (Syngnathidae, Actinopterygii) ». *Acta Zoologica* 102 (2): 206-19. https://doi.org/10.1111/azo.12328.

Elipot, Yannick, Hélène Hinaux, Jacques Callebert, Jean-Marie Launay, Maryline Blin, et Sylvie Rétaux. 2014. « A Mutation in the Enzyme Monoamine Oxidase Explains Part of the Astyanax Cavefish Behavioural Syndrome ». *Nature Communications* 5: 3647. https://doi.org/10.1038/ncomms4647.

Emerling, Christopher A., et Mark S. Springer. 2014. « Eyes Underground: Regression of Visual Protein Networks in Subterranean Mammals ». *Molecular Phylogenetics and Evolution* 78: 260-70. https://doi.org/10.1016/j.ympev.2014.05.016.

Eschmeyer, William N., Ronald Fricke, et Richard Van Der Laan. 2021. « Eschmeyer's Catalog of Fishes: Genera, Species, References. » Eschmeyer's Catalog of Fishes. 2021. http://biotaxa.org/Zootaxa/article/view/zootaxa.3882.1.1.

Espinasa, Luis, Claudia Patricia Ornelas-García, Laurent Legendre, Sylvie Rétaux, Alexandra Best, Ramses Gamboa-Miranda, Hector Espinosa-Pérez, et Peter Sprouse. 2020. « Discovery of Two New Astyanax Cavefish Localities Leads to Further Understanding of the Species Biogeography ». *Diversity* 12 (10): 368. https://doi.org/10.3390/d12100368.

Eyun, Seong-il, Hideaki Moriyama, Federico G. Hoffmann, et Etsuko N. Moriyama. 2016. « Molecular Evolution and Functional Divergence of Trace Amine–Associated Receptors ». *PLOS ONE* 11 (3): e0151023. https://doi.org/10.1371/journal.pone.0151023.

Fan, Yanhui, Wenjuan Wang, Guoji Ma, Lijing Liang, Qi Shi, et Shiheng Tao. 2007. « Patterns of Insertion and Deletion in Mammalian Genomes ». *Current Genomics* 8 (6): 370-78. https://doi.org/10.2174/138920207783406479.

Farber, Debora B., et Richard N. Lolley. 1974. « Cyclic Guanosine Monophosphate: Elevation in Degenerating Photoreceptor Cells of the C3H Mouse Retina ». *Science* 186 (4162): 449-51. https://doi.org/10.1126/science.186.4162.449.

Ferrando, Sara, Lorenzo Gallus, Laura Ghigliotti, Andrea Amaroli, Ghulam Abbas, et Marino Vacchi. 2017. « Clarification of the Terminology of the Olfactory Lamellae in Chondrichthyes ». *The Anatomical Record* 300 (11): 2039-45. https://doi.org/10.1002/ar.23632.

Froese, R. et D. Pauly, 2021. FishBase. World Wide Web electronic publication. www.fishbase.org, version (06/2021).

Fischer, Ruth M., Bruno M. Fontinha, Stephan Kirchmaier, Julia Steger, Susanne Bloch, Daigo Inoue, Satchidananda Panda, Simon Rumpel, et Kristin Tessmar-Raible. 2013. « Co-Expression of VAL- and TMT-Opsins Uncovers Ancient Photosensory Interneurons and Motorneurons in the Vertebrate Brain ». *PLoS Biology* 11 (6): e1001585. https://doi.org/10.1371/journal.pbio.1001585.

Fishelson, L., D. Golani, B. Galil, et M. Goren. 2010. « Comparison of the Nasal Olfactory Organs of Various Species of Lizardfishes (Teleostei: Aulopiformes: Synodontidae) with Additional Remarks on the Brain ». *International Journal of Zoology* 2010: e807913. https://doi.org/10.1155/2010/807913.

Formicki, Krzysztof, Agata Korzelecka-Orkisz, et Adam Tański. 2019. « Magnetoreception in Fish ». *Journal of Fish Biology* 95 (1): 73-91. https://doi.org/10.1111/jfb.13998.

Fortune, Eric S., Nicole Andanar, Manu Madhav, Ravikrishnan P. Jayakumar, Noah J. Cowan, Maria Elina Bichuette, et Daphne Soares. 2020. « Spooky Interaction at a Distance in Cave and Surface Dwelling Electric Fishes ». *Frontiers in Integrative Neuroscience* 14: 51. https://doi.org/10.3389/fnint.2020.561524.

Foulkes, N.S., D. Whitmore, D. Vallone, et C. Bertolucci. 2016. « Studying the Evolution of the Vertebrate Circadian Clock ». In *Advances in Genetics*, 95:1-30. Elsevier. https://doi.org/10.1016/bs.adgen.2016.05.002.

Frisch, K. v. 1938. « Zur Psychologie des Fisch-Schwarmes ». *Naturwissenschaften* 26 (37): 601-6. https://doi.org/10.1007/BF01590598.

Fritzsch, Bernd, et Karen L. Elliott. 2017. « Evolution and Development of the Inner Ear Efferent System: Transforming a Motor Neuron Population to Connect to the Most Unusual Motor Protein via Ancient Nicotinic Receptors ». *Frontiers in Cellular Neuroscience* 11. https://doi.org/10.3389/fncel.2017.00114.

Fu, Limin, Beifang Niu, Zhengwei Zhu, Sitao Wu, et Weizhong Li. 2012. « CD-HIT: accelerated for clustering the next-generation sequencing data ». *Bioinformatics* 28 (23): 3150-52. https://doi.org/10.1093/bioinformatics/bts565.

Fumey, Julien, Hélène Hinaux, Céline Noirot, Claude Thermes, Sylvie Rétaux, et Didier Casane. 2018. « Evidence for late Pleistocene origin of Astyanax mexicanus cavefish ». *BMC Evolutionary Biology* 18 (1): 43. https://doi.org/10.1186/s12862-018-1156-7.

Gal, A., U. Orth, W. Baehr, E. Schwinger, et T. Rosenberg. 1994. « Heterozygous Missense Mutation in the Rod CGMP Phosphodiesterase Beta-Subunit Gene in Autosomal Dominant Stationary Night Blindness ». *Nature Genetics* 7 (1): 64-68. https://doi.org/10.1038/ng0594-64.

Graening, Gary, Danté Fenolio, Matthew Niemiller, Arthur Brown, et Jonathan Beard. 2010. « The 30-year recovery effort for the Ozark cavefish (Amblyopsis rosae): Analysis of current distribution, population trends, and conservation status of this threatened species ». *Environmental Biology of Fishes* 87: 55-88. https://doi.org/10.1007/s10641-009-9568-2.

Graham, Laurie A., et Peter L. Davies. 2021. « Horizontal Gene Transfer in Vertebrates: A Fishy Tale ». *Trends in Genetics* 37 (6): 501-3. https://doi.org/10.1016/j.tig.2021.02.006.

Gregory, T. Ryan. 2021. « Animal Genome Size Database ». 2021. http://www.genomesize.com/.

Grimsditch, Gabriel, Ahmed Basheer, et D. E. P. Bryant. 2017. « Extreme White Colouration of Frogfish Antennarius Maculatus Due to Coral Bleaching Event ». *Coral Reefs* 36 (1): 167-167. https://doi.org/10.1007/s00338-016-1500-6.

Gross, Joshua B., Richard Borowsky, et Clifford J. Tabin. 2009. « A Novel Role for Mc1r in the Parallel Evolution of Depigmentation in Independent Populations of the Cavefish Astyanax Mexicanus ». *PLOS Genetics* 5 (1): e1000326. https://doi.org/10.1371/journal.pgen.1000326.

Gross, Joshua B., Allison Furterer, Brian M. Carlson, et Bethany A. Stahl. 2013. « An Integrated Transcriptome-Wide Analysis of Cave and Surface Dwelling Astyanax mexicanus ». *PLoS ONE* 8 (2): e55659. https://doi.org/10.1371/journal.pone.0055659.

Hansen, Anne, et Barbara S. Zielinski. 2005. « Diversity in the Olfactory Epithelium of Bony Fishes: Development, Lamellar Arrangement, Sensory Neuron Cell Types and Transduction Components ». *Journal of Neurocytology* 34 (3): 183-208. https://doi.org/10.1007/s11068-005-8353-1.

Herman, Adam, Yaniv Brandvain, James Weagley, William R. Jeffery, Alex C. Keene, Thomas J. Y. Kono, Helena Bilandžija, et al. 2018. « The role of gene flow in rapid and repeated evolution of cave related traits in Mexican tetra, Astyanax mexicanus ». *Molecular ecology* 27 (22): 4397-4416. https://doi.org/10.1111/mec.14877.

Hervant, F., J. Mathieu, et J. Durand. 2001. « Behavioural, Physiological and Metabolic Responses to Long-Term Starvation and Refeeding in a Blind Cave-Dwelling (Proteus Anguinus) and a Surface-Dwelling (Euproctus Asper) Salamander ». *The Journal of Experimental Biology* 204 (Pt 2): 269-81.

Hinaux, Hélène, Maryline Blin, Julien Fumey, Laurent Legendre, Aurélie Heuzé, Didier Casane, et Sylvie Rétaux. 2015. « Lens Defects in Astyanax Mexicanus Cavefish: Evolution of Crystallins and a Role for AlphaA-Crystallin ». *Developmental Neurobiology* 75 (5): 505-21. https://doi.org/10.1002/dneu.22239.

Hinaux, Hélène, Lucie Devos, Maryline Blin, Yannick Elipot, Jonathan Bibliowicz, Alexandre Alié, et Sylvie Rétaux. 2016. « Sensory evolution in blind cavefish is driven by early embryonic events during gastrulation and neurulation ». *Development* 143 (23): 4521-32. https://doi.org/10.1242/dev.141291.

Hughes, Graham M., Emma S. M. Boston, John A. Finarelli, William J. Murphy, Desmond G. Higgins, et Emma C. Teeling. 2018. « The Birth and Death of Olfactory Receptor Gene Families in Mammalian Niche Adaptation ». *Molecular Biology and Evolution* 35 (6): 1390-1406. https://doi.org/10.1093/molbev/msy028.

Hughes, Lily C., Guillermo Ortí, Yu Huang, Ying Sun, Carole C. Baldwin, Andrew W. Thompson, Dahiana Arcila, et al. 2018. « Comprehensive Phylogeny of Ray-Finned Fishes (Actinopterygii) Based on Transcriptomic and Genomic Data ». *Proceedings of the National Academy of Sciences* 115 (24): 6249-54. https://doi.org/10.1073/pnas.1719358115.

Hussain, Ashiq, Luis R. Saraiva, David M. Ferrero, Gaurav Ahuja, Venkatesh S. Krishna, Stephen D. Liberles, et Sigrun I. Korsching. 2013. « High-Affinity Olfactory Receptor for the Death-Associated Odor Cadaverine ». *Proceedings of the National Academy of Sciences* 110 (48): 19579-84. https://doi.org/10.1073/pnas.1318596110.

Hyacinthe, Carole, Joël Attia, et Sylvie Rétaux. 2019. « Evolution of Acoustic Communication in Blind Cavefish ». *Nature Communications* 10 (1). https://doi.org/10.1038/s41467-019-12078-9.

Irion, Uwe, et Christiane Nüsslein-Volhard. 2019. « The identification of genes involved in the evolution of color patterns in fish ». *Current Opinion in Genetics & Development* 57 (août): 31-38. https://doi.org/10.1016/j.gde.2019.07.002.

Ishikawa, Sohta A, Anna Zhukova, Wataru Iwasaki, et Olivier Gascuel. 2019. « A Fast Likelihood Method to Reconstruct and Visualize Ancestral Scenarios ». *Molecular Biology and Evolution* 36 (9): 2069-85. https://doi.org/10.1093/molbev/msz131.

IUCN. 2016. « Ailuropoda Melanoleuca Giant Panda ». IUCN.

IUCN. 2021. « The IUCN Red List of Threatened Species ». IUCN Red List of Threatened Species. 2021. https://www.iucnredlist.org/en.

Janvier, Philippe. 2010. « microRNAs revive old views about jawless vertebrate divergence and evolution ». *Proceedings of the National Academy of Sciences of the United States of America* 107 (45): 19137-38. https://doi.org/10.1073/pnas.1014583107.

Jeffery, William R. 2020. « Astyanax surface and cave fish morphs ». *EvoDevo* 11 (1): 14. https://doi.org/10.1186/s13227-020-00159-6.

Johnsen, Sönke. 2001. « Hidden in Plain Sight: The Ecology and Physiology of Organismal Transparency ». *The Biological Bulletin* 201 (3): 301-18. https://doi.org/10.2307/1543609.

Juan, Carlos, Michelle T. Guzik, Damià Jaume, et Steven J. B. Cooper. 2010. « Evolution in Caves: Darwin's 'Wrecks of Ancient Life' in the Molecular Era ». *Molecular Ecology* 19 (18): 3865-80. https://doi.org/10.1111/j.1365-294X.2010.04759.x.

Kalyaanamoorthy, Subha, Bui Quang Minh, Thomas K. F. Wong, Arndt von Haeseler, et Lars S. Jermiin. 2017. « ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates ». *Nature Methods* 14 (6): 587-89. https://doi.org/10.1038/nmeth.4285.

Kaskan, Peter M., Edna Cristina S. Franco, Elizabeth S. Yamada, Luiz Carlos de Lima Silveira, Richard B. Darlington, et Barbara L. Finlay. 2005. « Peripheral variability and central constancy in mammalian visual system evolution ». *Proceedings of the Royal Society B: Biological Sciences* 272 (1558): 91-100. https://doi.org/10.1098/rspb.2004.2925.

Kasumyan, Alexander O. 2019. « The Taste System in Fishes and the Effects of Environmental Variables ». *Journal of Fish Biology* 95 (1): 155-78. https://doi.org/10.1111/jfb.13940.

Katoh, Kazutaka, et Daron M. Standley. 2013. « MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability ». *Molecular Biology and Evolution* 30 (4): 772-80. https://doi.org/10.1093/molbev/mst010.

Keeling, Patrick J., et Jeffrey D. Palmer. 2008. « Horizontal Gene Transfer in Eukaryotic Evolution ». *Nature Reviews Genetics* 9 (8): 605-18. https://doi.org/10.1038/nrg2386.

Kim, Hyun Tae, Seung Woon Yun, et Jong Young Park. 2019. « Anatomy, ultrastructure and histology of the olfactory organ of the largemouth bass Micropterus salmoides, Centrarchidae ». *Applied Microscopy* 49 (1): 18. https://doi.org/10.1186/s42649-019-0023-3.

Kingston, Jennifer, Gil Rosenthal, et Michael Ryan. 2003. « The role of sexual selection in maintaining a colour polymorphism in the pygmy swordtail, Xiphophorus pygmaeus ». *Animal Behaviour* 65: 735-43. https://doi.org/10.1006/anbe.2003.2110.

Kishida, Takushi, JGM Thewissen, Takashi Hayakawa, Hiroo Imai, et Kiyokazu Agata. 2015. « Aquatic adaptation and the evolution of smell and taste in whales ». *Zoological Letters* 1 (1): 9. https://doi.org/10.1186/s40851-014-0002-z.

Klaassen, Hannah, Yongfu Wang, Kay Adamski, Nicolas Rohner, et Johanna E. Kowalko. 2018. « CRISPR Mutagenesis Confirms the Role of Oca2 in Melanin Pigmentation in Astyanax Mexicanus ». *Developmental Biology* 441 (2): 313-18. https://doi.org/10.1016/j.ydbio.2018.03.014.

Korsching, Sigrun. 2008. « The Molecular Evolution of Teleost Olfactory Receptor Gene Families ». In *Chemosensory Systems in Mammals, Fishes, and Insects*, édité par Sigrun Korsching et Wolfgang Meyerhof, 47:221-38. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/400_2008_11.

Korsching, Sigrun I., Silke Argo, Harald Campenhausen, Rainer W. Friedrich, Andreas Rummrich, et Franco Weth. 1997. « Olfaction in Zebrafish: What Does a Tiny Teleost Tell Us? » *Seminars in Cell & Developmental Biology* 8 (2): 181-87. https://doi.org/10.1006/scdb.1996.0136.

Koskiniemi, Sanna, Song Sun, Otto G. Berg, et Dan I. Andersson. 2012. « Selection-Driven Gene Loss in Bacteria ». *PLOS Genetics* 8 (6): e1002787. https://doi.org/10.1371/journal.pgen.1002787.

Kottler, Verena A., et Manfred Schartl. 2018. « The Colorful Sex Chromosomes of Teleost Fish ». *Genes* 9 (5): 233. https://doi.org/10.3390/genes9050233.

Krauss, Jana, Silke Geiger-Rudolph, Iris Koch, Christiane Nüsslein-Volhard, et Uwe Irion. 2014. « A Dominant Mutation in Tyrp1A Leads to Melanophore Death in Zebrafish ». *Pigment Cell & Melanoma Research* 27 (5): 827-30. https://doi.org/10.1111/pcmr.12272.

Kumar, Sudhir, Glen Stecher, Michael Suleski, et S. Blair Hedges. 2017. « TimeTree: A Resource for Timelines, Timetrees, and Divergence Times ». Molecular Biology and Evolution 34 (7): 1812-19. https://doi.org/10.1093/molbev/msx116.

Kumar, Vinod. 2017. *Biological Timekeeping: Clocks, Rhythms and Behaviour*. Springer.

Ladich, F. 2000. « Acoustic communication and the evolution of hearing in fishes. » *Philosophical Transactions of the Royal Society B: Biological Sciences* 355 (1401): 1285-88.

Laerm, Joshua, Brian R. Chapman, et Ford W. Mark. 2007. « Star-Nosed Mole, Condylura Cristata (Linnaeus, 1758) ». *The Land Manager's Guide to Mammals of the South*, 116.

Lamb, Trevor D., Edward N. Pugh, et Shaun P. Collin. 2008. « The Origin of the Vertebrate Eye ». *Evolution: Education and Outreach* 1 (4): 415-26. https://doi.org/10.1007/s12052-008-0091-2.

Lanfear, Robert, John J. Welch, et Lindell Bromham. 2010. « Watching the Clock: Studying Variation in Rates of Molecular Evolution between Species ». *Trends in Ecology & Evolution* 25 (9): 495-503. https://doi.org/10.1016/j.tree.2010.06.007.

Larhammar, Dan, Karin Nordström, et Tomas A. Larsson. 2009. « Evolution of vertebrate rod and cone phototransduction genes ». *Philosophical Transactions of the Royal Society B: Biological Sciences* 364 (1531): 2867-80. https://doi.org/10.1098/rstb.2009.0077.

Lastein, Stine, El Hassan Hamdani, et Kjell B. Døving. 2006. « Gender Distinction in Neural Discrimination of Sex Pheromones in the Olfactory Bulb of Crucian Carp, Carassius carassius ». *Chemical Senses* 31 (1): 69-77. https://doi.org/10.1093/chemse/bjj007.

Lecchini, D, et Y Nakamura. 2013. « Use of Chemical Cues by Coral Reef Animal Larvae for Habitat Selection ». *Aquatic Biology* 19 (3): 231-38. https://doi.org/10.3354/ab00532.

Li, Diyan, et Jianzhi Zhang. 2014. « Diet Shapes the Evolution of the Vertebrate Bitter Taste Receptor Gene Repertoire ». *Molecular Biology and Evolution* 31 (2): 303-9. https://doi.org/10.1093/molbev/mst219.

Li, Heng. 2011. « A Statistical Framework for SNP Calling, Mutation Discovery, Association Mapping and Population Genetical Parameter Estimation from Sequencing Data ». *Bioinformatics (Oxford, England)* 27 (21): 2987-93. https://doi.org/10.1093/bioinformatics/btr509.

Li, Heng, et Richard Durbin. 2009. « Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform ». *Bioinformatics (Oxford, England)* 25 (14): 1754-60. https://doi.org/10.1093/bioinformatics/btp324.

Li, Ruiqiang, Wei Fan, Geng Tian, Hongmei Zhu, Lin He, Jing Cai, Quanfei Huang, et al. 2010. « The Sequence and de Novo Assembly of the Giant Panda Genome ». *Nature* 463 (7279): 311-17. https://doi.org/10.1038/nature08696.

Li, Wen-Hsiung, Takashi Gojobori, et Masatoshi Nei. 1981. « Pseudogenes as a Paradigm of Neutral Evolution ». *Nature* 292 (5820): 237-39. https://doi.org/10.1038/292237a0.

Li, Wen-Hsiung, et Masatoshi Nei. 1977. « Persistence of Common Alleles in Two Related Populations or Species ». *Genetics* 86 (4): 901-14.

Li, Ying, Guang Li, Haifang Wang, Jiulin Du, et Jun Yan. 2013. « Analysis of a Gene Regulatory Cascade Mediating Circadian Rhythm in Zebrafish ». *PLOS Computational Biology* 9 (2): e1002940. https://doi.org/10.1371/journal.pcbi.1002940.

Liu, Ake, Funan He, Libing Shen, Ruixiang Liu, Zhijun Wang, et Jingqi Zhou. 2019. « Convergent degeneration of olfactory receptor gene repertoires in marine mammals ». *BMC Genomics* 20 (1): 977. https://doi.org/10.1186/s12864-019-6290-0.

Liu, Zhijin, Huamei Wen, Frank Hailer, Fang Dong, Zuomin Yang, Tao Liu, Ling Han, Fanglei Shi, Yibo Hu, et Jiang Zhou. 2019. « Pseudogenization of Mc1r Gene Associated with Transcriptional Changes Related to Melanogenesis Explains Leucistic Phenotypes in Oreonectes Cavefish (Cypriniformes, Nemacheilidae) ». *Journal of Zoological Systematics and Evolutionary Research* 57 (4): 900-909. https://doi.org/10.1111/jzs.12286.

Lorin, Thibault, Frédéric G. Brunet, Vincent Laudet, et Jean-Nicolas Volff. 2018. « Teleost Fish-Specific Preferential Retention of Pigmentation Gene-Containing Families After Whole Genome Duplications in Vertebrates ». *G3 (Bethesda, Md.)* 8 (5): 1795-1806. https://doi.org/10.1534/g3.118.200201.

Lynch, M., et John S Conery. 2000. « The Evolutionary Fate and Consequences of Duplicate Genes ». *Science* 290 (5494): 1151-55. https://doi.org/10.1126/science.290.5494.1151.

Lynch, Michael, et Avinash Kewalramani. 2003. « Messenger RNA Surveillance and the Evolutionary Proliferation of Introns ». *Molecular Biology and Evolution* 20 (4): 563-71. https://doi.org/10.1093/molbev/msg068.

Mackay, Donna S., Olivera B. Boskovska, Harry L. S. Knopf, Kirsten J. Lampi, et Alan Shiels. 2002. « A Nonsense Mutation in CRYBB1 Associated with Autosomal Dominant Cataract Linked to Human Chromosome 22q ». *American Journal of Human Genetics* 71 (5): 1216-21. https://doi.org/10.1086/344212.

Macqueen, Daniel J., et Ian A. Johnston. 2014. « A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification ». *Proceedings of the Royal Society B: Biological Sciences* 281 (1778): 20132881. https://doi.org/10.1098/rspb.2013.2881.

Malinsky, Milan, Hannes Svardal, Alexandra M. Tyers, Eric A. Miska, Martin J. Genner, George F. Turner, et Richard Durbin. 2018. « Whole-Genome Sequences of Malawi Cichlids Reveal Multiple Radiations Interconnected by Gene Flow ». *Nature Ecology & Evolution* 2 (12): 1940-55. https://doi.org/10.1038/s41559-018-0717-x.

Mammola, Stefano. 2019. « Finding Answers in the Dark: Caves as Models in Ecology Fifty Years after Poulson and White ». *Ecography* 42 (7): 1331-51. https://doi.org/10.1111/ecog.03905.

Mao, Ting-Ru, Ye-Wei Liu, Madhava Meegaskumbura, Jian Yang, Gajaba Ellepola, Gayani Senevirathne, Cheng-Hai Fu, Joshua B. Gross, et Marcio R. Pie. 2021. « Evolution in Sinocyclocheilus cavefish is marked by rate shifts, reversals, and origin of novel traits ». *BMC Ecology and Evolution* 21 (1): 45. https://doi.org/10.1186/s12862-021-01776-y.

Marshall, Justin. 2017. « Vision and Lack of Vision in the Ocean ». *Current Biology* 27 (11): R494-502. https://doi.org/10.1016/j.cub.2017.03.012.

Marshall, N. Justin, Fabio Cortesi, Fanny de Busserolles, Uli E. Siebeck, et Karen L. Cheney. 2019. « Colours and Colour Vision in Reef Fishes: Past, Present and Future Research Directions ». *Journal of Fish Biology* 95 (1): 5-38. https://doi.org/10.1111/jfb.13849.

Matsui, Atsushi, Yasuhiro Go, et Yoshihito Niimura. 2010. « Degeneration of Olfactory Receptor Gene Repertories in Primates: No Direct Link to Full Trichromatic Vision ». *Molecular Biology and Evolution* 27 (5): 1192-1200. https://doi.org/10.1093/molbev/msq003.

Maximino, Caio, Rhayra X. do Carmo Silva, Kimberly dos Santos Campos, Jeisiane S. de Oliveira, Sueslene P. Rocha, Maryana P. Pyterson, Dainara P. dos Santos Souza, et al. 2019. « Sensory Ecology of Ostariophysan Alarm Substances ». *Journal of Fish Biology* 95 (1): 274-86. https://doi.org/10.1111/jfb.13844.

Mayrose, Itay, Shing H. Zhan, Carl J. Rothfels, Karen Magnuson-Ford, Michael S. Barker, Loren H. Rieseberg, et Sarah P. Otto. 2011. « Recently Formed Polyploid Plants Diversify at Lower Rates ». *Science* 333 (6047): 1257-1257. https://doi.org/10.1126/science.1207205.

McGaugh, Suzanne E., Joshua B. Gross, Bronwen Aken, Maryline Blin, Richard Borowsky, Domitille Chalopin, Hélène Hinaux, et al. 2014. « The Cavefish Genome Reveals Candidate Genes for Eye Loss ». *Nature Communications* 5 (1): 10. https://doi.org/10.1038/ncomms6307.

McLaughlin, M. E., M. A. Sandberg, E. L. Berson, et T. P. Dryja. 1993. « Recessive Mutations in the Gene Encoding the Beta-Subunit of Rod Phosphodiesterase in Patients with Retinitis Pigmentosa ». *Nature Genetics* 4 (2): 130-34. https://doi.org/10.1038/ng0693-130.

Melin, Amanda D., Donald W. Kline, Chihiro Hiramatsu, et Tim Caro. 2016. « Zebra Stripes through the Eyes of Their Predators, Zebras, and Humans ». *PLOS ONE* 11 (1): e0145679. https://doi.org/10.1371/journal.pone.0145679.

Meng, Fanwei, Ingo Braasch, Jennifer B. Phillips, Xiwen Lin, Tom Titus, Chunguang Zhang, et John H. Postlethwait. 2013. « Evolution of the Eye Transcriptome under Constant Darkness in Sinocyclocheilus Cavefish ». *Molecular Biology and Evolution* 30 (7): 1527-43. https://doi.org/10.1093/molbev/mst079.

Meng, Fanwei, Yahui Zhao, John H. Postlethwait, et Chunguang Zhang. 2013. « Differentially-expressed opsin genes identified in Sinocyclocheilus cavefish endemic to China ». *Current zoology* 59 (2): 170-74.

Meredith, R. W., G. Zhang, M. T. P. Gilbert, E. D. Jarvis, et M. S. Springer. 2014. « Evidence for a Single Loss of Mineralized Teeth in the Common Avian Ancestor ». *Science* 346 (6215): 1254390-1254390. https://doi.org/10.1126/science.1254390.

Meredith, Robert W., John Gatesy, William J. Murphy, Oliver A. Ryder, et Mark S. Springer. 2009. « Molecular Decay of the Tooth Gene Enamelin (ENAM) Mirrors the Loss of Enamel in the Fossil Record of Placental Mammals ». *PLOS Genetics* 5 (9): e1000634. https://doi.org/10.1371/journal.pgen.1000634.

Miles, Alistair, Zamin Iqbal, Paul Vauterin, Richard Pearson, Susana Campino, Michel Theron, Kelda Gould, et al. 2016. « Indels, structural variation, and recombination drive genomic diversity in Plasmodium falciparum ». *Genome research* 26 (août). https://doi.org/10.1101/gr.203711.115.

Mitchell, R.W., W.H. Russell, et William Elliott. 1977. « Mexican eyeless characin fishes, genus Astyanax: Environment, distribution, and evolution ». *Spec. Publ. Mus. Texas Tech. Univ.* 12: 1-89.

Modrell, Melinda S., William E. Bemis, R. Glenn Northcutt, Marcus C. Davis, et Clare V. H. Baker. 2011. « Electrosensory Ampullary Organs Are Derived from Lateral Line Placodes in Bony Fishes ». *Nature Communications* 2 (1): 496. https://doi.org/10.1038/ncomms1502.

Moran, Damian, Rowan Softley, et Eric J. Warrant. 2015. « The Energetic Cost of Vision and the Evolution of Eyeless Mexican Cavefish ». *Science Advances* 1 (8): e1500363. https://doi.org/10.1126/sciadv.1500363.

Morimura, Hiroyuki, Florence Saindelle-Ribeaudeau, Eliot L. Berson, et Thaddeus P. Dryja. 1999. « Mutations in RGR , Encoding a Light-Sensitive Opsin Homologue, in Patients with Retinitis Pigmentosa ». *Nature Genetics* 23 (4): 393-94. https://doi.org/10.1038/70496.

Naisbett-Jones, Lewis C., Nathan F. Putman, Michelle M. Scanlan, David L. G. Noakes, et Kenneth J. Lohmann. 2020. « Magnetoreception in Fishes: The Effect of Magnetic Pulses on Orientation of Juvenile Pacific Salmon ». *The Journal of Experimental Biology*. https://doi.org/10.1242/jeb.222091.

Nature. 1948. « Water Movements and Earth Currents : Electrical and Magnetic Effects* ». *Nature* 161 (4084): 192-93. https://doi.org/10.1038/161192a0.

Near, Thomas J., Ron I. Eytan, Alex Dornburg, Kristen L. Kuhn, Jon A. Moore, Matthew P. Davis, Peter C. Wainwright, Matt Friedman, et W. Leo Smith. 2012. « Resolution of Ray-Finned Fish Phylogeny and Timing of Diversification ». *Proceedings of the National Academy of Sciences* 109 (34): 13698-703. https://doi.org/10.1073/pnas.1206625109.

Nei, Masatoshi, Xun Gu, et Tatyana Sitnikova. 1997. « Evolution by the Birth-and-Death Process in Multigene Families of the Vertebrate Immune System ». *Proceedings of the National Academy of Sciences* 94 (15): 7799-7806. https://doi.org/10.1073/pnas.94.15.7799.

Nei, Masatoshi, et Alejandro P. Rooney. 2005. « Concerted and Birth-and-Death Evolution of Multigene Families ». *Annual review of genetics* 39: 121-52. https://doi.org/10.1146/annurev.genet.39.073003.112240.

Nguyen, Dinh Truong, Kyooyeol Lee, Hojun Choi, Min-kyeung Choi, Minh Thong Le, Ning Song, Jin-Hoi Kim, et al. 2012. « The complete swine olfactory subgenome: expansion of the olfactory gene repertoire in the pig genome ». *BMC Genomics* 13 (1): 584. https://doi.org/10.1186/1471-2164-13-584.

Nguyen, Lam-Tung, Heiko A. Schmidt, Arndt von Haeseler, et Bui Quang Minh. 2015. « IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies ». *Molecular Biology and Evolution* 32 (1): 268-74. https://doi.org/10.1093/molbev/msu300.

Niemiller, Matthew L., Benjamin M. Fitzpatrick, Premal Shah, Lars Schmitz, et Thomas J. Near. 2013. « Evidence for Repeated Loss of Selective Constraint in Rhodopsin of Amblyopsid Cavefishes (teleostei: Amblyopsidae) ». *Evolution* 67 (3): 732-48. https://doi.org/10.1111/j.1558-5646.2012.01822.x.

Niemiller, Matthew L., Thomas J. Near, et Benjamin M. Fitzpatrick. 2012. « Delimiting Species Using Multilocus Data: Diagnosing Cryptic Diversity in the Southern Cavefish, Typhlichthys Subterraneus (ᴛᴇʟeostei: Amblyopsidae) ». *Evolution* 66 (3): 846-66. https://doi.org/10.1111/j.1558-5646.2011.01480.x.

Niimura, Yoshihito. 2009. « On the Origin and Evolution of Vertebrate Olfactory Receptor Genes: Comparative Genome Analysis Among 23 Chordate Species ». *Genome Biology and Evolution* 1: 34-44. https://doi.org/10.1093/gbe/evp003.

Niimura, Yoshihito, Atsushi Matsui, et Kazushige Touhara. 2014. « Extreme Expansion of the Olfactory Receptor Gene Repertoire in African Elephants and Evolutionary Dynamics of Orthologous Gene Groups in 13 Placental Mammals ». *Genome Research* 24 (9): 1485-96. https://doi.org/10.1101/gr.169532.113.

Niimura, Yoshihito, Atsushi Matsui, et Kazushige Touhara. 2018. « Acceleration of Olfactory Receptor Gene Loss in Primate Evolution: Possible Link to Anatomical Change in Sensory Systems and Dietary Transition ». *Molecular Biology and Evolution* 35 (6): 1437-50. https://doi.org/10.1093/molbev/msy042.

Niimura, Yoshihito, et Masatoshi Nei. 2005. « Evolutionary dynamics of olfactory receptor genes in fishes and tetrapods ». *Proceedings of the National Academy of Sciences* 102 (17): 6039-44.

Nikaido, Masato. 2019. « Evolution of *V1R* Pheromone Receptor Genes in Vertebrates: Diversity and Commonality ». *Genes & Genetic Systems* 94 (4): 141-49. https://doi.org/10.1266/ggs.19-00009.

Niven, Jeremy E., et Simon B. Laughlin. 2008. « Energy Limitation as a Selective Pressure on the Evolution of Sensory Systems ». *The Journal of Experimental Biology* 211 (Pt 11): 1792-1804. https://doi.org/10.1242/jeb.017574.

Orme, David, R.P. Freckleton, G.H. Thomas, T. Petzoldt, S.A. Fritz, et Nick Isaac. 2013. « CAPER: comparative analyses of phylogenetics and evolution in R ». Methods in Ecology and Evolution 3: 145-51.

Pejaver, Vikas, Jorge Urresti, Jose Lugo-Martinez, Kymberleigh A. Pagel, Guan Ning Lin, Hyun-Jun Nam, Matthew Mort, et al. 2020. « Inferring the Molecular and Phenotypic Impact of Amino Acid Variants with MutPred2 ». *Nature Communications* 11 (1): 5918. https://doi.org/10.1038/s41467-020-19669-x.

Penel, Simon, Anne-Muriel Arigon, Jean-François Dufayard, Anne-Sophie Sertier, Vincent Daubin, Laurent Duret, Manolo Gouy, et Guy Perrière. 2009. « Databases of Homologous Gene Families for Comparative Genomics ». BMC Bioinformatics 10 Suppl 6 : S3. https://doi.org/10.1186/1471-2105-10-S6-S3.

Peters, Robert, Tim Wessel, Bert Wollenberg, et Annelies Olijslagers. 2002. « The bioelectric field of the catfish Ictalurus nebulosus ». *Journal of physiology, Paris* 96: 397-404. https://doi.org/10.1016/S0928-4257(03)00018-4.

Pettibone, Marian H. 1985. « Polychaete Worms From A Cave In The Bahamas And From Experimental Wood Panels In Deep water Of The North atlantic (Polynoidae, Macellicephalinae, Harmothoinae) ». *Proceedings of the Biological Society of Washington*, 1985, 98 édition.

Pfeiffer, Wolfgang. 1968. « Das geruchsorgan der Polypteridae (Pisces, Brachiopterygii) ». *Zeitschrift für Morphologie der Tiere* 63 (1): 75-110. https://doi.org/10.1007/BF00343428.

Piotrowski, Tatjana, et Clare V. H. Baker. 2014. « The Development of Lateral Line Placodes: Taking a Broader View ». *Developmental Biology*, Placodes, 389 (1): 68-81. https://doi.org/10.1016/j.ydbio.2014.02.016.

Popper, Arthur N., Dennis T.T. Plachta, David A. Mann, et Dennis Higgs. 2004. « Response of clupeid fish to ultrasound: a review ». *ICES Journal of Marine Science* 61 (7): 1057-61. https://doi.org/10.1016/j.icesjms.2004.06.005.

Poulson, Thomas L. 1963. « Cave Adaptation in Amblyopsid Fishes ». *The American Midland Naturalist* 70 (2): 257-90. https://doi.org/10.2307/2423056.

Protas, Meredith E, Candace Hersey, Dawn Kochanek, Yi Zhou, Horst Wilkens, William R Jeffery, Leonard I Zon, Richard Borowsky, et Clifford J Tabin. 2006. « Genetic Analysis of Cavefish Reveals Molecular Convergence in the Evolution of Albinism ». *Nature Genetics* 38 (1): 107-11. https://doi.org/10.1038/ng1700.

Proudlove, Graham S. 2021. « Subterranean Fishes of the World ». 2021. https://cavefishes.org.uk.

Putland, Rosalyn L., John C. Montgomery, et Craig A. Radford. 2019. « Ecology of Fish Hearing ». *Journal of Fish Biology* 95 (1): 39-52. https://doi.org/10.1111/jfb.13867.

Rabosky, Daniel L., Jonathan Chang, Pascal O. Title, Peter F. Cowman, Lauren Sallan, Matt Friedman, Kristin Kaschner, et al. 2018. « An Inverse Latitudinal Gradient in Speciation Rate for Marine Fishes ». *Nature* 559 (7714): 392-95. https://doi.org/10.1038/s41586-018-0273-1.

Ramachandran, V. S., C. W. Tyler, R. L. Gregory, D. Rogers-Ramachandran, S. Duensing, C. Pillsbury, et C. Ramachandran. 1996. « Rapid Adaptive Camouflage in Tropical Flounders ». *Nature* 379 (6568): 815-18. https://doi.org/10.1038/379815a0.

Ranwez, Vincent, Sébastien Harispe, Frédéric Delsuc, et Emmanuel J. P. Douzery. 2011. « MACSE: Multiple Alignment of Coding SEquences Accounting for Frameshifts and Stop Codons ». *PLOS ONE* 6 (9): e22594. https://doi.org/10.1371/journal.pone.0022594.

Recknagel, Hans, Kathryn R. Elmer, et Axel Meyer. 2013. « A Hybrid Genetic Linkage Map of Two Ecologically and Morphologically Divergent Midas Cichlid Fishes (Amphilophus Spp.) Obtained by Massively Parallel DNA Sequencing (DdRADSeq) ». *G3: Genes, Genomes, Genetics* 3 (1): 65-74. https://doi.org/10.1534/g3.112.003897.

Rétaux, Sylvie, et Didier Casane. 2013. « Evolution of eye development in the darkness of caves: adaptation, drift, or both? » *EvoDevo* 4 (1): 26. https://doi.org/10.1186/2041-9139-4-26.

Rice, P., I. Longden, et A. Bleasby. 2000. « EMBOSS: The European Molecular Biology Open Software Suite ». *Trends in Genetics: TIG* 16 (6): 276-77. https://doi.org/10.1016/s0168-9525(00)02024-2.

Riddle, Misty R., Ariel C. Aspiras, Karin Gaudenz, Robert Peuß, Jenny Y. Sung, Brian Martineau, Megan Peavey, et al. 2018. « Insulin resistance in cavefish as an adaptation to a nutrient-limited environment ». *Nature* 555 (7698): 647-51. https://doi.org/10.1038/nature26136.

Roach, Jared C., Gustavo Glusman, Arian F.A. Smit, Chad D. Huff, Robert Hubley, Paul T. Shannon, Lee Rowen, et al. 2010. « Analysis of Genetic Inheritance in a Family Quartet by Whole Genome Sequencing ». *Science (New York, N.Y.)* 328 (5978): 636-39. https://doi.org/10.1126/science.1186802.

Ronco, Fabrizia, Michael Matschiner, Astrid Böhne, Anna Boila, Heinz H. Büscher, Athimed El Taher, Adrian Indermaur, et al. 2021. « Drivers and Dynamics of a Massive Adaptive Radiation in Cichlid Fishes ». *Nature* 589 (7840): 76-81. https://doi.org/10.1038/s41586-020-2930-4.

Roura, E., M.W. Baldwin, et K.C. Klasing. 2013. « The Avian Taste System: Potential Implications in Poultry Nutrition ». *Animal Feed Science and Technology* 180 (1-4): 1-9. https://doi.org/10.1016/j.anifeedsci.2012.11.001.

Rowland, W. J., C. L. Baube, et T. T. Horan. 1991. « Signalling of Sexual Receptivity by Pigmentation Pattern in Female Sticklebacks ». *Animal Behaviour* 42 (2): 243-49. https://doi.org/10.1016/S0003-3472(05)80555-X.

Ruzicka, Leyla, Douglas G. Howe, Sridhar Ramachandran, Sabrina Toro, Ceri E. Van Slyke, Yvonne M. Bradford, Anne Eagle, et al. 2019. « The Zebrafish Information Network: New Support for Non-Coding Genes, Richer Gene Ontology Annotations and the Alliance of Genome Resources ». *Nucleic Acids Research* 47 (D1): D867-73. https://doi.org/10.1093/nar/gky1090.

Salzberg, Steven L. 2019. « Next-generation genome annotation: we still struggle to get it right ». *Genome Biology* 20 (1): 92. https://doi.org/10.1186/s13059-019-1715-2.

Santini, Francesco, Luke J Harmon, Giorgio Carnevale, et Michael E Alfaro. 2009. « Did Genome Duplication Drive the Origin of Teleosts? A Comparative Study of Diversification in Ray-Finned Fishes ». *BMC Evolutionary Biology* 9 (1): 194. https://doi.org/10.1186/1471-2148-9-194.

Sato, Motoya, Mitsuru Nakazawa, Tomoaki Usui, Naoyuki Tanimoto, Haruki Abe, et Hiroshi Ohguro. 2005. « Mutations in the Gene Coding for Guanylate Cyclase-Activating Protein 2 (GUCA1B Gene) in Patients with Autosomal Dominant Retinal Dystrophies ». *Graefe's Archive for Clinical and Experimental Ophthalmology* 243 (3): 235-42. https://doi.org/10.1007/s00417-004-1015-7.

Sato, Yuki, Nobuhiko Miyasaka, et Yoshihiro Yoshihara. 2007. « Hierarchical Regulation of Odorant Receptor Gene Choice and Subsequent Axonal Projection of Olfactory Sensory Neurons in Zebrafish ». *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 27 (7): 1606-15. https://doi.org/10.1523/JNEUROSCI.4218-06.2007.

Sazima, Ivan, Lucélia Nobre Carvalho, Fernando Pereira Mendonça, et Jansen Zuanon. 2006. « Fallen leaves on the water-bed: diurnal camouflage of three night active fish species in an Amazonian streamlet ». *Neotropical Ichthyology* 4 (1): 119-22. https://doi.org/10.1590/S1679-62252006000100013.

Schulze, Lisanne, Jörg Henninger, Mykola Kadobianskyi, Thomas Chaigne, Ana Isabel Faustino, Nahid Hakiy, Shahad Albadri, et al. 2018. « Transparent Danionella Translucida as a Genetically Tractable Vertebrate Brain Model ». *Nature Methods* 15 (11): 977-83. https://doi.org/10.1038/s41592-018-0144-6.

Schwab, I R. 2018. « The evolution of eyes: major steps. The Keeler lecture 2017: centenary of Keeler Ltd ». *Eye* 32 (2): 302-13. https://doi.org/10.1038/eye.2017.226.

Shi, Peng, et Jianzhi Zhang. 2006. « Contrasting Modes of Evolution Between Vertebrate Sweet/Umami Receptor Genes and Bitter Receptor Genes ». *Molecular Biology and Evolution* 23 (2): 292-300. https://doi.org/10.1093/molbev/msj028.

Shi, Peng, et Jianzhi Zhang. 2007. « Comparative genomic analysis identifies an evolutionary shift of vomeronasal receptor gene repertoires in the vertebrate transition from water to land ». *Genome Research* 17 (2): 166-74. https://doi.org/10.1101/gr.6040007.

Shu, D.-G., S. Conway Morris, J. Han, Z.-F. Zhang, K. Yasui, P. Janvier, L. Chen, et al. 2003. « Head and Backbone of the Early Cambrian Vertebrate Haikouichthys ». *Nature* 421 (6922): 526-29. https://doi.org/10.1038/nature01264.

Simon, Noah, Suguru Fujita, Megan Porter, et Masato Yoshizawa. 2019. « Expression of Extraocular Opsin Genes and Light-Dependent Basal Activity of Blind Cavefish ». *PeerJ* 7: e8148. https://doi.org/10.7717/peerj.8148.

Simon, Victor, Romain Elleboode, Kélig Mahé, Laurent Legendre, Patricia Ornelas-Garcia, Luis Espinasa, et Sylvie Rétaux. 2017. « Comparing growth in surface and cave morphs of the species Astyanax mexicanus: insights from scales ». *EvoDevo* 8 (1): 23. https://doi.org/10.1186/s13227-017-0086-6.

Simon, Victor, Carole Hyacinthe, et Sylvie Rétaux. 2019. « Breeding behavior in the blind Mexican cavefish and its river-dwelling conspecific ». *PLoS ONE* 14 (2): e0212591. https://doi.org/10.1371/journal.pone.0212591.

Slater, Guy St C., et Ewan Birney. 2005. « Automated generation of heuristics for biological sequence comparison ». *BMC Bioinformatics* 6 (1): 31. https://doi.org/10.1186/1471-2105-6-31.

Spady, Tyrone C., Juliet W. L. Parry, Phyllis R. Robinson, David M. Hunt, James K. Bowmaker, et Karen L. Carleton. 2006. « Evolution of the Cichlid Visual Palette through Ontogenetic Subfunctionalization of the Opsin Gene Arrays ». *Molecular Biology and Evolution* 23 (8): 1538-47. https://doi.org/10.1093/molbev/msl014.

Stoletzki, Nina, et Adam Eyre-Walker. 2011. « The Positive Correlation between dN/dS and dS in Mammals Is Due to Runs of Adjacent Substitutions ». Molecular Biology and Evolution 28 (4): 1371-80. https://doi.org/10.1093/molbev/msq320.

Swamynathan, Shivalingappa K., Mary A. Crawford, W. Gerald Robison, Jyotshnabala Kanungo, et Joram Piatigorsky. 2003. « Adaptive differences in the structure and macromolecular compositions of the air and water corneas of the “four-eyed” fish (Anableps anableps) ». *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* 17 (14): 1996-2005. https://doi.org/10.1096/fj.03-0122com.

Takahashi, Joseph S., Hee-Kyung Hong, Caroline H. Ko, et Erin L. McDearmon. 2008. « The Genetics of Mammalian Circadian Order and Disorder: Implications for Physiology and Disease ». *Nature Reviews Genetics* 9 (10): 764-75. https://doi.org/10.1038/nrg2430.

Tanay, Amos, et Eric D. Siggia. 2008. « Sequence context affects the rate of short insertions and deletions in flies and primates ». *Genome Biology* 9 (2): R37. https://doi.org/10.1186/gb-2008-9-2-r37.

The UniProt Consortium. 2021. « UniProt: the universal protein knowledgebase in 2021 ». *Nucleic Acids Research* 49 (D1): D480-89. https://doi.org/10.1093/nar/gkaa1100.

Theisen, Birgit. 1970. « The morphology and vascularization of the olfactory organ in calamoichthys calabaricus, (pisces, polypteridae). » *Videnskabelige Meddelelser Dansk Naturhistorisk Forening*, *133*, 31-50.

Theodosius Dobzhansky; Nothing in Biology Makes Sense except in the Light of Evolution. The American Biology Teacher 1 March 1973; 35 (3): 125–129. doi: https://doi.org/10.2307/4444260

Tobler, Michael, Seth W. Coleman, Brian D. Perkins, et Gil G. Rosenthal. 2010. « Reduced opsin gene expression in a cave-dwelling fish ». *Biology Letters* 6 (1): 98-101. https://doi.org/10.1098/rsbl.2009.0549.

Trajano, Eleonora. 2000. « Cave Faunas in the Atlantic Tropical Rain Forest: Composition, Ecology, and Conservation1 ». *Biotropica* 32 (4b): 882-93. https://doi.org/10.1111/j.1744-7429.2000.tb00626.x.

Tsagkogeorga, Georgia, Steven Müller, Christophe Dessimoz, et Stephen J. Rossiter. 2017. « Comparative Genomics Reveals Contraction in Olfactory Receptor Genes in Bats ». *Scientific Reports* 7 (1): 259. https://doi.org/10.1038/s41598-017-00132-9.

Ueda, Hiroshi. 2021. « Chemoreception in Fishes ». In *Oxford Research Encyclopedia of Neuroscience*, par Hiroshi Ueda. Oxford University Press. https://doi.org/10.1093/acrefore/9780190264086.013.333.

Varatharasan, Nirupa, Roger P. Croll, et Tamara Franz-Odendaal. 2009. « Taste Bud Development and Patterning in Sighted and Blind Morphs of Astyanax Mexicanus ». *Developmental Dynamics* 238 (12): 3056-64. https://doi.org/10.1002/dvdy.22144.

Vatine, Gad, Daniela Vallone, Gothilf Yoav, et Nicholas S Foulkes. 2011. « It's Time to Swim! Zebrafish and the Circadian Clock ». *FEBS Letters* 585 (10). https://doi.org/10.1016/j.febslet.2011.04.007.

Vieira, Filipe G, Alejandro Sánchez-Gracia, et Julio Rozas. 2007. « Comparative genomic analysis of the odorant-binding protein family in 12 Drosophila genomes: purifying selection and birth-and-death evolution ». *Genome Biology* 8 (11): R235. https://doi.org/10.1186/gb-2007-8-11-r235.

Villacorta, Carlos, Damià Jaume, Pedro Oromí, et Carlos Juan. 2008. « Under the volcano: phylogeography and evolution of the cave-dwelling Palmorchestia hypogaea(Amphipoda, Crustacea) at La Palma (Canary Islands) ». *BMC Biology* 6 (1): 7. https://doi.org/10.1186/1741-7007-6-7.

Volff, J.-N. 2005. « Genome Evolution and Biodiversity in Teleost Fish ». *Heredity* 94 (3): 280-94. https://doi.org/10.1038/sj.hdy.6800635.

Wackermannova, Marie, Ludvik Pinc, et L Jebavý. 2016. « Olfactory sensitivity in mammalian species: a review ». *Physiological research / Academia Scientiarum Bohemoslovaca* 65.

Wakisaka, Noriko, Nobuhiko Miyasaka, Tetsuya Koide, Miwa Masuda, Towako Hiraki-Kajiyama, et Yoshihiro Yoshihara. 2017. « An Adenosine Receptor for Olfaction in Fish ». *Current Biology* 27 (10): 1437-1447.e4. https://doi.org/10.1016/j.cub.2017.04.014.

Waterhouse, Robert M., Mathieu Seppey, Felipe A. Simão, Mosè Manni, Panagiotis Ioannidis, Guennadi Klioutchnikov, Evgenia V. Kriventseva, et Evgeny M. Zdobnov. 2018. « BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics ». *Molecular Biology and Evolution* 35 (3): 543-48. https://doi.org/10.1093/molbev/msx319.

Webb, Paul. 2020. *Introduction to Oceanography*. https://rwu.pressbooks.pub/webboceanography/.

Wilkens, H., Ulrike Strecker, et Jill Yager. 1989. « Eye Reduction and Phylogenetic Age in Ophidiiform Cave Fish1 ». *Journal of Zoological Systematics and Evolutionary Research* 27 (2): 126-34. https://doi.org/10.1111/j.1439-0469.1989.tb00337.x.

Won, Yong-Jin, Arjun Sivasundar, Yong Wang, et Jody Hey. 2005. « On the origin of Lake Malawi cichlid species: A population genetic analysis of divergence ». *Proceedings of the National Academy of Sciences* 102 (suppl 1): 6581-86.

WWF. 2001. « The Status of Wild Atlantic Salmon : A River by River Assessment ». WWF.

Wynne, Jut. 2011. « Ecological Inventory of Lava Tube Caves, El Malpais National Monument, New Mexico ». Northern Arizona University. https://in.nau.edu/wp-content/uploads/sites/128/2018/08/NAU-252-ek.pdf.

Xu, Peng, Jian Xu, Guangjian Liu, Lin Chen, Zhixiong Zhou, Wenzhu Peng, Yanliang Jiang, et al. 2019. « The Allotetraploid Origin and Asymmetrical Genome Evolution of the Common Carp Cyprinus Carpio ». *Nature Communications* 10 (1): 4625. https://doi.org/10.1038/s41467-019-12644-1.

Yancey, Paul H., Mackenzie E. Gerringer, Jeffrey C. Drazen, Ashley A. Rowden, et Alan Jamieson. 2014. « Marine Fish May Be Biochemically Constrained from Inhabiting the Deepest

Ocean Depths ». *Proceedings of the National Academy of Sciences* 111 (12): 4461-65. https://doi.org/10.1073/pnas.1322003111.

Yang, Junxing, Xiaoli Chen, Jie Bai, Dongming Fang, Ying Qiu, Wansheng Jiang, Hui Yuan, et al. 2016. « The Sinocyclocheilus cavefish genome provides insights into cave adaptation ». *BMC Biology* 14 (1): 1. https://doi.org/10.1186/s12915-015-0223-4.

Yang, Z. 2007. « PAML 4: Phylogenetic Analysis by Maximum Likelihood ». *Molecular Biology and Evolution* 24 (8): 1586-91. https://doi.org/10.1093/molbev/msm088.

Yokoyama, Shozo. 2000. « Molecular Evolution of Vertebrate Visual Pigments ». *Progress in Retinal and Eye Research* 19 (4): 385-419. https://doi.org/10.1016/S1350-9462(00)00002-1.

Yoshihara, Yoshihiro. 2014. « Zebrafish Olfactory System ». In *The Olfactory System*, édité par Kensaku Mori, 71-96. Tokyo: Springer Japan. https://doi.org/10.1007/978-4-431-54376-3_5.

Yoshizawa, Masato, Špela Gorčiki, Daphne Soares, et William R. Jeffery. 2010. « Evolution of a behavioral shift mediated by superficial neuromasts helps cavefish find food in darkness ». *Current biology : CB* 20 (18): 1631-36. https://doi.org/10.1016/j.cub.2010.07.017.

Zapilko, Veronika, et Sigrun I. Korsching. 2016. « Tetrapod V1R-like Ora Genes in an Early-Diverging Ray-Finned Fish Species: The Canonical Six Ora Gene Repertoire of Teleost Fish Resulted from Gene Loss in a Larger Ancestral Repertoire ». *BMC Genomics* 17: 83. https://doi.org/10.1186/s12864-016-2399-6.

Zeng, Zishuo, et Yana Bromberg. 2019. « Predicting Functional Effects of Synonymous Variants: A Systematic Review and Perspectives ». *Frontiers in Genetics* 10. https://doi.org/10.3389/fgene.2019.00914.

Zhao, Haiyu, Giuseppe Di Mauro, Sebastian Lungu-Mitea, Pietro Negrini, Andrea Maria Guarino, Elena Frigato, Thomas Braunbeck, et al. 2018. « Modulation of DNA Repair Systems in Blind Cavefish during Evolution in Constant Darkness ». *Current Biology: CB* 28 (20): 3229-3243.e4. https://doi.org/10.1016/j.cub.2018.08.039.

Zhao, Huabin, Jian-Rong Yang, Huailiang Xu, et Jianzhi Zhang. 2010. « Pseudogenization of the Umami Taste Receptor Gene Tas1r1 in the Giant Panda Coincided with its Dietary Switch to Bamboo ». *Molecular Biology and Evolution* 27 (12): 2669-73. https://doi.org/10.1093/molbev/msq153.

Zhuang, Xuan, Chun Yang, Katherine R. Murphy, et C.-H. Christina Cheng. 2019. « Molecular Mechanism and History of Non-Sense to Sense Evolution of Antifreeze Glycoprotein Gene in Northern Gadids ». *Proceedings of the National Academy of Sciences* 116 (10): 4400-4405. https://doi.org/10.1073/pnas.1817138116.

Zuckerkandl, Emile, et Linus Pauling. 1965. « Evolutionary Divergence and Convergence in Proteins ». In *Evolving Genes and Proteins*, 97-166. Elsevier. https://doi.org/10.1016/B978-1-4832-2734-4.50017-6.

A- Articles publiés en premier auteur

1) **Contrasting Gene Decay in Subterranean Vertebrates: Insights from Cavefishes and Fossorial Mammals**

Contrasting Gene Decay in Subterranean Vertebrates: Insights from Cavefishes and Fossorial Mammals

Maxime Policarpo,[1] Julien Fumey,[†,1] Philippe Lafargeas,[1] Delphine Naquin,[2] Claude Thermes,[2] Magali Naville,[3] Corentin Dechaud,[3] Jean-Nicolas Volff,[3] Cedric Cabau,[4] Christophe Klopp,[5] Peter Rask Møller,[6] Louis Bernatchez,[7] Erik García-Machado,[7,8] Sylvie Rétaux,[*,9] and Didier Casane[*,1,10]

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

[2]Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), 91198, Gif-sur-Yvette, France

[3]Institut de Génomique Fonctionnelle de Lyon, Université de Lyon, CNRS UMR 5242, Ecole Normale Supérieure de Lyon, Université Claude Bernard Lyon 1, Lyon, France

[4]SIGENAE, GenPhySE, INRAE, ENVT, Université de Toulouse, Castanet Tolosan, France

[5]INRAE, SIGENAE, Genotoul Bioinfo, MIAT UR875, Castanet Tolosan, France

[6]Natural History Museum of Denmark, University of Copenhagen, Copenhagen Ø, Denmark

[7]Department of Biology, Institut de Biologie Intégrative et des Systèmes, Université Laval, Québec City, QC, Canada

[8]Centro de Investigaciones Marinas, Universidad de La Habana, La Habana, Cuba

[9]CNRS, Institut des Neurosciences Paris-Saclay, Université Paris-Saclay, Gif-sur-Yvette, France

[10]UFR Sciences du Vivant, Université de Paris, Paris, France

[†]Present address: Human Genetics and Cognitive Functions, Institut Pasteur, CNRS UMR 3571, Université de Paris, Paris, France

***Corresponding authors**: E-mails: sylvie.retaux@inaf.cnrs-gif.fr; didier.casane@egce.cnrs-gif.fr.

Associate editor: Chang Belinda

Abstract

Evolution sometimes proceeds by loss, especially when structures and genes become dispensable after an environmental shift relaxes functional constraints. Subterranean vertebrates are outstanding models to analyze this process, and gene decay can serve as a readout. We sought to understand some general principles on the extent and tempo of the decay of genes involved in vision, circadian clock, and pigmentation in cavefishes. The analysis of the genomes of two Cuban species belonging to the genus *Lucifuga* provided evidence for the largest loss of eye-specific genes and nonvisual opsin genes reported so far in cavefishes. Comparisons with a recently evolved cave population of *Astyanax mexicanus* and three species belonging to the Chinese tetraploid genus *Sinocyclocheilus* revealed the combined effects of the level of eye regression, time, and genome ploidy on eye-specific gene pseudogenization. The limited extent of gene decay in all these cavefishes and the very small number of loss-of-function mutations per pseudogene suggest that their eye degeneration may not be very ancient, ranging from early to late Pleistocene. This is in sharp contrast with the identification of several vision genes carrying many loss-of-function mutations in ancient fossorial mammals, further suggesting that blind fishes cannot thrive more than a few million years in cave ecosystems.

Key words: **cavefishes, eye-specific genes, pseudogenization, machine learning, relaxed selection, molecular dating.**

Introduction

The evolution of organisms confronted with drastic environmental shifts results sometimes in profound phenotypic changes. Constructive evolution involved in adaptation to new environments, relying on novelties at the phenotypic and genetic levels, has attracted much interest (Rose and Lauder 1996). Nevertheless, it has become evident that regressive evolution, which is often nonadaptive and which occurs by loss of structures and functions and the corresponding genes, accounts for a nonnegligible component of the evolutionary process (Lahti et al. 2009; Albalat and Cañestro 2016). Selection may be involved in regressive evolution, as suggested by a QTL analysis of eye degeneration in a cavefish (Protas et al. 2007). More generally, the evolution of animals in a dark environment is a particularly useful model to analyze regressive evolution because it has occurred many times in many taxa and some challenges, such as the absence of light, are always present, which allows us to analyze convergent evolution. A dark environment is expected to release purifying selection on light-related genes such as those involved in vision, the circadian clock, and pigmentation.

© The Author(s) 2020. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Open Access

Downloaded from https://academic.oup.com/mbe/article/38/2/589/5918... on 29 June 2021

Article

A careful examination of gene decay on a genome-wide scale has been performed in obligate fossorial mammals. It has been shown that several independent lineages with degenerate eyes have lost many genes involved in visual perception (Kim et al. 2011; Emerling and Springer 2014; Fang, Nevo, et al. 2014; Fang, Seim, et al. 2014; Emerling 2018).

In order to better understand the modalities, tempo, extent, and limits of molecular decay of light-related genetic systems at the scale of subterranean vertebrates, it is informative to compare gene decay in fossorial mammals with another large group of subterranean vertebrates, the cavefishes. Cavefishes represent the largest and most diverse group of cave vertebrates (Culver and Pipan 2009), but gene decay has not been surveyed on a genome-wide scale in relevant species. On the one hand, in the reference genome of the Mexican cavefish *Astyanax mexicanus*, very few pseudogenes have been found among the light-related genes (McGaugh et al. 2014). The retention of almost all of the eye-specific genes in this blindfish is a paradox under the hypothesis that it evolved millions of years ago, but would be expected under the hypothesis that all *A. mexicanus* cave populations are very recent (Fumey et al. 2018). On the other hand, in the genomes of three fishes belonging to the Chinese genus *Sinocyclocheilus*, that is, *S. grahami*, a surface fish with large eyes, *S. anshuiensis*, a blind cavefish, and *S. rhinocerous*, a small-eyed cavefish, many loss-of-function (LoF) mutations were found (Yang et al. 2016), but their tetraploid genomes hampered the identification of those LoF mutations related to the shift from surface to cave. After a whole-genome duplication (WGD), the pairs of paralogs resulting from this process (ohnologs) are most often redundant and one ohnolog can be pseudogenized without reducing fitness. Accordingly, several eye-specific genes contain LoF mutations in the large-eyed *S. grahami*. However, no thorough analysis of differential gene losses in relation to the level of eye degeneration has been carried out (Yang et al. 2016).

As the very recent origin of *A. mexicanus* cavefish populations and the tetraploidy of *Sinocyclocheilus* species did not allow the analysis of gene decay in cavefishes, it was necessary to examine the genomes of cavefishes that are millions of years old and which have not undergone a recent WGD. Species belonging to the genus *Lucifuga* (cave brotulas from the Bahamas and Cuba) were identified as good candidates. The genus *Lucifuga* is divided into two clades, one comprising only blind species and the other only small-eyed species (García-Machado et al. 2011). As no close surface relative has been identified and large genetic distances have been found between some species, within and between these clades, this genus of cavefishes is likely relatively ancient and the last common ancestor of the extant species was probably a cave-adapted fish (García-Machado et al. 2011; Hernández et al. 2020).

In order to compare eye-specific gene decay among cavefishes and between cavefishes and fossorial mammals, here we first sequenced the genomes of two Cuban cavefishes: one specimen, belonging to *L. dentata*, was blind and depigmented, the other one, belonging to *L. gibarensis* (Hernández et al. 2020), had small eyes and was pigmented. For all cavefishes for which genomes were available, and for some closely related surface species, we looked for likely LoF mutations (i.e., premature STOP codons, losses of START and STOP codons, losses of intron splice sites, and small indels leading to frameshifts) and for signatures of relaxed purifying selection on nonsynonymous mutations. The comparison of gene decay among cavefishes was extended to nonvisual opsin genes and two large sets of genes involved in the circadian clock and pigmentation. Contrasting patterns of gene loss indicated that eye-specific genes and nonvisual opsin genes have been much less constrained than circadian clock and pigmentation genes. The level of eye-specific gene decay was related to several factors such as the time the fishes have spent in the subterranean environment, the level of eye degeneration, and the level of genome ploidy. Nevertheless, no eye-specific genes with many LoF mutations were found in any cavefish, in sharp contrast to the highly degenerated eye-specific genes found in some fossorial mammals, suggesting that eye degeneration in the cavefishes is much more recent.

Results

Vision, Circadian Clock, and Pigmentation Gene Sets

In the zebrafish, *Danio rerio*, we identified 63 eye-specific genes, that is, those expressed only in the eyes, and 32 genes coding for nonvisual opsins. As the same pattern of gene decay was observed for both gene sets, and because a large majority of these genes are involved in vision, they were pooled in a single group thereafter called vision genes (fig. 1 and supplementary fig. S1, Supplementary Material online). In addition, we defined two other gene sets, 42 genes involved in the circadian clock and 257 genes involved in pigmentation (fig. 1).

An annotated draft genome and a transcriptome were obtained for one *L. dentata* cavefish (eyeless, supplementary fig. S2, Supplementary Material online) and the reads of one *L. gibarensis* cavefish (small-eyed, supplementary fig. S2, Supplementary Material online) were mapped on the genome of *L. dentata*. A detailed description of the results is given in supplementary data S1, Supplementary Material online. This allowed us to identify genes belonging to the gene sets defined above in the two *Lucifuga* species using the program EXONERATE and the coding sequences of the zebrafish. Using the same method, homologs were retrieved from the published genomes of three surface fishes (*Brotula barbata*, *Carapus acus*, and *Lamprogrammus exutus*) belonging to the same order, Ophidiiformes; from the published genomes of an *A. mexicanus* cavefish from Pachón cave and two surface fishes (a conspecific *A. mexicanus* surface fish and the piranha *Pygocentrus nattereri*), all belonging to the order Characiformes; and from the published genomes of two *Sinocyclocheilus* cavefishes (*S. anshuiensis* and *S. rhinocerous*) and a surface fish (*S. grahami*), all belonging to the order Cypriniformes (fig. 2).

As some genes have been duplicated in the terminal lineage leading to zebrafish (used as the reference to establish the gene lists), only one copy was expected to be found in other fishes. On the other hand, gene duplications, gene deletions as

Downloaded from https://academic.oup.com/mbe/article/38/2/589/5912537 by guest on 29 June 2021

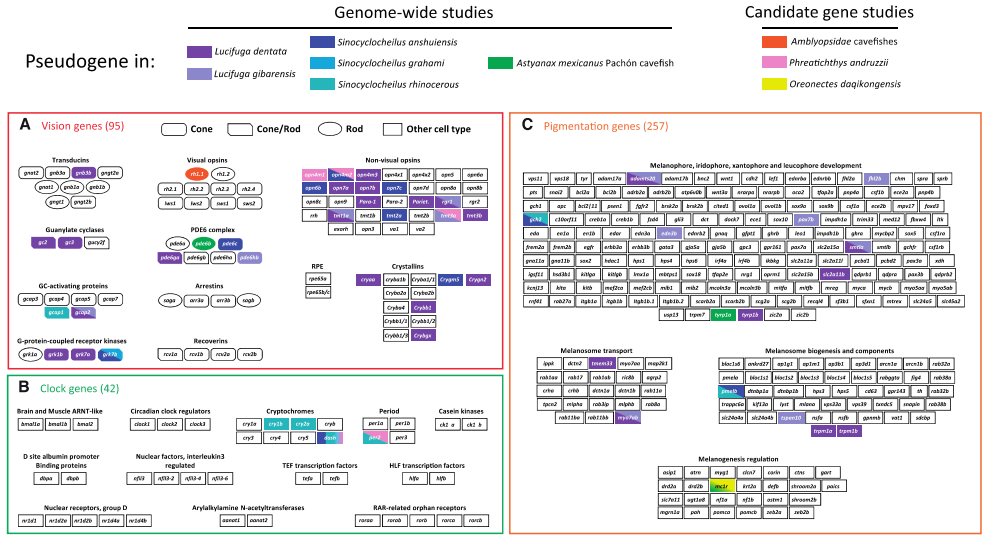


FIG. 1. Gene sets and pseudogenes. (*A*) Vision genes. (*B*) Circadian clock genes. (*C*) Pigmentation genes. The number of genes in zebrafish for each set in shown in round brackets. Each box represents a gene, except the box rpe65b/c which represents two genes present only in zebrafish. The gene *mc1r* is duplicated in *Astyanax mexicanus* and only one copy carries a LoF mutation in the *A. mexicanus* cavefish genome. Pseudogenes are colored according the species in which they were found. In candidate gene studies, only few genes were examined, whereas most genes were studied in genome-wide analyses.

well as WGDs occurred in other lineages. Therefore, the number of genes retrieved was highly variable among genomes (fig. 2).

Identification of LoF Mutations

Gene sequences were classified as complete if the whole-coding DNA sequence (CDS) could be retrieved or otherwise were classified as incomplete. Incomplete genes were discarded as it was most often impossible to know if they corresponded to sequencing gaps, assembly artifacts, or true large deletions. In the case of the *A. mexicanus* cavefish genome, among 45 missing exons in 19 incomplete genes, 36 exons of 16 genes could be amplified using primers designed with the exon sequences present in *A. mexicanus* surface fish genome. These data suggest that 80% of incomplete CDS in the *A. mexicanus* cavefish genome are not the result of large deletions but are assembly artifacts. Only complete sequences were further analyzed for the following LoF mutations: premature STOP codon, loss of the initiation codon, loss of the STOP codon, indel leading to a frameshift, mutations at intron splice sites. Other mutations in noncoding and coding sequences that could lead to a nonfunctional gene were not investigated as they cannot be readily identified. For example, several in-frame indel mutations were found in *A. mexicanus* but their functional consequences were not clear (Berning et al. 2019). The numbers of pseudogenes reported are therefore underestimates of the true numbers of nonfunctional genes.

Vision Pseudogenes

Among the list of 95 zebrafish vision genes, 76 genes were retrieved from the two *Lucifuga* spp. (cavefishes), 75 from *B. barbata*, 72 from *C. acus*, and 73 from *La. exutus* (surface fishes) (fig. 2 and supplementary fig. S1, Supplementary Material online). In eyed ophidiiform fishes, no vision pseudogene was found (*B. barbata* and *C. acus*) or only one (*gcap1* in *La. exutus*), but five pseudogenes were identified in the small-eyed cavefish *L. gibarensis* and 19 pseudogenes in the eyeless *L. dentata*. The nonvisual opsin *rgr1* was pseudogenized in the common ancestor of the two *Lucifuga* species, as the same mutation (at a splice site of intron 4) was found in both genomes (fig. 2 and supplementary fig. S1, Supplementary Material online). Examination of the read coverage of LoF mutations indicated that the specimen of *L. gibarensis* sequenced was heterozygous for LoF mutations found at two different sites in the *gcap2* gene (supplementary table S1, Supplementary Material online). In the transcriptome of *L. dentata*, transcripts corresponding to nine pseudogenes were found (three nonvisual opsins, three crystallins, and three genes involved in the phototransduction pathway), whereas no transcripts were found for 10 other pseudogenes (supplementary table S1, Supplementary Material online). In agreement with a recent WGD, two copies (ohnologs) of most vision genes were retrieved from the genomes of *Sinocyclocheilus* species (fig. 2 and supplementary fig. S1, Supplementary Material online). In the large-eyed *S. grahami*, ~10% of the retrieved vision genes were pseudogenized (18/173 genes carried at least one LoF mutation),

Downloaded from https://academic.oup.com/mbe/article/38/2/589/5912537 by guest on 29 June 2021

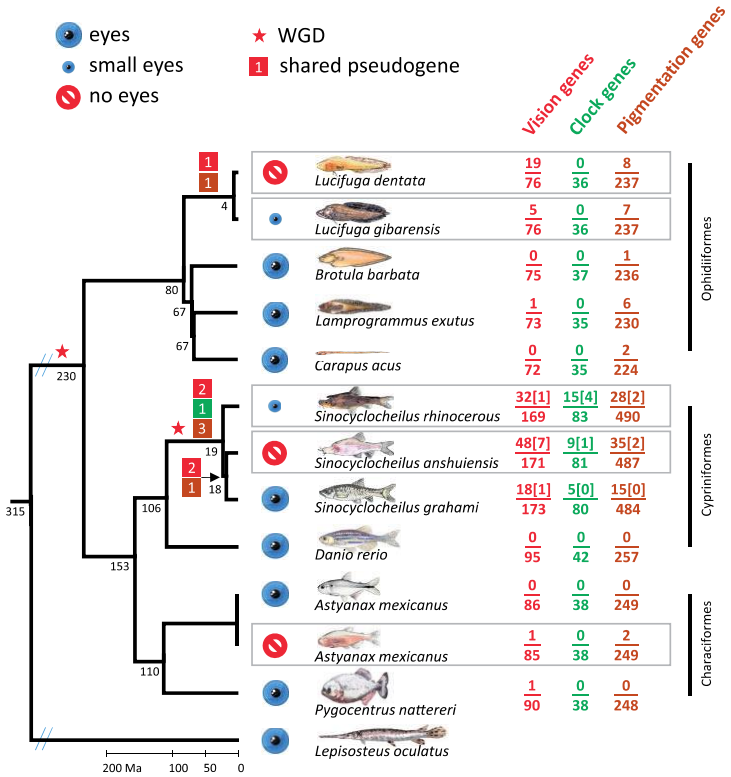


FIG. 2. Phylogeny and pseudogene mapping. For each gene set, the number of pseudogenes found and the number of genes examined in a species are given to the right of the species name. For tetraploid species, the number of pairs of pseudogenes that are ohnologs is shown in square brackets. The number of shared pseudogenes is in boxes colored according to the gene set they belong to. Divergence times were obtained from timetree (http://www.timetree.org/), except between *Lucifuga* species (present study). Red star, whole-genome duplication.

compared with 19% (32/169) in the small-eyed *S. rhinocerous* and 28% (48/171) in the eyeless cavefish *S. anshuiensis*. Only one pair of ohnologs was pseudogenized in the eyed *S. grahami* and the small-eyed *S. rhinocerous*, whereas seven pairs of ohnologs were pseudogenized in the blind *S. anshuiensis* (fig. 2 and supplementary fig. S1, Supplementary Material online). Two premature STOP codons and a frameshift in *sws1* were shared by the three *Sinocyclocheilus* species, as well as a mutation at the donor site of the third intron of *gc3*; *S. anshuiensis* and *S. grahami* shared a frameshift in *crygm5* and a frameshift plus a premature STOP codon in *grk7b* (fig. 2). Thus, most LoF mutations occurred in the terminal lineages leading to these species but a few LoF mutations occurred in different common ancestors.

In *A. mexicanus*, 86 and 85 vision genes were retrieved from the surface fish and the Pachón cavefish genome, respectively. Only one pseudogene was found in the Pachón cavefish genome, which is due to a deletion of 11 bp in *pde6b* (figs. 1 and 2).

In summary, although no or very few vision genes are pseudogenized in surface fishes and *A. mexicanus* cavefish, many vision pseudogenes were found in other cavefishes, up to 25% in *L. dentata*.

Circadian Clock Pseudogenes

Among the list of 42 zebrafish circadian clock genes, 36 genes were retrieved from *Lucifuga* genomes and 38 from *Astyanax* genomes. No pseudogene was found in these species. On the other hand, 5, 15, and 9 pseudogenes were identified among 80, 83, and 81 genes retrieved from the genomes of *S. grahami* (eyed), *S. rhinocerous* (small-eyed), and *S. anshuiensis* (blind), respectively. Both ohnologs of *cry-dash* were independently pseudogenized in *S. rhinocerous* and *S. anshuiensis*, a gene also pseudogenized in the Somalian cavefish *Phreatichthys andruzzii* (Zhao et al. 2018). Three other pairs of ohnologs (*cry1b*, *cry2a*, and *per2*) carried LoF mutations in *S. rhinocerous* (in *Ph. andruzzii*, the most abundant transcript of *per2* encodes a

Downloaded from https://academic.oup.com/mbe/article/38/2/589/5912537 by guest on 29 June 2021

truncated protein; Ceinos et al. 2018). These data suggest that the circadian clock has most likely been impaired in *S. rhinocerous* but less so in *S. anshuiensis* (figs. 1 and 2).

Pigmentation Genes

Among the list of 257 zebrafish pigmentation genes, 237 genes were retrieved from *Lucifuga* genomes, 8 being pseudogenized in *L. dentata* and 7 in *L. gibarensis*. Although *smtla* and *myo7ab* have been lost independently in the two lineages, a premature STOP codon and an insertion are shared in *adamts20*. The number of pseudogenes in these cavefishes does not greatly differ from those found in some surface relatives, as 6 pseudogenes were identified among 230 pigmentation genes in *La. exutus* (figs. 1 and 2). Among *Sinocyclocheilus* species, only 3% (15/484) of pseudogenes were found in *S. grahami*, whereas 6% (28/490) were found in *S. rhinocerous* and 7% (35/487) in *S. anshuiensis* (fig. 2). Thus, after the WGD, the retention of pigmentation genes seems to have been much higher than among the vision genes in the two cavefishes but also in the surface fish (compared with 10%, 19%, and 28% of vision pseudogenes, respectively). Such a high percentage of retention of pigmentation genes has been found also after the Salmonid-specific WGD (Lorin et al. 2018). Strikingly, although no pair of ohnologs was found pseudogenized in *S. grahami*, the same two pairs of ohnologs (*gch2* and *pmelb*) were independently pseudogenized in *S. anshuiensis* and *S. rhinocerous* (fig. 1). The very small number of pseudogenes and the independent pseudogenization of the same genes in these two species suggest that only a limited subset of genes involved in pigmentation can be lost in these cavefishes.

In *A. mexicanus* cavefish, 2 pseudogenes were found among 249 pigmentation genes: *mc1r* which has already been reported in the literature (Gross et al. 2009) and which is also pseudogenized in the Chinese cavefish *Oreonectes daqikongensis* (Liu et al. 2019), and *tyrp1a* (figs. 1 and 2). The gene *mc1r* is duplicated in *A. mexicanus*, one copy is pseudogenized and the other could be functional (Gross et al. 2018).

The reliability of the LoF mutations identified above was assessed by different approaches. First, we found no or very few LoF mutations in at least one gene set in published genomes used in the present study (figs. 1 and 2), indicating that if some LoF mutations are artifacts, they are very rare in these genomes. In *L. dentata* and *L. gibarensis* genomes, no LoF mutations were found in circadian clock genes. Second, in *L. dentata*, LoF mutations were also found in transcripts when available. Third, the sequencing depth was high at the positions where LoF mutations were found in *L. dentata* and *L. gibarensis* (supplementary table S1, Supplementary Material online). Altogether, these lines of evidence suggest that LoF mutations identified in cavefish genomes are reliable.

Some LoF mutations listed above have been found in genes for which gene knockout experiments in model species suggest that their pseudogenization could be involved in eye regression, circadian clock disruption, or depigmentation. Other LoF mutations have been found in genes for which pseudogenization is involved in evolution of these traits in other cavefishes. A summary of the putative effects of these LoF mutations can be found in supplementary data S2, Supplementary Material online.

Estimation of the Number of Neutral Vision Genes Based on the Distribution of LoF Mutations per Pseudogene in Cave Brotulas

Among vision pseudogenes, some accumulated more than one LoF mutation, but in most of the cases only one LoF mutation was found (supplementary fig. S1, Supplementary Material online). In order to test if the whole set, or only a subset, of vision genes could accumulate LoF mutations in cavefishes, we compared the distribution of the number of LoF mutations per pseudogene with those expected under these different hypotheses. Expected distributions were obtained using either a simple analytical model assuming that all genes have the same probability to mutate, or a more complex model that takes into account that different genes do not have the same probability to mutate because they have different length and do not contain the same number of introns. In the latter case, the computation of expected distributions was based on simulations (see Materials and Methods for a detailed description of both methods). Very similar expected distributions were obtained with both approaches (fig. 3). This analysis could be performed only with *Lucifuga* species, as only one LoF mutation was found in *A. mexicanus* cavefish vision genes and a WGD allowed LoF mutations in many vision genes in *Sinocyclocheilus* species, including *S. grahami* which is a surface fish with large functional eyes.

In the eyeless *L. dentata*, 22 LoF mutations were distributed among 19 vision pseudogenes. More precisely, among the 76 genes retrieved, there were 57 genes without LoF mutation, 16 with 1 mutation, and 3 with 2 mutations. This distribution was compared with expected distributions obtained for different numbers of neutral genes ranging from 19 to 76 (fig. 3*A*). A better fit between the observed and expected distribution was found when at least 60 genes were assumed to be neutral sequences in which LoF could be found (fig. 3*A*). Using the same approach, we compared the distribution of the number of LoF mutations per pseudogene in the small-eyed *L. gibarensis* (71 genes without LoF mutation, 3 with 1 mutation, 2 with 2 mutations) with expected distributions assuming a number of neutral genes ranging from 5 to 76 (fig. 3*B*). In this case, the best fit was obtained when ~15 vision genes were free to accumulate LoF mutations (fig. 3*B*). These results suggested that most genes, if not all, are dispensable in the blind *L. dentata* whereas only a small subset can be lost in the small-eyed *L. gibarensis*.

Evidence of Relaxed Selection on Nonsynonymous Mutations in Cavefish Vision Genes

To confirm these findings on LoF mutations, we searched for other signatures of relaxed selection using methods based on changes in ω (the ratio of the mean number of nonsynonymous substitutions per nonsynonymous site to the mean number of synonymous substitutions per synonymous site,

Downloaded from https://academic.oup.com/mbe/article/38/2/589/5912537 by guest on 29 June 2021

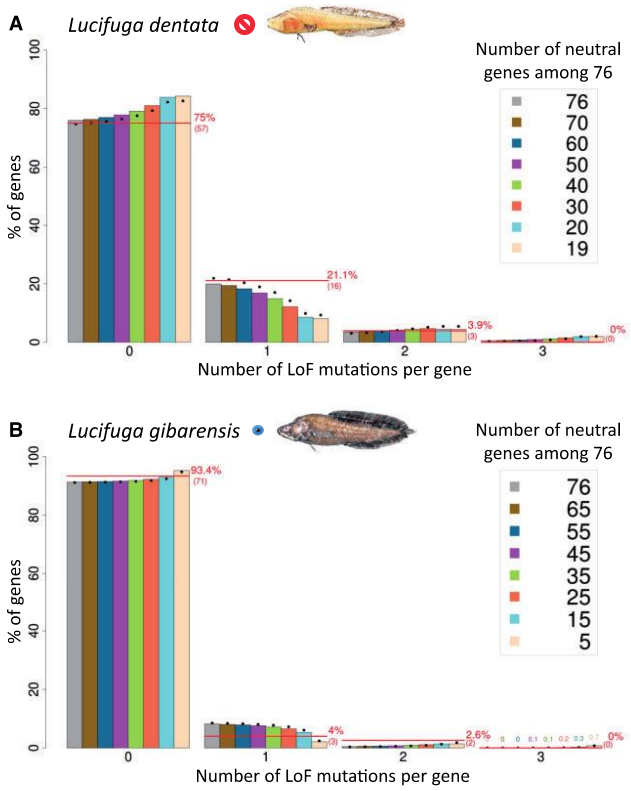


FIG. 3. Observed and expected distributions of LoF mutations per gene. (A) *Lucifuga dentata*. (B) *L. gibarensis*. Red lines, observed distribution. Expected distributions were obtained using an analytical model (dots) and 10,000 simulations (histograms). Expected distributions were obtained for different numbers of neutral genes among 76 genes: between 19 and 76 in *L. dentata* and between 5 and 76 in *L. gibarensis*.

also known as d*n*/ds and Ka/Ks). This ratio is expected to be lower than one under purifying selection, equal to one under neutral evolution, and larger than one under adaptive selection. As gene divergence was <0.9% between *L. dentata* and *L. gibarensis*, and <0.2% between the two *A. mexicanus* morphs, the number of nucleotide differences per gene was very low and often no sequence change was observed (supplementary fig. S3, Supplementary Material online). Therefore, ω was computed for three sets of concatenated gene sequences (vision, circadian clock, and pigmentation genes), to obtain more reliable estimates and to compare the shift of selective pressure on these gene sets in different fish lineages. In order to obtain a reliable distribution of ω in surface fishes, gene sequences were retrieved from other fish genomes: tetraodon (*Dichotomyctere nigroviridis*), cod (*Gadus morhua*), stickleback (*Gasterosteus aculeatus*), spotted gar (*Lepisosteus oculatus*), tilapia (*Oreochromis niloticus*), medaka (*Oryzias latipes*), platyfish (*Xiphophorus maculatus*). Taking into account the known fish phylogenetic relationships (supplementary fig. S4, Supplementary Material online) and using the PAML package (Yang 2007), we compared three nested branch models assuming: 1) only one ω; 2) one ω_{CF} for blind cavefishes (*A. mexicanus* and *L. dentata*) and one ω_{SF} for the other fishes (two-ratio model); and 3) one ω for each branch (free-ratio model). For each data set, likelihood ratio tests suggested that the two-ratio model was better than the one-ratio model ($\omega_{CF} > \omega_{SF}$) and the best model was the free-ratio model (supplementary table S2, Supplementary Material online). With the free-ratio model, *L. dentata* had the highest ω (0.409) for vision genes. For circadian clock genes, both *A. mexicanus* blind cavefish and *L. dentata* had the highest ω (0.29). For pigmentation genes, ω was similar in cave and surface fishes (fig. 4 and supplementary fig. S4, Supplementary Material online).

Independently, we analyzed the same sets of genes in *Sinocyclocheilus* species. For each species, ohnologs were concatenated into two series of gene sequences. For the two-ratio model, we assumed one ω_{CF} for the blind cavefish, *S. anshuiensis*, and one ω_{SF} for the two-eyed species, *S. rhinocerous* and *S. grahami*. For each data set, likelihood

Downloaded from https://academic.oup.com/mbe/article/38/2/589/5912537 by guest on 29 June 2021

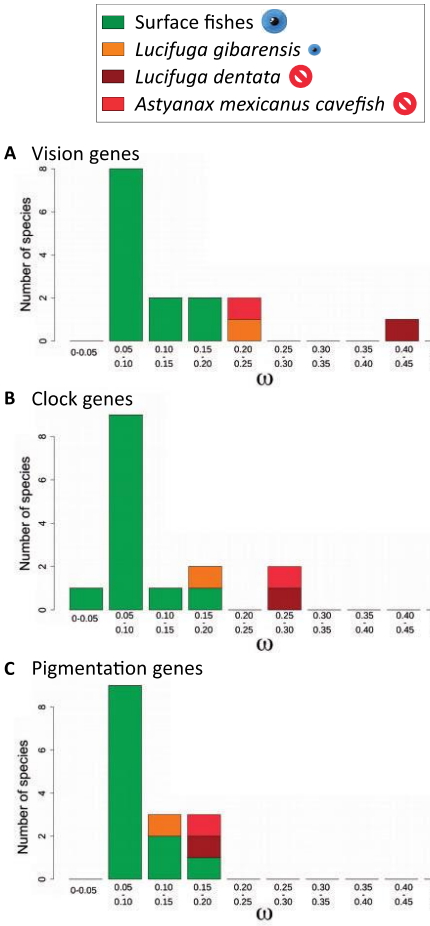


FIG. 4. Distribution of ω in surface and cave fishes. (*A*) Vision genes. (*B*) Circadian clock genes. (*C*) Pigmentation genes. Surface fishes: *Astyanax mexicanus* surface morph, *Brotula barbata*, *Carapus acus*, *Danio rerio*, *Dichotomyctere nigroviridis*, *Gadus morhua*, *Gasterosteus aculeatus*, *Lamprogrammus exutus*, *Oreochromis niloticus*, *Oryzias latipes*, *Xiphophorus maculatus*, and *Pygocentrus nattereri*.

ratio tests suggested that the two-ratio model was better than the one-ratio model ($\omega_{CF} > \omega_{SF}$) and the best model was the free-ratio model (supplementary table S2, Supplementary Material online). With the free-ratio model, ω was higher in the blind *S. anshuiensis* (0.36) than in the small-eyed *S. rhinocerous* (0.32) and the eyed *S. grahami* (0.23) with vision genes. With circadian clock genes, ω was higher in the blind *S. anshuiensis* (0.38) and the small-eyed *S. rhinocerous* (0.37) than in the eyed *S. grahami* (0.25). With pigmentation genes, ω was higher in the small-eyed *S. rhinocerous* (0.32) and the blind *S. anshuiensis* (0.29) than in the eyed *S. grahami* (0.25) (supplementary fig. S4, Supplementary Material online). Thus, ω was consistently higher in cavefishes than in surface fishes, the shift being larger for vision genes than for circadian clock and pigmentation genes. These results suggest a larger reduction of purifying selection on the vision gene set than on the other gene sets in cavefishes.

We then used another approach implemented in RELAX which computes the values and distribution of three ω using a branch-site model, the convergence of the three ω toward one in a lineage being a signature of relaxed purifying selection (Wertheim et al. 2015). The magnitude of convergence depends on a parameter, k, which tends to zero as selection tends to complete relaxation. RELAX detected relaxed purifying selection on *L. dentata* vision genes with an important shift toward $\omega = 1$ as $k = 0.2$, and this was also true to a lesser extent in *A. mexicanus* cavefish ($k = 0.5$). For pigmentation genes, the largest shift was also observed in *L. dentata* ($k = 0.48$). No shift was observed with cavefish circadian clock genes, suggesting that most of these genes are under strong purifying selection in cavefishes (supplementary figs. S5–S7, Supplementary Material online).

Finally, with the aim of finding additional and independent evidence of relaxed purifying selection in cavefishes, in particular on *A. mexicanus* vision genes for which the number of mutations is particularly low and thus, the estimate of ω was not accurate, we developed a novel approach. First, nonsynonymous mutations in different lineages were inferred using the aaml program from the PAML package. The deleterious impact of these mutations was then estimated using a machine learning method implemented in MutPred2 (Pejaver V, Urresti J, Lugo-Martinez J, Pagel KA, Lin GN, Nam H-J, Mort M, Cooper DN, Sebat J, Iakoucheva LM, Mooney SD and Radivojac P 2017. MutPred2: inferring the molecular and phenotypic impact of amino acid variants, unpublished data. Available from: https://www.biorxiv.org/content/10.1101/134981v1) which gives a score between 0 (not deleterious) and 1 (very deleterious). The kernel density estimation of the distributions of the scores in vision, circadian clock, and pigmentation genes were obtained for each terminal lineage leading to surface fishes and cavefishes, as well as for computer simulations of substitutions in the same gene sets under a neutral model. With all surface fishes, the kernel density estimation was similarly right-skewed (fig. 5), suggesting that most substitutions in surface fishes have a low impact on fitness. This was confirmed by the shape of the distribution of the scores in simulations of substitutions without selection (equivalent to the distribution of the scores before selection) which was very different to those of surface fishes, that is almost uniform, suggesting that the most deleterious mutations had been removed by selection in surface fishes. Before selection, the score distribution was slightly different for the different sets of genes, probably reflecting different selective constraints on the sequences belonging to these gene sets (fig. 5A–C, gray and black curves). The transitions/transversions (Ts/Tv) ratio used in simulations of substitutions under a neutral model had no impact on the distribution of the scores (supplementary fig. S8, Supplementary Material online). In the cavefishes on the other hand, the score

Downloaded from https://academic.oup.com/mbe/article/38/2/589/5912537 by guest on 29 June 2021

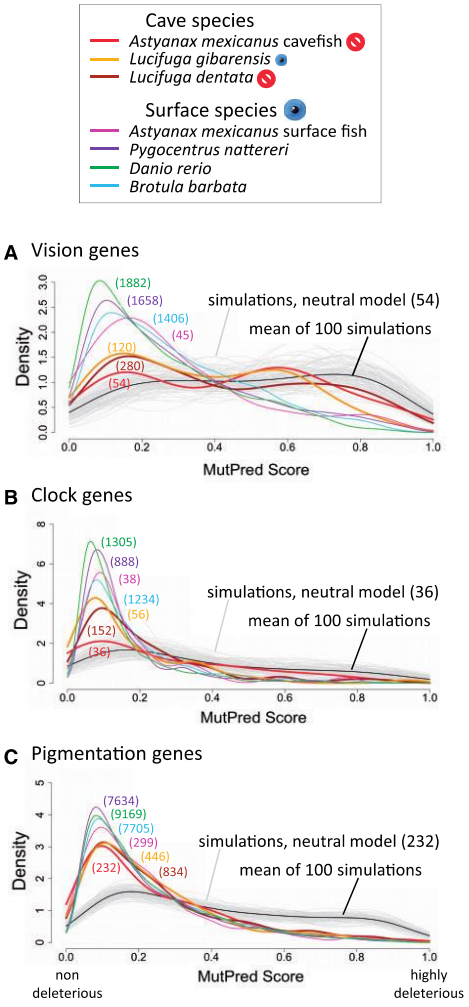


FIG. 5. Distributions of MutPred2 scores in different fish lineages and in simulations of mutations without selection. This score ranges between 0 (nondeleterious mutation) and 1 (highly deleterious mutation). The number of mutations in each lineage is given in round brackets. One hundred simulations were performed on each gene set. In each simulation, 54 nonsynonymous mutations were generated in vision genes, 36 in circadian clock genes, and 232 in pigmentation genes, those numbers corresponding to the numbers of nonsynonymous mutations found in *Astyanax mexicanus* cavefish.

distribution was very variable, depending on the cavefish species and the set of genes (fig. 5*A*–*C*). Pairwise comparisons of empirical cumulative distribution functions (ECDF) were performed using the nonparametric Kolmogorov–Smirnov (KS) test (supplementary fig. S9, Supplementary Material online). The same approach was attempted using Grantham's distances (Grantham 1974) instead of MutPred2 scores but the contrast between the distributions of the distances with and without selection was much less discriminant and not analyzed further (supplementary fig. S10, Supplementary Material online).

In order to refine the analysis of the score distribution in cavefishes, mixtures of different proportions of substitutions picked from two distributions, one under neutral evolution (from the simulations) and the other under purifying selection (in the lineage leading to zebrafish) were also obtained to compare with cavefish distributions (supplementary fig. S11, Supplementary Material online).

Vision Genes

For the *A. mexicanus* cavefish (red curve, fig. 5*A*), the distribution was not statistically different from that expected if all substitutions were neutral in this lineage (KS test, $P = 0.2$; supplementary fig. S9, Supplementary Material online), yet the best fit was with a mixture distribution with 24% of substitutions from the distribution under purifying selection (supplementary fig. S11, Supplementary Material online). For *L. dentata* (brown curve, fig. 5*A*) and *L. gibarensis* (orange curve, fig. 5*A*), distributions departed from the neutral distribution (KS test, $P = 1.4 \times 10^{-5}$ and $P = 4 \times 10^{-6}$, respectively) (supplementary fig. S9, Supplementary Material online) and the best fit was obtained with 34% and 60% of the substitutions from the distribution under purifying selection, respectively (supplementary fig. S11, Supplementary Material online). For all *Sinocyclocheilus* species, the score distribution was different from those of surface fishes (supplementary fig. S12, Supplementary Material online). This was even true for the eyed *S. grahami*, most likely because after the WGD, purifying selection on nonsynonymous mutations was partially relaxed on one or both ohnologs. However, the ECDF of *S. rhinocerous* and *S. anshuiensis* were more shifted toward the neutral distribution than the ECDF of *S. grahami*, suggesting that the two cavefishes experienced a more neutral regime than the surface fish (supplementary fig. S12, Supplementary Material online).

Circadian Clock Genes

No cavefish ECDF fit with the expected distribution under neutral evolution (fig. 5*B* and supplementary fig. S9, Supplementary Material online). However, the ECDF of *A. mexicanus* cavefish was different from those of surface fishes and the best fit was obtained with a mixture of 59% of the substitutions from the distribution under purifying selection (fig. 5*B* and supplementary fig. S11, Supplementary Material online). For *L. dentata* and *L. gibarensis*, the best fit involved the mixture of 69% or 93% of the substitutions from the distribution under purifying selection (fig. 5*B* and supplementary fig. S11, Supplementary Material online). In accordance with the number of pseudogenes found in *S. rhinocerous*, the ECDF was the closest to a neutral distribution among the three *Sinocyclocheilus* species, with the best fit found with a mixture of 39% of substitutions

Downloaded from https://academic.oup.com/mbe/article/38/2/589/5912537 by guest on 29 June 2021

from the distribution under purifying selection (supplementary figs. S12 and S13, Supplementary Material online).

Pigmentation Genes

No cavefish ECDF fitted with the expected distribution under neutral evolution (fig. 5C). All cavefish distributions were very similar to those of surface fishes, in accordance with the hypothesis that very few pigmentation genes can be lost, even after cave colonization and/or genome duplication (see also supplementary figs. S9 and S11–S13, Supplementary Material online).

In summary, three different approaches consistently suggested different levels of relaxed purifying selection on the set of vision genes that are related with the levels of eye degeneration in cavefishes, whereas most circadian clock and pigmentation genes remained under strong purifying selection.

Dating Relaxation of Purifying Selection on Vision Genes in *L. dentata*

In order to make compatible the results suggesting that most vision genes are dispensable and those suggesting that selection is not totally relaxed in the *L. dentata* lineage, we postulated two successive periods of evolution, one under purifying selection followed by another under completely relaxed selection. Three independent approaches were used to estimate when purifying selection was relaxed in the *L. dentata* lineage.

First, we used the number of vision pseudogenes and an estimate of the LoF mutation rate per gene per generation. Using the numbers of premature STOP codons, frameshifts, losses of splice site, losses of START and STOP codons in this species, and the method described in supplementary data S3, Supplementary Material online, we estimated the relative rates of these LoF mutations in *Lucifuga* species: 0.031, 0.0143, 0.0212, 0.0028, and 0.0023 μ, respectively (where μ is the nucleotide mutation rate per site per generation). In vertebrates, very few estimates of μ are available, the most accurate value being $\sim 10^{-8}$ in humans (Roach et al. 2010). Two independent values were recently obtained in cichlid fishes, one lower (3.5×10^{-9}) and one higher (6.6×10^{-8}) (Recknagel et al. 2013; Malinsky et al. 2018). Assuming that $\mu = 10^{-8}$ and taking into account that the average length of a vision gene equals 1,091 bp, the LoF mutation rate per gene per generation was 0.78×10^{-6}. With the analytical model described in Materials and Methods and this rate of pseudogenization, the highest probability of finding 19 pseudogenes among 76 neutral genes was obtained with a complete relaxation of selection starting 367,779 generations ago (probability $> 5\%$ in a range between 273,990 and 480,980 generations) (fig. 6, red curve). Assuming that only 50 vision genes could accumulate LoF mutations, this time was pushed back to 611,132 (445,950–813,580) generations (fig. 6, pale red curve). Simulations were also performed to take into account variations of gene length and number of introns per gene, codon usage, transition/transversion ratio ($r = 4.57$), and effective population size (N_e) in a range between 100 and 1,000. These simulations gave estimates very similar to those obtained with the analytical model, showing that the effects of N_e and per gene LoF mutation rate variation due to differences in gene length and number of introns were marginal (fig. 6, black, green, and blue curves; only simulations assuming 76 neutral genes are shown).

Second, two dating methods were used (Li et al. 1981; Meredith et al. 2009), both based on the hypothesis of a shift of ω from a value lower than 1 to 1 after purifying selection was relaxed in a lineage. We assumed a divergence time of 80 My between *Lucifuga* and *Brotula* (http://www.timetree.org/) reflecting their rather distant position in two different families, Bythitidae and Ophidiidae, within Ophidiiformes (Møller et al. 2016). Vision genes of *Lucifuga* species and *B. barbata* were individually aligned and alignments concatenated. With one method (Li et al. 1981), the divergence time between *L. dentata* and *L. gibarensis* was estimated to 4.1 Ma and the time since nonfunctionalization of vision genes in *L. dentata* was 1.5 My (table 1). With the other method (Meredith et al. 2009), ω was estimated to 0.27 in the lineage leading to *L. gibarensis* and 0.50 in the lineage leading to *L. dentata*. Assigning these ratios respectively to functional branches and a mixed branch, the time since nonfunctionalization was estimated to 1.3 My (table 1).

Third, we assumed that in the lineage leading to *L. dentata*, there is a mixture of 66% of the mutations that accumulated under completely relaxed selection and 34% under purifying selection (supplementary fig. S11, Supplementary Material online), $\omega = 0.27$ under purifying selection (that is ω estimated in *L. gibarensis*), $\omega = 1$ under completely relaxed selection and the divergence between *L. dentata* and *L. gibarensis* occurred 4.1 Ma (estimated above). Using the method described in Materials and Methods, we obtained an estimate, 1.4 My, of the age of relaxation of purifying selection (table 1). Thus, different methods for dating relaxation of purifying selection in the *L. dentata* lineage converged to $\sim$1.3–1.5 Ma. These estimates are compatible with the estimated time since vision genes could accumulate LoF mutations, that is $\sim$370,000 generations ago, if we assume a generation time of $\sim$4 years in *L. dentata*.

Distribution of LoF Mutations in Vision Pseudogenes of Cavefishes versus Fossorial Mammals

An extensive study of the regression of visual protein networks in three fossorial mammals, the Cape golden mole *Chrysochloris asiatica*, the naked mole-rats *Heterocephalus glaber*, and the star-nosed mole *Condylura cristata*, has been published (Emerling and Springer 2014). From this publication, we retrieved the number of pseudogenes, their names, and the number of LoF mutations per pseudogene in the three species. In the Cape golden mole, 18 pseudogenes were found among 65 vision genes, whereas only 11 pseudogenes were found in the naked mole-rat and seven in the star-nosed mole. Several independent LoF mutations were found in orthologous vision genes of fossorial mammals and cavefishes. The distributions of LoF mutations per pseudogene in these mammals and two blind cavefishes (*L. dentata* and *S. anshuensis*) were compared (fig. 7). *Astyanax mexicanus* cavefish, which is also blind, was not included in this comparison because there is only one LoF mutation in one gene in

Downloaded from https://academic.oup.com/mbe/article/38/2/589/5912537 by guest on 29 June 2021

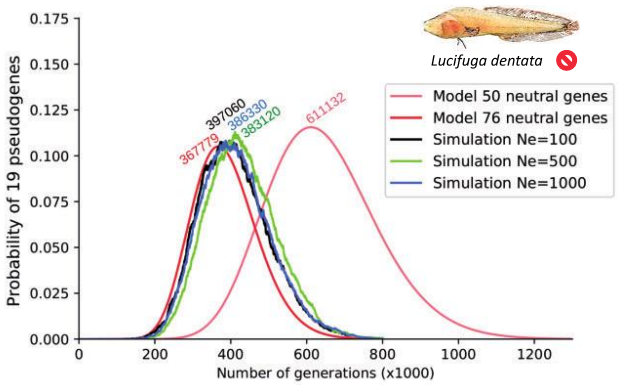
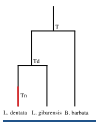


FIG. 6. Probability of finding 19 vision pseudogenes in *Lucifuga dentata* according to the period of neutral evolution. Red and pink lines, based on an analytical model assuming 76 and 50 neutral genes, respectively; other lines, estimates based on 10,000 simulations, assuming 76 neutral genes and taking into account the length and number of introns in each vision gene and considering different effective population sizes. The number of generations for which the highest probability was found is reported above each line.

Table 1. Estimates of the Period of Neutral Evolution of *Lucifuga dentata* Vision Genes.

	Method	Tn/Td	Tn
	Meredith et al. (2009)	0.32	1.3 My
	Based on Pejaver et al. (unpublished data)	0.34	1.4 My
	Based on Li and Nei (1977)	n/a	367,779 g
	Li et al. (1981)	0.36	1.5 My

NOTE.—Tn, period of neutral evolution; Td, divergence time of *L. dentata* and *L. gibarensis*; T, divergence time of *Brotula barbata* and *Lucifuga* spp.; My, million years; n/a, not applicable; g, generations.

this species. The distributions were sharply contrasted when comparing mammals and fishes. In fossorial mammals, most pseudogenes carried several LoF mutations, up to 28 mutations in two pseudogenes of the golden mole and 54 mutations in a single pseudogene of the star-nosed mole (fig. 7). On the contrary, in fishes, very few LoF mutations were found in each pseudogene (fig. 7), the maximum being five LoF mutations in a pseudogene of *S. anshuiensis*. This comparison strongly supports the hypothesis that some vision genes of fossorial mammals have been under completely relaxed purifying selection for a much longer period of time than any cavefish vision genes.

Discussion

When selection for maintaining a functional protein is completely relaxed, theory predicts that LoF mutations in coding and regulatory sequences can reach fixation by random genetic drift (Lynch and Conery 2000; Lahti et al. 2009). In an isolated population, among a set of dispensable genes, the longer the period of neutral evolution, the higher the expected number of pseudogenes, each carrying at least one LoF mutation. Eventually, all the genes under relaxed selection will be pseudogenized. At the level of a single gene, the longer the period of neutral evolution, the higher the expected number of LoF mutations. Thus, after a very long period of neutral evolution, all the neutrally evolving genes will carry many LoF mutations. The pace of this gene decay depends essentially on the pace of the appearance of LoF mutations (Li and Nei 1977). In the present study, we focused on a subset of LoF mutations that could be readily detected in genomes, that is, mutations that generate premature STOP codons, eliminate START or STOP codons, or disrupt intron splice sites, as well as small insertions/deletions (indels) causing translation frameshifts. Although this approach inevitably leads to an underestimation of the number of nonfunctional genes, it allowed us to carry out comparative studies and molecular dating of the relaxation of purifying selection. We analyzed molecular decay among three sets of genes in relation to the level of regression of three traits: eyes, circadian clock, and pigmentation. For most genes, assigning a gene to a category was straightforward, yet for some genes, it was more ambiguous. Vision genes corresponded primarily to a set of genes expressed only in eyes, however, fishes also express many nonvisual opsins genes that we assigned to this category on the basis of their homology to visual opsins and because most of them are likely just as dispensable in the absence of light as most eye-specific genes. Indeed, a similar pattern of gene decay was observed in these two gene sets and they were pooled for the quantitative analyses of relaxed selection and dating. Genes known for being involved in the circadian clock were assigned to a second set of genes. Some nonvisual opsins are involved in this process. Pigmentation genes comprised a large set of genes involved in several processes from pigment cell differentiation to pigment synthesis. Our a priori hypothesis was that vision genes should be more prone to degeneration in blind fishes as they are only expressed in eyes or involved in light sensing in various tissues, whereas many circadian clock and pigmentation genes may

Downloaded from https://academic.oup.com/mbe/article/38/2/589/5912537 by guest on 29 June 2021

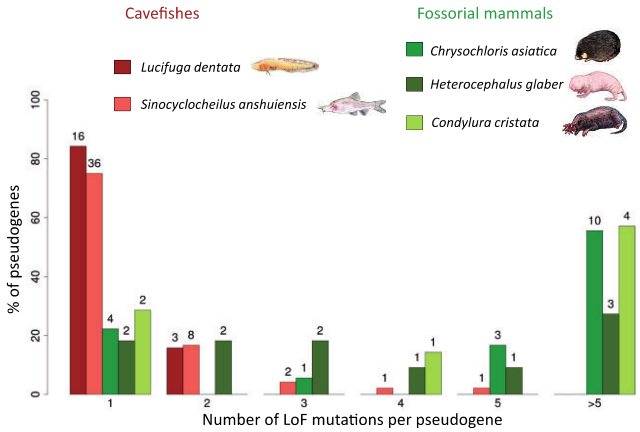


FIG. 7. Distributions of the number of LoF mutations per vision pseudogene in blind cavefishes and fossorial mammals. The number of pseudogenes is given above the bar.

be maintained as their expression is not always restricted to regressed structures and functions. In accordance with this assumption, many pseudogenes were found among vision genes of the oldest blind cavefish, *L. dentata*, but only few pseudogenes were found among circadian clock and pigmentation genes in cavefishes.

Below, we discuss the repeated loss of a few genes involved in circadian clock and pigmentation. Then, we show how the loss of many vision genes in *L. dentata* sheds new light on gene decay in relation to eye regression in cavefishes. At a broader phylogenetic scale, we examine the contrasting vision gene decay in cavefishes and fossorial mammals.

Molecular Evidence of Circadian Clock Disruption in Several Cavefishes

No LoF mutations were found in the set of circadian clock genes of both *Lucifuga* species. However, the nonvisual opsin *tmt3a* is pseudogenized in *L. gibarensis* and the loss of this gene is involved in the disruption of the circadian clock in the Somalian cavefish *Ph. andruzzii* (Cavallari et al. 2011). Thus, our survey of LoF mutations in Cuban cave brotulas suggests the loss of the circadian clock in *L. gibarensis*, but not in *L. dentata*. The maintenance of purifying selection on most circadian clock genes in both species is further supported by the analysis of nonsynonymous mutations showing similar accumulation of deleterious mutation in these species and in surface fishes. As expected, no LoF mutations in both ohnologs of circadian clock genes and nonvisual opsin genes were found in *S. grahami*, which is a surface fish. On the other hand, the small-eyed *S. rhinocerous* has accumulated more LoF mutations in both ohnologs (*per2*, *cry-dash*, *cry1b*, *cry2a*) than the blind *S. anshuiensis* (*cry-dash*). Interestingly, in both *Lucifuga* and *Sinocyclocheilus*, the molecular decay of the circadian clock is not correlated with the level of eye regression as the small-eyed species (*L. gibarensis* and *S. rhinocerous*) carry more pseudogenes than the blind species (*L. dentata* and *S. anshuiensis*).

LoF mutations were found repeatedly in a very small number of circadian clock genes, some of them already known to be involved in circadian clock disruption in other species. It suggests that only a small subset of genes may be involved in circadian clock loss, in particular those belonging to the cryptochromes and period families, which are light-inducible genes.

A Small Set of Pigmentation Pseudogenes in Cavefishes

A similar trend was observed in the large set of pigmentation genes. Independent LoF mutations were found in *myo7ab* and *smtla* of *L. dentata* and *L. gibarensis* and both ohnologs of *gch2* and *pmelb* carried independent LoF mutations in *S. anshuiensis* and *S. rhinocerous*. Recurrent pseudogenization of only a couple of genes suggests that few pigmentation genes can be lost in cavefishes. They could belong to a very small subset of genes that may be involved in the depigmented phenotype without having additional and strong deleterious side effects.

Many Pseudogenes among Cavefish Vision Genes

In sharp contrast with the very few pseudogenes found among circadian clock and pigmentation genes, many vision genes are pseudogenized in cavefishes. Before the present study, there was no evidence of decay in large numbers of both eye-specific genes and nonvisual opsin genes in a blind cavefish. Despite some cave populations belonging to *A. mexicanus* were assumed to be very ancient, that is several millions of years old, no LoF mutations were found in eye-specific genes of these fish with highly degenerate eyes (McGaugh et al. 2014). To explain this, very unlikely processes were assumed, such as "hidden functions" for these genes expressed only in eyes, and/or a high gene flow from the

Downloaded from https://academic.oup.com/mbe/article/38/2/589/5912537 by guest on 29 June 2021

surface preventing the fixation of pseudogenes. But a high gene flow implies strong selection at other loci to maintain blindness in very small cavefish populations. However, these unlikely hypotheses are not necessary if we assume that *A. mexicanus* cave populations are very recent (Fumey et al. 2018) and that there has simply been not enough time for the appearance of many vision pseudogenes.

Rapid and extreme eye degeneration without eye-specific gene losses further questions the nature of the developmental mechanisms involved in eye loss, the pace of eye degeneration, and the correlation of eye degeneration with eye-specific gene decay in cavefishes. A clear refutation of the hypothesis that eye-specific genes and nonvisual opsin genes are under purifying selection in blind cavefishes was provided by the analysis of the genome of *L. dentata*, as 25% of these genes carry LoF mutations in these species. Moreover, the distribution of LoF among genes was consistent with the neutral evolution of most, if not all, vision genes in this species. The dispensability of most vision genes was further supported by the fact that other vision genes were identified with LoF mutations in other cavefishes. We predict that with more blind cavefish genomes becoming available we will find that most of these genes have been lost in at least one species. On the other hand, in *L. gibarensis*, which has small but functional eyes, most vision genes seemed under purifying selection, yet the partial degeneration of its visual system was correlated with the loss of several genes that were well conserved in eyed fishes. These data allowed us to propose a two-step scenario for the release of purifying selection on vision genes in *Lucifuga*. The last common ancestor of *L. dentata* and *L. gibarensis* was a cavefish that had accumulated a small number of pseudogenes in relation to life in darkness, which were not among the eye-specific genes studied here. In the lineage leading to *L. gibarensis*, only a few vision genes have been under relaxed selection whereas in the lineage leading to *L. dentata*, purifying selection has been relaxed on most vision genes. The small-eyed *L. gibarensis* may not be an intermediary stage to blindness as observed in *L. dentata*, but another stable state. Small-eyes and eyeless phenotypes could correspond to different ecological parameters such as different exposures to light. In accordance with this hypothesis, *L. gibarensis* belongs to an ancient monophyletic clade of small-eyed species living in Cuba and Bahamas (García-Machado et al. 2011; Hernández et al. 2020).

The lack of correlation between the degree of eye regression and the number of eye-specific pseudogenes suggests that the extent of eye regression should not be taken as a proxy of the evolutionary age of cavefish populations or species.

Dating *L. dentata* Blindness

Although phylogenetic evidence suggested an ancient diversification in the cavefish genus *Lucifuga* and a monophyletic clade of distantly related blind species (García-Machado et al. 2011), there was no estimate of the time since this phenotype evolved. Similar estimates were obtained using several well-established methods and new approaches. With three methods relying on the shift of ω from a value lower than one (a signature of purifying selection) to one (a signature of neutral evolution), we found that the time since purifying selection was completely released on *L. dentata* vision genes is between 1.3 and 1.5 Ma. Using the number of pseudogenes in the set of vision genes, we estimated that *L. dentata* settled in caves ~370,000 generations ago. The generation time of this fish is unknown, and translating the number of generations into years is difficult. However, assuming that the generation time is ~4 years, which is realistic if we consider that these fish have a long lifespan and could reproduce for ~10 years, the above independent estimates of relaxed selection would be compatible. They suggest that the loss of vision in the lineage leading to *L. dentata* occurred in the middle of the Pleistocene.

Pattern of LoF Mutations in Tetraploids Cavefishes with Different Levels of Eye Regression: The Case of *Sinocyclocheilus*

The genus *Sinocyclocheilus*, which is endemic to southwestern karst areas in China, is the largest cavefish genus known to date (Xiao et al. 2005). In a genome-wide analysis, LoF mutations were found in many genes of three species, one species (*S. anshuiensis*) being blind and depigmented, another species (*S. rhinocerous*) having small eyes and being pigmented, and the last one (*S. grahami*) showing no cave-related traits (Yang et al. 2016). These species share a recent WGD with other cyprinids such as the common carp *Cyprinus carpio* (David et al. 2003; Yuan et al. 2010) which can explain why even the surface fish *S. grahami* carries many LoF mutations in vision, circadian clock, and pigmentation genes (Yang et al. 2016). However, no thorough comparisons were performed to relate differences in gene decay with the level of eye regression. We found that the number of vision pseudogenes in the blind *S. anshuiensis* is much higher than in the small-eyed *S. rhinocerous* and the eyed *S. grahami*, a result supporting the cumulative effect of tetraploidy and cave settlement on the rate of pseudogenization. As most genes are present twice, a gene function is most likely lost if, and only if, at least one LoF mutation is present in each ohnolog. With this criterion, seven vision genes were lost in *S. anshuiensis*, but only one in *S. rhinocerous* and *S. grahami*, indicating that selection maintaining functional vision genes is weaker in the blind species than in the two-eyed species. This conclusion was further supported by estimations of the strength of purifying selection on nonsynonymous mutations in vision genes, which showed that it was higher in the fish with large eyes.

Contrasting Dynamics of Pseudogenization in Fossorial Mammals and Cavefishes

An extensive study of LoF mutations in genes coding for proteins involved in retinal networks using the genomes of three independently evolved fossorial mammals has been previously reported (Emerling and Springer 2014). All three species have functional eyes, but star-nosed moles often leave their burrows and thus have a greater exposure to light than naked mole-rats and Cape golden moles, which are entirely subterranean. The eyes of Cape golden moles are

Downloaded from https://academic.oup.com/mbe/article/38/2/589/5912537 by guest on 29 June 2021

subcutaneous. More pseudogenes were identified in the Cape golden mole than in the naked mole-rat genome and the lowest number of pseudogenes was found in the star-nosed mole genome, suggesting that the decrease in retinal exposure to light allowed the decay of more vision genes. A striking difference between these fossorial mammals and the cavefishes studied here was that several pseudogenes of fossorial mammals carried a large number of LoF mutations, whereas cavefish pseudogenes accumulated at most five LoF mutations. This is in accordance with the adaptation of the fossorial mammals to the subterranean environment in the Oligocene, ~25 Ma (Emerling and Springer 2014), whereas colonization of the dark environment by the cavefishes occurred much later, in the Pleistocene.

Conclusion

Our analyses suggest that blind cavefishes for which genomes are available are not very ancient and that they all lost their eyes during the Pleistocene. The oldest, the Cuban *L. dentata*, in the middle of the Pleistocene and the most recent, the Mexican *A. mexicanus*, during the late Pleistocene or even later in the Holocene. Two gene-centered studies, based on molecular evolution of one and two genes, respectively, suggested that other blind cavefishes could be much older: some North American amblyopsid cavefishes may have accumulated LoF mutations over the last 10.3 My (Niemiller et al. 2013) and the Somalian cavefish *Ph. andruzzii* was estimated to be 5.3 My old (Calderoni et al. 2016). These estimations were based on very small numbers of nonsynonymous mutations and very few LoF mutations. Genome-wide analyses using several independent approaches are necessary to confirm that these cavefishes evolved much earlier than the cavefishes we analyzed here. If all cavefishes are actually relatively recent, the sequencing of a large number of blind cavefish genomes will be necessary to identify the whole set of eye-specific genes that are dispensable when eyes are highly degenerated, and the small subset of genes that has been repeatedly involved in the circadian clock disruption and depigmentation. Finding a blind cavefish genome in which most vision genes are pseudogenized and carry many LoF mutations would refute our current working hypothesis that blind cavefishes cannot thrive more than a few million years in cave ecosystems.

Materials and Methods

Genomic Resources for Two Cuban Cave Brotulas

A draft genome and a transcriptome were obtained for a specimen belonging to *L. dentata*. The reads of a specimen belonging to a closely related species, *L. gibarensis*, were mapped onto the *L. dentata* genome. A detailed description of the materials, methods, and results are given in supplementary data S1, Supplementary Material online. In the present study, these genomic resources were used to retrieve exon sequences and intron splice sites. The transcriptome allowed us to assess the reliability of the LoF mutations found in the genome of *L. dentata*.

Vision, Circadian Clock, and Pigmentation Gene Sets

The set of vision genes included all opsins, visual opsins that are expressed in eye photoreceptor cells (cone and rods) but also nonvisual opsins that are expressed in various tissues. It also comprised eye-specific crystallin genes that were selected using expression patterns reported in zebrafish from the ZFINdatabase (https://zfin.org/) and in *A. mexicanus* (Hinaux et al. 2015). However, *crygm2* crystallin genes were excluded from the analyses. Indeed, many copies were found in fish genomes (>50 copies in *A. mexicanus*) and relaxed purifying selection on some copies could have occurred independently in response to any environmental shift. The set of vision genes also included genes coding for proteins involved in the phototransduction cascade and whose expression was restricted to the retina and/or the pineal complex: RPE65, Arrestins, Recoverins, Transducins, PDE6, CNGA3 and CNGB3, GCAPs, zGCs, and GRKs. Sets of circadian clock and pigmentation genes were defined on the basis of gene lists established in previous studies (Li et al. 2013; Lorin et al. 2018). The set of circadian clock genes was completed with *ck1δa* and *ck1δb* genes which are specific kinases of *cry* and *per* genes (Takahashi et al. 2008) and *aanat1* and *aanat2* genes whose expressions are regulated by the circadian clock in zebrafish (Vatine et al. 2011). The complete list of genes with their standardized identifiers is given in figure 1.

Construction of Data Sets

The sequences of visual and nonvisual opsins of zebrafish were retrieved from (Davies et al. 2015). Other vision genes, circadian clock, and pigmentation genes of zebrafish were retrieved from GenBank.

A series of BlastN and TBlastX (Altschul et al. 1990) with zebrafish sequences were performed against *A. mexicanus* surface and Pachón cave genomes (GCF_000372685.2 and GCF_000372685.1, respectively), *S. grahami*, *S. rhinocerous*, *S. anshuiensis*, *P. nattereri*, *B. barbata*, *C. acus*, and *La. exutus* genomes (GCF_001515645.1, GCF_001515625.1, GCF_001515605.1, GCF_001682695.1, GCA_900303265.1, GCA_900312935.1, and GCA_900312555.1, respectively), and *L. dentata* and *L. gibarensis* genomes (this study). Matching regions were extracted using samtools (Li 2011) and CDS were predicted using EXONERATE with protein sequences of zebrafish (Slater and Birney 2005).

Given the low quality of the *A. mexicanus* cavefish genome assembly compared with the surface one (see in Results for an assessment of the quality of the *A. mexicanus* cavefish genome assembly), in order to analyze reliable gene sequences, cavefish reads were retrieved and mapped onto the genome of the surface fish using the NCBI remapping service with the default parameter values. This approach identified an opsin gene repertoire (36 genes) slightly larger than the one recently published (33 genes) using only the cavefish genome (Simon et al. 2019). Similarly, as a poor quality genome assembly was obtained for *L. gibarensis*, the reads were mapped on the high-quality *L. dentata* genome.

Incomplete CDS could be assembly artifacts, so they were not further analyzed. In order to estimate the percentage of missing exons in the *A. mexicanus* cavefish genome that were

Downloaded from https://academic.oup.com/mbe/article/38/2/589/5912537 by guest on 29 June 2021

assembly artifacts, primers were designed to PCR amplify these exons in a cavefish using homologous sequences identified in the *A. mexicanus* surface fish genome.

For comparisons of the strength of purifying selection (ω) in surface fishes and cavefishes, orthologous CDS from well-annotated fish genomes—tetraodon (*D. nigroviridis*), cod (*G. morhua*), stickleback (*Gas. aculeatus*), spotted gar (*Le. oculatus*), tilapia (*Oreochromis niloticus*), medaka (*Ory. latipes*), platyfish (*X. maculatus*)—were downloaded from ENSEMBL 93, excepted visual opsin CDS which were retrieved from a study of their evolution in ray-finned fishes (Lin et al. 2017).

In order to compare the decay of vision genes in cavefishes and fossorial mammals, the number of pseudogenes, and the number of LoF mutations per pseudogene, in a set of genes coding for proteins involved in retinal networks were retrieved from a publication (Emerling and Springer 2014) for three fossorial mammals (the Cape golden mole *Ch. asiatica*, the naked mole-rat *H. glaber*, and star-nosed mole *Co. cristata*).

Phylogenetic Analyses

Orthologous and paralogous relationships between genes were inferred through phylogenetic analyses. First, coding sequences were aligned using MUSCLE (Edgar 2004), after having taken into account indels (i.e., adding N where nucleotides were missing or removing additional nucleotides). For each alignment, DNA sequences were translated into protein sequences and a maximum likelihood phylogenetic tree was inferred using IQ-TREE (Nguyen et al. 2015) with the optimal model found by ModelFinder (Kalyaanamoorthy et al. 2017) and the robustness of the nodes was evaluated with 1,000 ultrafast bootstraps (Hoang et al. 2018). The trees were rooted and visualized using iTOL (Letunic and Bork 2007).

Estimation of the Proportion of Neutral Genes Using the Distribution of LoF Mutations per Gene

In order to estimate the number of genes (V) under completely relaxed selection in a sample of (T) vision genes, we compared the observed distribution of LoF mutations per gene with the expected distribution taking into account that V genes can accumulate LoF mutations and $T-V$ genes cannot carry LoF mutations. Assuming that an LoF mutation has a probability $1/V$ to be in a given gene among V genes, the probability that a gene contains X LoF mutations can be computed as follows:

$$p(X=0)=\frac{V}{T}\left(1-\frac{1}{V}\right)^{m}+\frac{T-V}{T} \quad \text{if } i=0$$
$$p(X=i)=\frac{V}{T}\frac{m!}{i!(m-i)!}\left(\frac{1}{V}\right)^{i}\left(1-\frac{1}{V}\right)^{m-i} \quad \text{if } i\neq 0 \quad ,$$

where m is the total number of LoF mutations.

However, LoF mutations do not have the same probability of occurring in different genes if the size of the gene and the number of introns are variable. An LoF mutation is more likely in a gene with several large exons than in a gene with only one short exon. Simulations of the LoF mutation distribution were performed, taking into account the length of the coding sequence and the number of introns of each gene to estimate its relative mutation rate. We ran 10,000 simulations of the distribution of m mutations in a random sample of V genes taken at random among T vision genes. The distributions of the number of LoF mutations per gene in *L. dentata* and *L. gibarensis* were compared with expected distributions obtained with the two methods described above and for different values of V.

Other Analyses of the Relaxation of Purifying Selection in Cavefishes

For each diploid species, gene sequences belonging to the same gene set (vision, circadian clock, or pigmentation) were concatenated. In order to analyze gene sequences of the tetraploid *Sinocyclocheilus* species, another alignment was produced. For each species and each gene set, gene sequences were concatenated after having taken at random one gene from each pair of ohnologs. This sampling process produced for each gene set two concatenated gene sequences for each species. With these six alignments of concatenated sequences, maximum likelihood estimates of ω were obtained using the program codeml from the PAML package Version 4.9h (Yang 2007). We considered three nested branch models: a "one-ratio" model assuming the same ω for all branches in the phylogeny; a "two-ratio" model assuming one ω for blind cavefishes (ω_{CF}) and one ω for the other fishes (ω_{SF}); a free-ratio model allowing a different ω for each branch. Likelihood ratio tests were performed to compare the likelihood values of the different models. Each likelihood ratio (LR) was calculated as twice the difference of log likelihood between the two models compared. The significance of the LR was determined by using the χ^2 distribution with the number of degrees of freedom equal to the difference between the number of estimated parameters in the two models.

A further approach for detecting relaxed selection was carried out using the program RELAX (Wertheim et al. 2015), assigning the surface fishes as the reference. Each cavefish was independently assigned as the test branch. A value of the parameter k significantly different from one ($k<1$ if selection is relaxed and $k>1$ if selection is intensified) suggests a change in the selective regime in the cavefish lineage.

Inferring the Deleterious Impact of Amino Acid Variants with MutPred2

Maximum likelihood inference of amino acids substitutions was performed using the program aaml from the PAML package Version 4.9h (Yang 2007). For each amino acid substitution, MutPred2 scores (Pejaver et al., unpublished data) and Grantham's distances (Grantham 1974) were computed to estimate its deleterious impact.

In order to compare the distribution of scores (or distances) for a set of genes within a branch with the distribution expected under completely relaxed purifying selection, simulations of random mutations of nucleotides were generated, taking into account the length of the coding sequence of each

Downloaded from https://academic.oup.com/mbe/article/38/2/589/5912537 by guest on 29 June 2021

gene and the transition/transversion ratio (https://github.com/MaximePolicarpo/Molecular-decay-of-light-processing-genes-in-cavefishes/blob/master/Neutral_evolution_for_mutpred.py).

Dating Relaxation of Purifying Selection on Vision Genes in *L. dentata* Using the Number of Pseudogenes

In absence of purifying selection, the probability of fixation of an LoF mutation in a gene after t generations is:

$$p(t) = 1 - e^{-\mu_{LoF}t} \text{ if } N_e \ll 1/\mu_{LoF},$$

where μ_{LoF} is the rate of LoF mutation per gene per generation and N_e is the effective population size (Li and Nei 1977).

For a set of T genes, assuming that each gene has the same rate of LoF mutation, the probability that D genes have fixed an LoF after t generations is:

$$p(X = D) = \frac{T!}{D!(T-D)!}\left(1 - e^{-\mu_{LoF}t}\right)^D \left(e^{-\mu_{LoF}t}\right)^{T-D}.$$

The derivative of this function with respect to t allows to find for which value of t the probability $p(X = D)$ is maximal:

$$t = \frac{1}{\mu_{LoF}} \ln\left(\frac{T}{T-D}\right).$$

The estimation of μ_{LoF} taking into account the length of the CDS and the number of introns is described in supplementary data S3, Supplementary Material online.

For assessing the effect of the variation of μ_{LoF} among genes on dating, a program was written to simulate gene decay through accumulation of LoF mutations, taking into account the length and the number of introns to compute a μ_{LoF} for each gene (more details are given in supplementary data S3, Supplementary Material online).

Other Methods for Dating Relaxation of Purifying Selection on Vision Genes in *L. dentata*

Vision genes from the two Cuban cave brotulas (*L. dentata* and *L. gibarensis*) and an outgroup (*B. barbata*) were concatenated and aligned. We assumed that vision genes have been under purifying selection along the branches of the phylogenetic tree, except in the lineage leading to *L. dentata* which is a mixed branch (with a period under purifying selection followed by a period under completely relaxed selection). The time since purifying selection was relaxed in this branch was estimated using two slightly different methods (Li et al. 1981; Meredith et al. 2009), and assuming that *B. barbata* and Cuban cave brotulas diverged 80 Ma (http://www.timetree.org/).

As an alternative approach, we used the distribution of MutPred2 scores in the lineage leading to *L. dentata*. First, we computed the proportions of two distributions, one under purifying selection as in the zebrafish lineage (p_s) and one without selection as in simulated data (p_n), that produce a mixture distribution that best fit the distribution of MutPred2 scores in the lineage leading to *L. dentata*. We assumed that ω_s under purifying selection shifted to ω_n when purifying selection was relaxed. We called T_d the time since the separation of *L. dentata* and *L. gibarensis*, t_s the period under purifying selection, and t_n the period under relaxed selection in the lineage leading to *L. dentata* (table 1). In this lineage, the proportion of nonsynonymous substitutions that accumulated under selection depends on ω_s and t_s and the proportion of nonsynonymous substitutions that accumulated under relaxed selection depends ω_n and t_n. Thus, $\frac{p_s}{p_n} = \frac{t_s\omega_s}{t_n\omega_n}$ or $t_n = \frac{\omega_s}{\omega_n}\frac{p_n}{p_s}t_s$.

Ethics approval

Animals were treated according to the French and European regulations for handling of animals in research.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

This work was supported by a collaborative grant from Agence Nationale de la Recherche (BLINDTEST to S.R. and D.C.) and from Institut Diversité Ecologie et Evolution du Vivant (to S.R. and D.C). We thank Yan Jaszczyszyn, Jean Mainguy, Nina Paffoni, and Isabelle Germon for their help in sequencing and analyzing the genomes of *Lucifuga dentata* and *L. gibarensis*, Patrick Laurenti for animal drawings, and Cushla Metcalfe and a reviewer for improving the English. We also thank Carlsbergfondet for financial support (Grant No. 2013_01_0501) for sampling *L. gibarensis*.

Sampling Authorization

Lucifuga dentata: a permit (LH 112 AN (135) 2013) was provided to the Centro de Investigaciones Marinas, University of Havana by the Cuban authorities in December 2013 to study cave species diversity including nematodes, crustaceans, and fishes. As the species was listed Vulnerable (VU) by the IUCN, only two adult individuals (MFP 18.000278) were sampled (12 January 2014) from one of its largest and demographically stable populations (Emilio Cave, Las Cañas, Artemisa Province, Cuba). *Lucifuga gibarensis*: a permit (PE 2014/82) was provided to the Centro de Investigaciones Marinas, University of Havana by the Cuban authorities in November 2014 to study cave species diversity including nematodes, crustaceans, and fishes. A single adult fish (MFP 18.000279) was sampled (20 November 2014) from the Macigo Cave (Aguada de Macigo del Jobal), Gibara, Holguín Province, Cuba.

Data Availability

Lucifuga dentata: Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession VXCM00000000. The version described in this article is version VXCM01000000. *Lucifuga dentata*: Transcriptome Shotgun Assembly project has been deposited at DDBJ/EMBL/GenBank under the accession GIAU00000000. The version described in this article is the first version, GIAU01000000. *Lucifuga gibarensis*: Raw sequences were

Downloaded from https://academic.oup.com/mbe/article/38/2/589/5912537 by guest on 29 June 2021

submitted to the SRA Bioproject: PRJNA610231. The original GFF3 annotation file of *L. dentata* and scaffolds smaller than 200 bp are available in Supplementary Material online. Python programs and R scripts used in this article can be found in: https://github.com/MaximePolicarpo/Molecular-decay-of-light-processing-genes-in-cavefishes. All the data and the results of the analyses performed in this study are available for download at figshare (https://doi.org/10.6084/m9.figshare.12666758). File Data_Supp1.xlsx: summary of the number of genes retrieved from each species and for each gene set, along with the number of pseudogenes and the number of LoF mutations; file Data_Supp2.xlsx: sequences predicted with EXONERATE and ID of sequences retrieved from ENSEMBL; file Data_Supp3.xlsx: results obtained with different methods for dating relaxed selection on vision genes in *L. dentata*; folder Concatenated_Alignments: concatenated alignments for vision, circadian clock, and pigmentation genes; folder Divergence_values: pairwise nucleotidic distances between species for each gene set; folder Lucifuga_Supplementary_files_Genome: original GFF3 file with functional annotations and scaffolds smaller than 200 bp not uploaded to NCBI; folder MutPred2_Results: raw output of MutPred2 and parsed results files to be used with the script provided in github (MutPred2_Script.R); folder Phylogenies: Gene phylogenies computed with iQTree and displayed with iTOL, the model used for each phylogeny can be found in the "Models" folder.

References

Albalat R, Cañestro C. 2016. Evolution by gene loss. *Nat Rev Genet*. 17(7):379–391.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol*. 215(3):403–410.

Berning D, Adams H, Luc H, Gross JB. 2019. In-frame indel mutations in the genome of the blind Mexican cavefish, *Astyanax mexicanus*. *Genome Biol Evol*. 11(9):2563–2573.

Calderoni L, Rota-Stabelli O, Frigato E, Panziera A, Kirchner S, Foulkes NS, Kruckenhauser L, Bertolucci C, Fuselli S. 2016. Relaxed selective constraints drove functional modifications in peripheral photoreception of the cavefish *P. andruzzii* and provide insight into the time of cave colonization. *Heredity* 117(5):383–392.

Cavallari N, Frigato E, Vallone D, Fröhlich N, Lopez-Olmeda JF, Foà A, Berti R, Sánchez-Vázquez FJ, Bertolucci C, Foulkes NS. 2011. A blind circadian clock in cavefish reveals that opsins mediate peripheral clock photoreception. *PLoS Biol*. 9(9):e1001142.

Ceinos RM, Frigato E, Pagano C, Fröhlich N, Negrini P, Cavallari N, Vallone D, Fuselli S, Bertolucci C, Foulkes NS. 2018. Mutations in blind cavefish target the light-regulated circadian clock gene, period 2. *Sci Rep*. 8(1):8754.

Culver DC, Pipan T. 2009. The biology of caves and other subterranean habitats. Oxford: Oxford University Press.

David L, Blum S, Feldman MW, Lavi U, Hillel J. 2003. Recent duplication of the common carp (*Cyprinus carpio* L.) genome as revealed by analyses of microsatellite loci. *Mol Biol Evol*. 20(9):1425–1434.

Davies WIL, Tamai TK, Zheng L, Fu JK, Rihel J, Foster RG, Whitmore D, Hankins MW. 2015. An extended family of novel vertebrate photopigments is widely expressed and displays a diversity of function. *Genome Res*. 25(11):1666–1679.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 32(5):1792–1797.

Emerling CA. 2018. Regressed but not gone: patterns of vision gene loss and retention in subterranean mammals. *Integr Comp Biol*. 58(3):441–451.

Emerling CA, Springer MS. 2014. Eyes underground: regression of visual protein networks in subterranean mammals. *Mol Phylogenet Evol*. 78:260–270.

Fang X, Nevo E, Han L, Levanon EY, Zhao J, Avivi A, Larkin D, Jiang X, Feranchuk S, Zhu Y, et al. 2014. Genome-wide adaptive complexes to underground stresses in blind mole rats *Spalax*. *Nat Commun*. 5(1):3966.

Fang X, Seim I, Huang Z, Gerashchenko MV, Xiong Z, Turanov AA, Zhu Y, Lobanov Alexei V, Fan D, Yim SH, et al. 2014. Adaptations to a subterranean environment and longevity revealed by the analysis of mole rat genomes. *Cell Rep*. 8(5):1354–1364.

Fumey J, Hinaux H, Noirot C, Thermes C, Rétaux S, Casane D. 2018. Evidence for late Pleistocene origin of *Astyanax mexicanus* cavefish. *BMC Evol Biol*. 18(1):43.

García-Machado E, Hernández D, García-Debrás A, Chevalier-Monteagudo P, Metcalfe C, Bernatchez L, Casane D. 2011. Molecular phylogeny and phylogeography of the Cuban cave-fishes of the genus Lucifuga: Evidence for cryptic allopatric diversity. Mol Phylogenetics Evol. 61(2):470–483.

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

Gross JB, Borowsky R, Tabin CJ. 2009. A novel role for Mc1r in the parallel evolution of depigmentation in independent populations of the cavefish *Astyanax mexicanus*. *PLoS Genet*. 5(1):e1000326.

Gross JB, Weagley J, Stahl BA, Ma L, Espinasa L, McGaugh SE. 2018. A local duplication of the melanocortin receptor 1 locus in *Astyanax*. *Genome* 61(4):254–265.

Hernández D, Møller PR, Casane D, García-Machado E. 2020. A new species of the cave-fish genus *Lucifuga* (Ophidiiformes, Bythitidae), from eastern Cuba. *ZooKeys* 946:17–35.

Hinaux H, Blin M, Fumey J, Legendre L, Heuze A, Casane D, Retaux S. 2015. Lens defects in *Astyanax mexicanus* cavefish: evolution of crystallins and a role for alphaA-crystallin. *Dev Neurobiol*. 75(5):505–521.

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: improving the ultrafast bootstrap approximation. *Mol Biol Evol*. 35(2):518–522.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods*. 14(6):587–589.

Kim EB, Fang X, Fushan AA, Huang Z, Lobanov AV, Han L, Marino SM, Sun X, Turanov AA, Yang P, et al. 2011. Genome sequencing reveals insights into physiology and longevity of the naked mole rat. *Nature* 479(7372):223–227.

Lahti DC, Johnson NA, Ajie BC, Otto SP, Hendry AP, Blumstein DT, Coss RG, Donohue K, Foster SA. 2009. Relaxed selection in the wild. *Trends Ecol Evol*. 24(9):487–496.

Letunic I, Bork P. 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* 23(1):127–128.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27(21):2987–2993.

Li W-H, Gojobori T, Nei M. 1981. Pseudogenes as a paradigm of neutral evolution. *Nature* 292(5820):237–239.

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

Li Y, Li G, Wang H, Du J, Yan J. 2013. Analysis of a gene regulatory cascade mediating circadian rhythm in zebrafish. *PLoS Comput Biol*. 9(2):e1002940.

Lin J-J, Wang F-Y, Li W-H, Wang T-Y. 2017. The rises and falls of opsin genes in 59 ray-finned fish genomes and their implications for environmental adaptation. *Sci Rep*. 7(1):15568.

Liu Z, Wen H, Hailer F, Dong F, Yang Z, Liu T, Han L, Shi F, Hu Y, Zhou J. 2019. Pseudogenization of *Mc1r* gene associated with transcriptional changes related to melanogenesis explains leucistic phenotypes in

Downloaded from https://academic.oup.com/mbe/article/38/2/589/5912537 by guest on 29 June 2021

Oreonectes cavefish (Cypriniformes, Nemacheilidae). *J Zool Syst Evol Res*. 57(4):900–909.
Lorin T, Brunet FG, Laudet V, Volff J-N. 2018. Teleost fish-specific preferential retention of pigmentation gene-containing families after whole genome duplications in vertebrates. *G3 (Bethesda)* 8:1795–1806.
Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* 290(5494):1151–1155.
Malinsky M, Svardal H, Tyers AM, Miska EA, Genner MJ, Turner GF, Durbin R. 2018. Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. Nat Ecol Evol. 2(12):1940–1955.
McGaugh SE, Gross JB, Aken B, Blin M, Borowsky R, Chalopin D, Hinaux H, Jeffery WR, Keene A, Ma L, et al. 2014. The cavefish genome reveals candidate genes for eye loss. *Nat Commun*. 5(1):5307.
Meredith RW, Gatesy J, Murphy WJ, Ryder OA, Springer MS. 2009. Molecular decay of the tooth gene Enamelin (ENAM) mirrors the loss of enamel in the fossil record of placental mammals. *PLoS Genet*. 5(9):e1000634.
Møller PR, Knudsen SW, Schwarzhans W, Nielsen JG. 2016. A new classification of viviparous brotulas (Bythitidae) – with family status for Dinematichthyidae – based on molecular, morphological and fossil data. *Mol Phylogenet Evol*. 100:391–408.
Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*. 32(1):268–274.
Niemiller ML, Fitzpatrick BM, Shah P, Schmitz L, Near TJ. 2013. Evidence for repeated loss of selective constraint in rhodopsin of amblyopsid cavefishes (Teleostei: Amblyopsidae). *Evolution* 67(3):732–748.
Protas M, Conrad M, Gross JB, Tabin C, Borowsky R. 2007. Regressive evolution in the Mexican cave tetra, *Astyanax mexicanus*. *Curr Biol*. 17(5):452–454.
Recknagel H, Elmer KR, Meyer A. 2013. A hybrid genetic linkage map of two ecologically and morphologically divergent midas cichlid fishes (*Amphilophus* spp.) obtained by massively parallel DNA sequencing (ddRADSeq). *G3 (Bethesda)* 3:65–74.
Roach JC, Glusman G, Smit AFA, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M, et al. 2010. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* 328(5978):636–639.
Rose MR, Lauder GV. 1996. Adaptation. San Diego: Academic Press.
Simon N, Fujita S, Porter M, Yoshizawa M. 2019. Expression of extraocular opsin genes and light-dependent basal activity of blind cavefish. *PeerJ* 7:e8148.
Slater GSC, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6(1):31.
Takahashi JS, Hong H-K, Ko CH, McDearmon EL. 2008. The genetics of mammalian circadian order and disorder: implications for physiology and disease. *Nat Rev Genet*. 9(10):764–775.
Vatine G, Vallone D, Gothilf Y, Foulkes NS. 2011. It's time to swim! Zebrafish and the circadian clock. *FEBS Lett*. 585(10):1485–1494.
Wertheim JO, Murrell B, Smith MD, Kosakovsky Pond SL, Scheffler K. 2015. RELAX: detecting relaxed selection in a phylogenetic framework. *Mol Biol Evol*. 32(3):820–832.
Xiao H, Chen S-y, Liu Z-M, Zhang R-D, Li W-X, Zan R-G, Zhang Y-P. 2005. Molecular phylogeny of *Sinocyclocheilus* (Cypriniformes: Cyprinidae) inferred from mitochondrial DNA sequences. *Mol Phylogenet Evol*. 36(1):67–77.
Yang J, Chen X, Bai J, Fang D, Qiu Y, Jiang W, Yuan H, Bian C, Lu J, He S, et al. 2016. The *Sinocyclocheilus* cavefish genome provides insights into cave adaptation. *BMC Biol*. 14(1):13.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 24(8):1586–1591.
Yuan J, He Z, Yuan X, Jiang X, Sun X, Zou S. 2010. Speciation of polyploid Cyprinidae fish of common carp, crucian carp, and silver crucian carp derived from duplicated Hox genes. *J Exp Zool*. 314B(6):445–456.
Zhao H, Di Mauro G, Lungu-Mitea S, Negrini P, Guarino AM, Frigato E, Braunbeck T, Ma H, Lamparter T, Vallone D, et al. 2018. Modulation of DNA repair systems in blind cavefish during evolution in constant darkness. *Curr Biol*. 28(20):3229–3243.

Downloaded from https://academic.oup.com/mbe/article/38/2/589/5912537 by guest on 29 June 2021

2) Evolutionary Dynamics of the OR Gene Repertoire in Teleost Fishes: Evidence of an Association with Changes in Olfactory Epithelium Shape

Evolutionary Dynamics of the OR Gene Repertoire in Teleost Fishes: Evidence of an Association with Changes in Olfactory Epithelium Shape

Maxime Policarpo,[1] Katherine E. Bemis,[2] James C. Tyler,[3] Cushla J. Metcalfe,[4] Patrick Laurenti,[5] Jean-Christophe Sandoz,[1] Sylvie Rétaux,[6] and Didier Casane,[1,7,*]

[1]Université Paris-Saclay, CNRS, IRD, UMR Évolution, Génomes, Comportement et Écologie, Gif-sur-Yvette, France
[2]NOAA National Systematics Laboratory, National Museum of Natural History, Smithsonian Institution, Washington, DC, USA
[3]Department of Paleobiology, National Museum of Natural History, Smithsonian Institution, Washington, DC, USA
[4]Independent Researcher, Nambour, QLD, PO Box 21, Australia
[5]Université de Paris, Laboratoire Interdisciplinaire des Energies de Demain, Paris, France
[6]Université Paris-Saclay, CNRS, Institut des Neurosciences Paris-Saclay, Gif-sur-Yvette, France
[7]Université de Paris, UFR Sciences du Vivant, Paris, France

*Corresponding author: E-mail: didier.casane@egce.cnrs-gif.fr.
Associate Editor: David Irwin

Abstract

Teleost fishes perceive their environment through a range of sensory modalities, among which olfaction often plays an important role. Richness of the olfactory repertoire depends on the diversity of receptors coded by homologous genes classified into four families: OR, TAAR, VR1, and VR2. Herein, we focus on the OR gene repertoire. While independent large contractions of the OR gene repertoire associated with ecological transitions have been found in mammals, little is known about the diversity of the OR gene repertoire and its evolution in teleost fishes, a group that includes more than 34,000 living species. We analyzed genomes of 163 species representing diversity in this large group. We found a large range of variation in the number of functional OR genes, from 15 in the Broad-nose Pipefish *Syngnathus typhle* and the Ocean Sunfish *Mola mola*, to 429 in the Zig-zag Eel *Mastacembelus armatus*. The number of OR genes was higher in species when a multilamellar olfactory rosette was present. Moreover, the number of lamellae was correlated with the richness of the OR gene repertoire. While a slow and balanced birth-and-death process generally drives the evolution of the OR gene repertoire, we inferred several episodes of high rates of gene loss, sometimes followed by large gains in the number of OR genes. These gains coincide with morphological changes of the olfactory organ and suggest a strong functional association between changes in the morphology and the evolution of the OR gene repertoire.

Key words: **OR genes, olfactory rosette, teleosts, gene family dynamics.**

Introduction

Olfaction is an important sensory modality in many animals because it serves essential functions such as feeding, reproduction, migration, kin recognition, and predator avoidance. In vertebrates, odorant molecules are primarily detected by olfactory sensory neurons, each expressing an olfactory receptor, following the one neuron—one receptor rule, although some exceptions have been reported (Sato et al. 2007). Olfactory receptors are G protein-coupled receptors coded by a large gene family (Buck and Axel 1991). The diversity of odors that an individual can discriminate depends on the richness of its olfactory gene repertoire, which results from the number and diversity of functional genes present in its genome. There is a great range of variation in the number and diversity of olfactory receptor genes among species (Nei et al. 2008; Niimura 2009; Khan et al. 2015; Sharma et al. 2019), which may reflect differences in the importance of olfaction relative to other sensory systems. Expansion and contraction of the olfactory receptor gene repertoire in a lineage are governed by a birth-and-death process; that is, recurrent gene duplications and losses that are involved in both the adaptive and nonadaptive evolution of olfaction (Nei et al. 2008).

Vertebrate olfactory receptors genes are coded by four large multigene families named olfactory receptor (OR), trace amine-associated receptor (TAAR), vomeronasal receptor type 1 and 2 (V1R and V2R) (Nei et al. 2008). OR and TAAR genes share sequence similarity and have no introns in the coding sequence. Studies of the evolution of the

© The Author(s) 2021. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Open Access

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab145/6254830 by guest on 29 June 2021

vertebrate olfactory gene repertoire have focused mainly on mammalian OR genes, and far less research has been undertaken on the OR genes of bony fishes (Osteichthyes). The average number of functional OR genes is about 800 in mammals, but the range of variation is very large, from 58 in the Common Bottlenose Dolphin *Tursiops truncatus* to 2,514 in the African Elephant *Loxodonta africana* (Hughes et al. 2018). In teleosts, the average number of OR genes is an order of magnitude smaller. Although few species of teleosts have been analyzed in prior studies, a wide range of variation has been observed (Niimura 2009; Jiang et al. 2019).

Vertebrate OR genes are classified into two types and eleven families that diverged before the separation of actinopterygians and sarcopterygians: Type 1 (α, β, γ, δ, ε, and ζ) and Type 2 (η, $\theta 1$, $\theta 2$, κ and λ). There is a clear separation between Type 1 and Type 2 genes, but the phylogenetic relationships among the gene families belonging to each type are not well established (Niimura and Nei 2005; Niimura 2009). In tetrapods, the γ family is greatly expanded and, to a lesser extent, the α family also is expanded. Sizes of the other gene families decreased and some disappeared in some lineages. In amniotes, most OR genes belong to the α or γ families. In contrast, teleosts lack α genes and only one γ gene exists in some species, but all other gene families are present (Lv et al. 2019). These observations led to the hypothesis that α and γ genes serve to detect airborne odorants while other gene families detect water-soluble odorants (Niimura and Nei 2005). Hence, gene families would have expanded or contracted differentially after the transition from an ancestral aquatic and aerial olfactory system, in the last common ancestor of bony fishes, to either only aerial olfaction in tetrapods or only aquatic olfaction in teleosts. At smaller phylogenetic scales, high rates of OR gene death have been linked to major ecological transitions. For example, cetaceans lost many OR genes because of their return to the sea (Liu et al. 2019). In primates, an acceleration of OR gene loss that occurred in the ancestral branch of haplorhines is associated with the acquisition of acute vision, and a high rate of OR gene loss was also observed at the ancestral branch of leaf-eating colobines related to the dietary transition from frugivory to folivory (Niimura et al. 2018). Conversely, an extreme expansion of the OR gene repertoire occurred in elephants, although the reason for this is unknown (Niimura et al. 2014). A recent study analyzed how the OR gene repertoire has undergone expansion and contraction across Mammalia with respect to several ecological adaptations (Hughes et al. 2018). Another study found a correlation between the number of OR genes and the relative size of the cribriform plate, a part of the ethmoid bone, and the area of its foramina, which are indicators of the relative number of olfactory sensory neurons (Bird et al. 2018).

In contrast, we know little about the evolution of OR in teleosts. With >34,000 species, teleosts represent about half of the extant vertebrate species and more than 95% of aquatic vertebrates and are an ideal model to study the evolution of the olfactory system in aquatic environments. Although fossils clearly assignable to Teleostei do not appear until the Mesozoic (252 Ma), the compilation of 24 molecular datings (Kumar et al. 2017) suggest that the last common ancestor of extant teleosts may have lived earlier, 284 Ma (244–294 Ma). Regardless of the exact age of teleosts, a vast period of time allowed great diversification of morphology, physiology, behavior, ecology, and habitat. Switches in sensory modalities have involved large numbers of gene births and deaths in teleosts. For example, several deep-sea fishes independently evolved rod opsin-based dim light vision relying on the expansion of RH1 gene repertoires (Musilova et al. 2019). In some cavefishes living in total darkness, loss of vision allowed the decay of many eye-specific genes (Policarpo et al. 2021). Much less is known about the evolution of olfaction and the dynamics of the OR gene repertoire size.

Since the seminal work of Burne at the beginning of the 20th century (Burne 1909), the relative importance of olfaction among sensory systems in teleosts has been assessed by studying the shape and organization of the olfactory epithelium. Most teleosts have a multilamellar olfactory epithelium that forms a rosette, with each lamella increasing the epithelial surface, suggesting that the olfactory system is well developed (Kasumyan 2004). The presence of a rosette is the ancestral state for teleosts (Hansen and Zielinski 2005), indicating that the sense of smell was well developed in their last common ancestor. However, simpler organizations have been observed in many species. For example, the olfactory epithelium is flat (0 lamellae) in the Broad-nose Pipefish *Syngnathus typhle* (Dymek et al. 2021), and there is only one lamella in the Round Goby *Neogobius melanostomus* (Hansen and Zielinski 2005). When a multilamellar olfactory rosette is present, the number of lamellae is highly variable, with up to 230 in the Barred Pargo *Hoplopagrus guentherii* (Hara 1975). Species with a simple olfactory epithelium (a flat olfactory epithelium or an olfactory rosette with very few lamellae), such as the Three-spined Stickleback *Gasterosteus aculeatus*, Crowned Seahorse *Hippocampus coronatus*, Green Pufferfish *Dichotomyctere fluviatilis* are assumed to have degenerate olfaction. Conversely, species that possess olfactory rosettes with many lamellae, such as the Japanese Eel *Anguilla japonica*, Zig-zag Eel *Mastacembelus armatus*, Channel Catfish *Ictalurus punctatus*, and Catla *Labeo catla*, may rely primarily on olfaction to perceive features of their environment. Although many teleost genomes are now available, correlation between the complexity of the olfactory rosette and the richness of the OR gene repertoire has not been investigated. To date, OR gene repertoires of teleosts have been reported for only a handful of species (Niimura and Nei 2005; Niimura 2009; Gao et al. 2017; Jiang et al. 2019; Lv et al. 2019), and no comparative analyses have been made.

The purpose of our study was to describe the evolutionary dynamics of the OR gene repertoire across the diversity of teleosts. We conducted large-scale phylogenetic analyses of the OR gene repertoires of 163 species to examine correlations between contractions and expansions of the OR gene family and changes in morphological traits, ecological parameters, and genome size. We found a positive correlation between the richness of the OR gene repertoire and both the presence of an olfactory rosette and the number of lamellae. In particular, we discovered that the two highest death rates

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab145/6265480 by guest on 29 June 2021

of OR genes coincide with two independent transitions from a multilamellar olfactory rosette to a simple olfactory epithelium. Moreover, a reacquisition of a multilamellar olfactory rosette coincides with a re-expansion of the OR repertoire. These observations suggest a strong functional link between changes in the morphology of the olfactory organ and the evolution of the OR gene repertoire.

Results

Diversity and evolution of OR gene repertoire in teleost fishes

The genomes of 307 teleost species and two nonteleostean actinopterygians (the Sterlet Sturgeon *Acipenser ruthenus* and the Spotted Gar *Lepisosteus oculatus*) were downloaded from NCBI and ENSEMBL databases.

We discarded six genomes for which the genome assembly size was not similar to the genome size estimated with other methods and compiled in the Animal Genome Size Database (Gregory 2020) (fig. 1A and supplementary fig. S1, Supplementary Material online). The completeness of the remaining 301 genomes was assessed using BUSCO (Waterhouse et al. 2018). We discarded all genomes for which <90% of the BUSCO genes were retrieved as complete coding sequences. This threshold allowed us to select a priori 163 teleost and two nonteleost genomes for which we could expect to identify a nearly complete set of OR genes. A higher threshold would have removed more species from consideration and would not have greatly improved the average quality of the genomes analyzed (fig. 1B).

We found 27,320 OR genes in the genomes of these 165 species. Among the OR genes, 19,248 were complete with seven transmembrane domains, and we refer to these as functional olfactory receptors; 4,868 genes carried at least one frameshift or a stop codon (also called loss-of-function [LoF] mutations) and were classified as pseudogenes. We also identified 2,344 truncated genes, which were incomplete sequences without any LoF mutations, and 860 edge genes, which were incomplete coding sequences located at scaffold edges (fig. 2A).

The number of truncated genes per genome was correlated with the number of pseudogenes ($r = 0.62$, P value = 2.2e−16), whereas the number of edge genes was not correlated with the number of pseudogenes ($r = 0.045$, P value = 0.57), or with the number of truncated genes ($r = 0.036$, P value = 0.65) (respectively supplementary fig. S2A–C, Supplementary Material online). These results suggest that a large fraction of truncated genes are due to large deletions and correspond to actual pseudogenes, whereas edge genes are mainly assembly artifacts.

We found a positive correlation between the number of pseudogenes and the number of functional genes ($r = 0.42$, P value = 2.5e−08) (supplementary fig. S3A, Supplementary Material online). The phylogenetic signal (i.e., the tendency of closely related species to resemble each other more than species drawn at random from the same phylogenetic tree) for both the number of functional ORs and the number of pseudogenes was strong (Pagel's $\lambda = 0.77$ and 0.94, respectively, and both significantly different from 0). Therefore, we applied a phylogenetic correction, after which the correlation still holds true ($r = 0.36$, P value = 2.1e−6) (supplementary fig. S3B, Supplementary Material online). This suggests that OR repertoire remodeling is a slow process. Conversely, there was no correlation between the proportion of pseudogenes and the number of functional genes ($r = -0.036$, $P = 0.65$) (supplementary fig. S2D, Supplementary Material online), suggesting that the rate of gene loss does not increase with the number of functional genes. These results are similar to those obtained with the OR genes of placental mammals (Niimura et al. 2014).

We assessed the relative efficacy of our pipeline to identify the OR gene repertoire in a genome, including both functional and nonfunctional sequences, by comparing the number of genes retrieved in the present study and in previous analyses of the OR repertoire in teleosts, that is the Zebrafish *Danio rerio*, Three-spined Stickleback *Gasterosteus aculeatus*, Medaka *Oryzias latipes*, Tiger Pufferfish *Takifugu rubripes*, Green Pufferfish *Dichotomyctere nigroviridis* (Niimura 2009), Channel Catfish *Ictalurus punctatus* (Gao et al. 2017), Chinese Perch *Siniperca chuatsi* (Lv et al. 2019), and Mariana Snailfish *Pseudoliparis swirei* (Jiang et al. 2019). We systematically found more OR genes than in these previous reports, and, for several species, many more functional sequences (table 1). This is probably because of the higher quality of more recent genome versions when compared with older versions used in previous studies. This may also result from a better pipeline implementation compared to other analyses based on the same genome versions (more details are given in supplementary fig. S4A–H, Supplementary Material online).

OR genes were classified into seven previously defined families (Niimura and Nei 2005): α, β, γ, ζ, ε, δ (Type 1) and η (Type 2). Genes of the α family are absent in teleosts, but they are present in the two nonteleost outgroups, *Acipenser ruthenus* and *Lepisosteus oculatus* (fig. 2A). Type 2 genes belonging to the families θ, κ, and λ were not analyzed because they may not be OR genes, most often present as a unique copy in vertebrates when they are not absent, with very rare cases of gene gains and losses (Niimura 2009).

To visualize gene sequence diversity within each gene family, we reconstructed a phylogenetic tree with the functional OR sequences of 38 species belonging to 38 teleost orders, each species having the most complete genome within that order. At the scale of teleosts, this phylogeny (supplementary fig. S5, Supplementary Material online) shows that the δ family contains many more genes and is much more diverse than the other families, followed by the ζ and η families. The β and ε families are much less diverse, and diversity is very low in the γ family. This observation also holds at the species level (fig. 2A).

Contrasting Richness of the OR Repertoire among Teleost Fishes

We found an unanticipated large range of variation in the number of OR genes among teleosts (fig. 2, supplementary fig. S6, Supplementary Material online). The mean number of

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab145/6265480 by guest on 29 June 2021

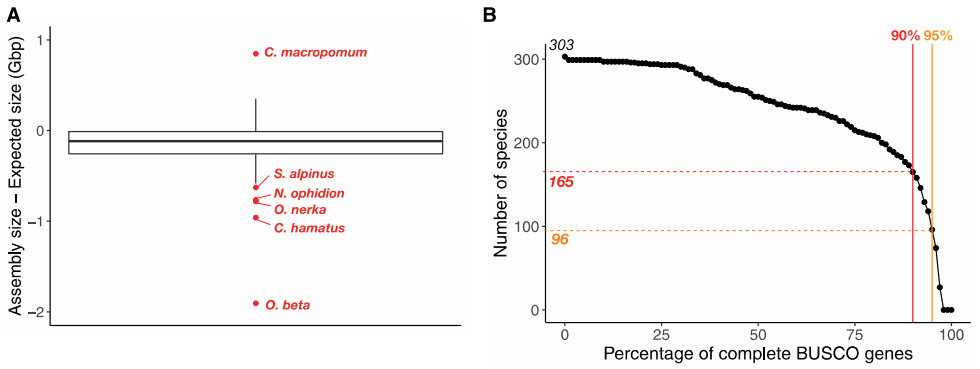


FIG. 1. Genome selection. (*A*) Congruence of genome assembly size and expected genome size based on haploid DNA content (C-value × 0.978 × 10^9) for 309 actinopterygian fishes (307 teleosts and 2 nonteleosts); six incongruent estimates of genome size are shown in red and species names indicated (more details in supplementary fig. S3, Supplementary Material online). (*B*) Number of genomes according to a completeness threshold (% of complete BUSCO genes). Out of 303 genomes with congruent estimates of genome size, 165 (163 teleosts and 2 nonteleosts) show a completeness > 90%, and 96 genomes (96 teleosts) a completeness > 95%.

functional OR genes per species was 117, with the smallest number (15) in the Broad-nose Pipefish *Syngnathus typhle* (Syngnathiformes) and the Ocean Sunfish *Mola mola* (Tetraodontiformes) and the highest (429) in the Zig-zag Eel *Mastacembelus armatus* (Synbranchiformes). The other Syngnathiformes species studied (*Syngnathus acus* and *Hippocampus comes*) were also OR-poor, and only a few δ and η genes were present (fig. 2*B*). In Tetraodontiformes, including *Mola mola*, *Thamnaconus septentrionalis*, *Dichotomyctere nigroviridis*, and three *Takifugu* species, the number of OR genes was more variable (fig. 2*C*).

The number of gene gains and losses along each branch of the phylogenetic tree was inferred using the gene tree—species tree reconciliation method (Chen et al. 2000), with one gene tree per OR gene family (fig. 2*A*). Birth and death rates were computed using these inferred numbers of gene gains and losses (supplementary fig. S7, Supplementary Material online). The mean birth and death rates were very similar, 0.0071 and 0.0073 per gene per million years, respectively (fig. 3), but the death rate was particularly high in the branches leading to Tetraodontiformes (0.3) and Syngnathiformes (0.32).

We also found two cases of high birth rates. The highest birth rate (0.2) was at the level of a deep internal node, in the common ancestor of Labriformes and Cyprinodontiformes while the second-highest birth rate (0.15) was observed in the common ancestor of *Perca* + *Sander* (fig. 2, supplementary fig. S7, Supplementary Material online).

The number of gene losses inferred in terminal branches correlated with the sum of the number of pseudogenes and truncated genes (fig. 4). Interestingly, the correlation was higher when considering only the shortest terminal branches ($r > 0.7$ with maximum branch length < 7 Myr) and decreased progressively when longer terminal branches were added ($r = 0.17$ with all terminal branches, see also supplementary fig. S8, Supplementary Material online). Similar results were obtained using either the number of pseudogenes or the number of truncated genes, but no correlation was found with edge genes (supplementary fig. S9, Supplementary Material online).

We ran simulations of sequence evolution involving nucleotide substitutions and small indels taking into account their relative frequencies estimated in a study of gene decay in teleosts (Policarpo et al. 2021). We found that about half of the pseudogenes were not detected after ~ 40 million generations (supplementary fig. S10, Supplementary Material online). If we also consider that large deletions can remove large fragments of nonfunctional genes and even complete sequences, then the time required for a large fraction of pseudogenes to fade away is probably much shorter. The continuous process of divergence and removal of nonfunctional genes may explain why the expected correlation between the inferred number of gene losses and the observed number of pseudogenes and truncated genes is high only when short terminal branches are analyzed.

Evolution of Olfactory Epithelium Folding in Teleost Fishes

Using the most recent hypotheses about teleost phylogeny and a compilation of descriptions on the olfactory organ, we inferred the evolution of the olfactory epithelium shape using maximum likelihood and parsimony methods. We conducted a literature survey of the shape of the olfactory epithelium for 220 actinopterygians beginning with the work of Burne (1909) (supplementary Data S1, Supplementary Material online). Our analysis indicated that a multilamellar olfactory rosette is ancestral, which was expected (Hansen and Zielinski 2005), but we also refined aspects of its evolution. In particular, the multilamellar olfactory rosette has been lost several times and regained in some lineages (fig. 2,

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab145/6265480 by guest on 29 June 2021

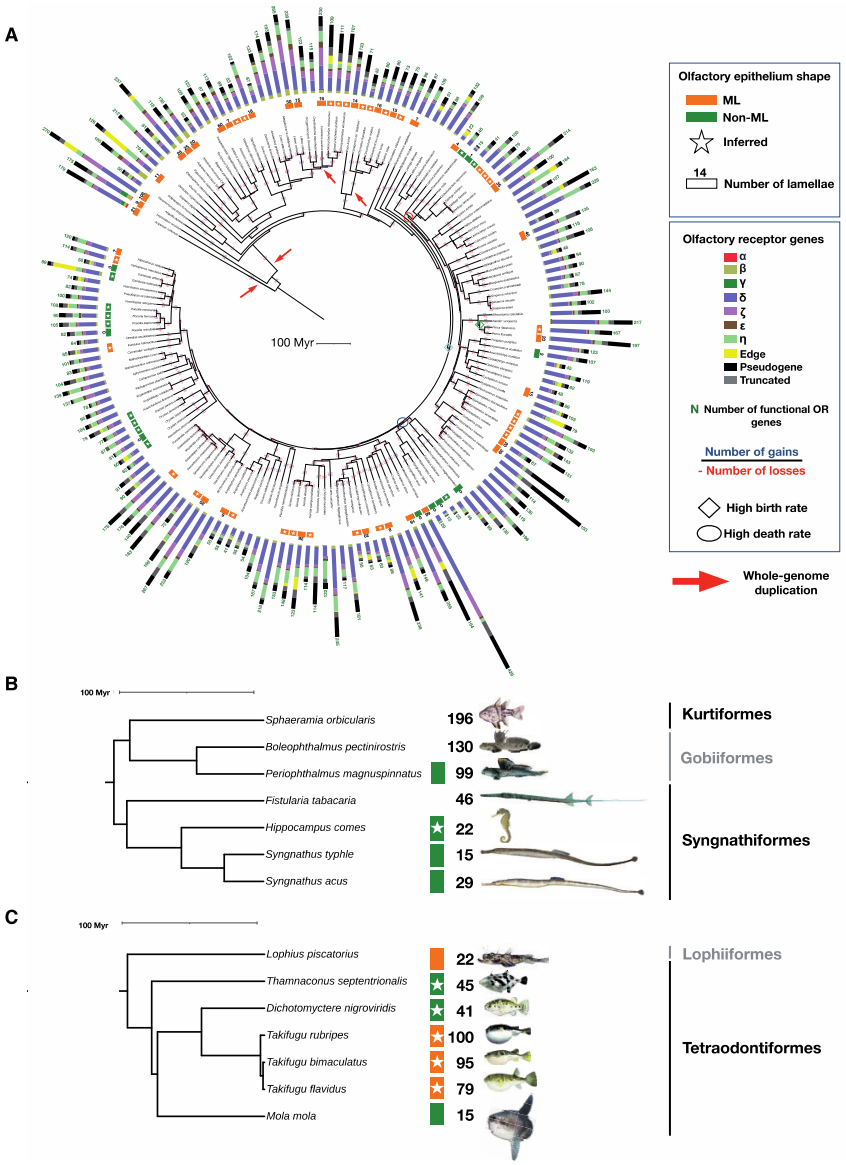


Fig. 2. Chronograms. (*A*) Dated phylogeny for 163 teleost fishes and two outgroups (from https://fishtreeoflife.org/). For each species, the total number of functional OR genes, the proportion of genes from each gene family, the proportion of truncated and edge genes, and the proportion of pseudogenes are provided. When available from the literature, presence or absence of a multilamellar olfactory rosette is reported (ML/non-ML). If found in the literature, the number of olfactory lamellae is also indicated. Presence or absence of a multilamellar olfactory rosette was inferred for some species without data on the shape of the olfactory epithelium using information available for a congeneric species (white star); the number of lamellae was not inferred because of intrageneric variation. Inferred number of gene gains and losses are provided on each branch (corresponding birth and death rates are given in supplementary fig. 7, Supplementary Material online). The branches associated with the two highest birth rates and the two highest death rates are indicated by diamond and oval symbols, respectively. Whole-genome duplications are indicated by red arrows. The tree was annotated and visualized using iTOL. (*B*) Focus on Syngnathiformes and related species. (*C*) Focus on Tetraodontiformes and related species.

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab145/6265480 by guest on 29 June 2021

Table 1. Comparison of the number of OR genes retrieved from the genome of eight species of teleosts in the present and previous studies.

	Present study			Previous studies			References
	F	I	P	F	I	P	
Danio rerio	162	4	8	154	1	21	Niimura (2009)
Gasterosteus aculeatus	107	11	33	102	5	52	Niimura (2009)
Oryzias latipes	81	3	5	68	6	24	Niimura (2009)
Takifugu rubripes	100	17	12	47	39	39	Niimura (2009)
Dichotomyctere nigroviridis	41	17	18	11	4	19	Niimura (2009)
Siniperca chuatsi	144	12	43	123	N/A	29	Lv et al. (2019)
Ictalurus punctatus	87	11	14	27	N/A	20	Gao et al. (2017)
Pseudoliparis swirei	48	9	10	43	0	10	Jiang et al. (2019)

NOTE.—F, functional genes; I, incomplete genes (truncated and edge); P, pseudogenes; N/A, data not available.
More details on this analysis can be found in supplementary fig. S4, Supplementary Material online.

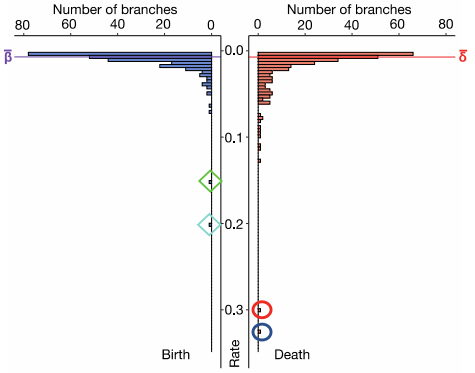


FIG. 3. Distribution of birth and death rates along the branches of the phylogenetic tree shown in fig. 2A. The mean values are indicated and the two highest birth rates and the two highest death rates are highlighted with diamond and oval frames, respectively.

supplementary fig. S11, Supplementary Material online). Maximum likelihood and parsimony methods gave similar results, except for a few internal nodes, such as the most recent common ancestor of Syngnathiformes and Gobiiformes + Kurtiformes, which is inferred to have a multilamellar olfactory rosette with the maximum likelihood MPPA and parsimony with DELTRAN (supplementary fig. S11*A* and *C*, Supplementary Material online), undefined with DOWNPASS (supplementary fig. S11*B*, Supplementary Material online) but a flat olfactory epithelium was inferred with ACCTRAN (supplementary fig. S11*D*, Supplementary Material online).

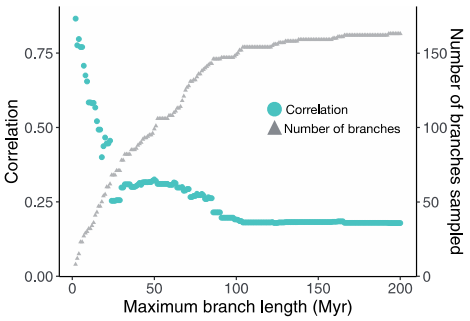


FIG. 4. Variation in the correlation between the number of gene losses inferred using a gene tree—species tree reconciliation method and the number of observed pseudogenes and truncated genes in the external branches of the phylogenetic tree according to a maximum branch length threshold ranging from 2 to 200 Myr. For each threshold, the number of branches is indicated by a gray triangle.

Correlation of the size of the OR repertoire with morphological and ecological traits

We examined correlations between olfactory epithelium shape, the number of olfactory lamellae, and the size of the OR gene repertoire. There is enormous diversity in the anatomy of olfactory epithelia, but herein we distill this information to a binary character for our analysis: (1) multilamellar olfactory rosette present (3 or more primary olfactory lamellae); (0) Multilamellar olfactory rosette absent (0–2 primary olfactory lamellae).

Among the 220 species for which we had information about olfactory epithelium shape, we had estimates of the size of the OR gene repertoire for 41 species. For 37 of the 41 species, we found information about the mean number of primary olfactory lamellae. For 38 additional species for which we had an estimate of the size of the OR gene repertoire, we inferred olfactory epithelium shape using information available for at least one species in the same genus (fig. 2, supplementary Data S1, Supplementary Material online), assuming that closely related species have similar olfactory epithelium (presence or absence of multilamellar olfactory rosette). However, we did not use information about the number of lamellae in a given species to infer this character at the genus level because it is more labile (Kasumyan 2004).

Species without a multilamellar olfactory rosette had significantly fewer functional OR genes than species with a multilamellar olfactory rosette (fig. 5*A* and *B*). Olfactory epithelium shape had a strong phylogenetic signal (Pagel's $\lambda = 0.999$, significantly different from 0). To account for the phylogenetic signal, we used a phylogenetic logistic model; the correlation between the olfactory epithelium shape and the number of OR genes was still significant (fig. 5*A* and *B*, supplementary fig. S12, Supplementary Material online). This correlation was statistically significant with the whole data set (*P* value = 0.035, 79 species, including the 38 species for which we inferred olfactory epithelium shape, fig. 5*A*) and with the reduced data set containing only 41 species for which

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab145/6265480 by guest on 29 June 2021

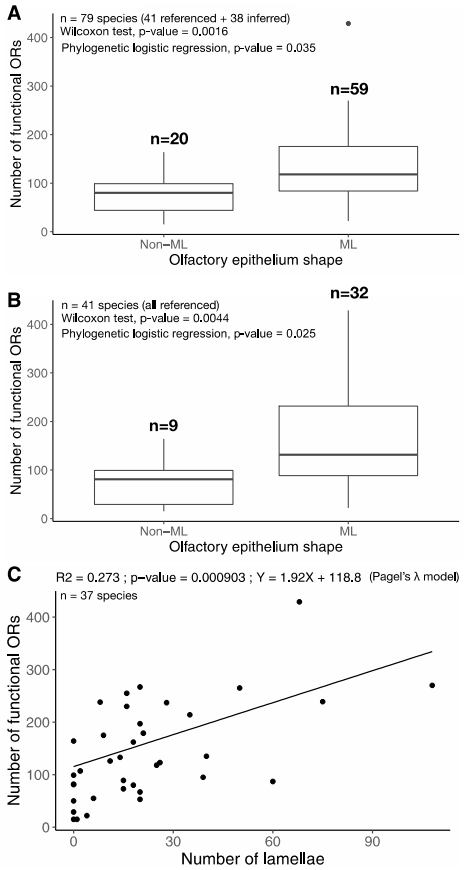


FIG. 5. Correlation between the number of functional OR genes and olfactory epithelium characteristics. (*A*) Correlation with absence (non-ML) or presence (ML) of a multilamellar olfactory epithelium, without taking into account the phylogeny (Wilcoxon test) or taking into account the phylogeny (phylogenetic logistic regression) and considering species for which data on the olfactory epithelium are either available or are available for a congeneric species. (*B*) Same analysis as (*A*), but restricted to species for which data are available. (*C*) Correlation between the number of lamellae and the number of functional OR genes taking into account the phylogeny (phylogenetic linear regression with a Pagel's λ model).

olfactory epithelium shape was directly known from the literature (*P* value = 0.025, fig. 5*B*).

We also found a positive correlation between the number of lamellae and the number of functional OR genes ($r = 0.58$, *P* value = 0.00015), which was also significant after phylogenetic correction with Pagel's λ model ($r = 0.52$, *P* value = 0.0009, AIC = 432.8) or an Ornstein-Uhlenbeck model ($r = 0.5$, *P* value = 0.0018, AIC = 431.2), and a Brownian motion model ($r = 0.33$, *P* value = 0.0432, AIC = 435.9) (fig. 5*C*). However, there was no correlation between the number of pseudogenes and olfactory epithelium shape, or the number of lamellae (supplementary fig. S13, Supplementary Material online).

We tested for a correlation between the number of functional OR genes and other morphological characters (maximum length of the fish, relative eye size (eye diameter/standard length of the fish), ecological parameters (trophic level, preferred temperature, and maximum depth) which were retrieved from FishBase (https://www.fishbase.us), and genome size. No significant correlations were found (supplementary fig. S14, Supplementary Material online).

Discussion

We examined the evolutionary dynamics of the size of the OR gene repertoire in a broad sample of teleost fishes and found a larger range of variation than previously recognized (Niimura 2009; Zhu et al. 2017; Jiang et al. 2019; Lv et al. 2019). Large repertoires do not seem to result from bursts of gene duplication, but we identified several episodic high rates of gene pseudogenization that occurred at the root of clades with species that have very few OR genes. Our analysis indicates that the richness of the OR gene repertoire correlates with the presence of a multilamellar olfactory rosette. We found that some large gene losses coincide with a simplification of the olfactory epithelium and that several cases of secondary complexifications of the olfactory epithelium coincide with secondary expansions of the OR gene repertoire.

OR Functions in Teleost Fishes

In the Zebrafish *Danio rerio*, five types of olfactory sensory neurons (ciliated, microvillus, crypt, kappe, and pear) with distinct morphologies, connections, and functions, are found in the olfactory epithelium. Segregated neural pathways are responsible for coding and processing information from different types of olfactory stimuli, at least at the level of the olfactory bulb (Friedrich and Korsching 1997; Biechl et al. 2017; Calvo-Ochoa and Byrd-Jacobs 2019). Less is known about molecular components and functional organization of the olfactory system in other teleosts, but similar studies in the Channel Catfish *Ictalurus punctatus* (Hansen et al. 2003) and the Goldfish *Carassius auratus* (Hansen et al. 2004) indicate that they may be largely conserved in teleosts. OR genes are expressed in ciliated neurons that recognize bile acids and project to medial glomeruli in the olfactory bulb (Yoshihara 2014). Bile acids are biliary steroids synthesized in the liver, secreted in the intestine and released to the environment. These molecules are used as pheromones in the context of social and reproductive behaviors and convey information about the status of the emitter (Doyle and Meeks 2018). For example, one of the most studied and common bile acids, taurocholic acid, acts as an attractant at very low concentration (10^{-8} M), to which females of *D. rerio* are more sensitive to than males (Michel and Lubomudrov 1995).

We know little about the molecules recognized by the different ORs in teleost fishes because no OR has been deorphanized so far. Although current evidence suggests that OR genes may be expressed in sensory neurons involved in social behaviors (e.g., reproduction, migration, kin

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab145/6265480 by guest on 29 June 2021

recognition, and social status) rather than foraging and feeding, more functional data are needed to describe the odorant response range of teleost ORs.

Diversity and evolution of the OR gene repertoire in teleost fishes

We used the many genomes now available to identify OR gene repertoires in 163 high-quality genomes from species representative of teleost diversity. We evaluated the accuracy of our OR gene mining pipeline by comparing our results with similar analyses performed previously on eight species. In comparison to prior studies, we systematically recovered more OR genes, sometimes finding that the initial estimate greatly underestimated the number of OR genes. We found that some species previously thought to have few functional OR genes, such as *Dichotomyctere nigroviridis* (11 genes) and *Ictalurus punctatus* (27 genes) possess larger OR gene repertoires (41 and 87, respectively) and are not among the fishes with the lowest number of OR genes.

We found an unanticipated large range of variation of OR gene number among teleosts, with an average number of 117 OR genes. The lowest number of OR genes (15) was observed in two distantly related species, the Broad-nose Pipefish *Syngnathus typhle* and the Ocean Sunfish *Mola mola*. In both cases, the few remaining intact ORs primarily belong to the δ and η gene families, which are among the three largest families in teleosts. In both *S. typhle* and *M. mola*, very few pseudogenes were identified, suggesting that contraction of the OR gene repertoire is ancient.

The highest number of OR genes (429) was found in the Zig-zag Eel *Mastacembelus armatus* (Synbranchiformes). This is primarily the result of duplications of genes belonging to δ and ζ families, that are among the three largest OR gene families. The high percentage of pseudogenes indicates high gene turnover in this lineage.

Together, these results suggest that in the most extreme cases of reduction and expansion of the OR gene repertoire, the implication of different OR gene families is not strongly biased.

Within teleosts, gene turnover seems to be driven by slow birth-and-death evolutionary dynamics involving similarly low birth and death rates. However, we identified a few instances of high birth and death rates. The highest birth rate was observed at the level of a deep internal branch (fig. 2). Because it was followed by various death rates in more recent branches, it is not correlated with extant species that have a rich OR gene repertoire. The second highest birth rate occurred in the common ancestor of *Sander* + *Perca* spp., a split that occurred relatively recently, and, in this case, is associated with high numbers of OR genes, 167 in the Yellow Perch *Perca flavescens* and 217 in the Zander *Sander lucioperca*, compared to the teleost average of 117. More striking is the link between the two highest death rates and the two species with the smallest OR gene repertoires, indicating ancient massive gene losses without compensation by gene duplications in some lineages.

An extensive study of the dynamics of OR gene repertoires in mammals found a low correlation ($r = 0.31$) between the number of gene losses inferred using the reconciled-tree method and the number of observed pseudogenes (Niimura et al. 2014). Our analyses suggest an explanation for this counter-intuitive result. Indeed, using all terminal branches, we found a similarly low correlation ($r = 0.17$) for teleosts. However, using only data for short terminal branches, which corresponds to the most recent evolutionary events in independent lineages, the correlation was much higher ($r > 0.7$ when using only branches shorter than 7 Myr). This suggests that the oldest pseudogenes are not always identified because they are too divergent or because they are removed from genomes by large deletions, hampering accurate estimation of the number of pseudogenes in long terminal branches. Analyses at smaller phylogenetic scales showed an expected congruence of estimates of the rate of gene losses based on these independent approaches.

Number of OR Genes and Correlation with Olfactory Epithelium Complexity

We searched for correlations between the size of the OR gene repertoire and morphological traits, ecological parameters and genome size. We did not find any correlation with fish maximum length, relative eye size, maximum depth range, trophic level, temperature preference, or genome size. In contrast, the shape of the olfactory epithelium (absence or presence of a multilamellar olfactory rosette) and, when a multilamellar olfactory rosette is present, the number of its lamellae, were positively correlated with the number of OR genes. These correlations suggest that epithelial folding allows for the presence of more sensory neurons and also a more diversified OR repertoire. Moreover, these correlations indicate a relationship between morphological and genomic changes in accordance with the relative importance of olfaction among the suite of sensory systems.

Case Studies

It is instructive to study cases of extreme OR gene repertoire reduction and subsequent evolution at the molecular and morphological levels.

One case concerns the Broad-nosed Pipefish *Syngnathus typhle* and closely related species. The inference of gene gains and losses indicates a large number of gene losses in the common ancestor of the clade Syngnathiformes + (Kurtiformes + Gobiiformes). If the phylogenetic relationship between Syngnathiformes and (Kurtiformes + Gobiiformes) is correct, then the disappearance of the multilamellar olfactory rosette in Syngnathiformes is closely related with a massive reduction of the OR gene repertoire. In the common ancestor of Syngnathiformes, the number of OR genes also greatly declined. The tiny repertoire of OR genes could be, with a flat olfactory epithelium, a shared character of Syngnathidae (seahorses, pipefishes, and seadragons). Behavioral studies suggest that vision is the leading sense during mating, and that olfaction is less important in *S. typhle* (Berglund and Rosenqvist 2001; Lindqvist et al. 2011). In the clade Gobiiformes + Kurtiformes, a re-expansion of the OR gene repertoire could have occurred at different rates in different lineages, but no data are available to relate these OR

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab145/6265480 by guest on 29 June 2021

gene repertoire expansions with changes in olfactory epithelium shape.

The diversity of olfactory epithelium within Tetraodontiformes presents a potential model to study evolution of the olfactory system and the OR gene repertoires. The last common ancestor of Tetraodontiformes most likely had a multilamellar olfactory rosette that is present in three families, Triacanthodidae, Triacanthidae, and Triodontidae (Tyler 1980; Arcila and Tyler 2017) (see also supplementary fig. S11*E–I*, Supplementary Material online). Repeated losses of the multilamellar olfactory rosette occurred within Tetraodontiformes, in particular in Monacanthidae and Tetraodontidae. In both cases, loss of the multilamellar olfactory rosette is associated with a contraction of the OR gene repertoire to 45 functional OR genes in the monacanthid *Thamnaconus septentrionalis* and 41 functional OR genes in the tetraodontid *Dichotomyctere nigroviridis*. *Takifugu* spp., also tetraodontids, have a multilamellar olfactory rosette, but the series of folds do not have the ancestral rosette shape, suggesting that it might be a secondary re-acquisition. We infer an expansion of the OR gene repertoire in the common ancestor of *Takifugu* because the functional OR repertoire size ranges from 79 in *Takifugu flavidus* to 100 in *Takifugu rubripes*. Finally, the small number of functional OR genes (15) in the Ocean Sunfish *Mola mola* corresponds with their anatomy. *Mola mola* has minute nostrils almost flush with the surface of the skin (Tyler 1980). We examined an adult specimen of *M. mola* (129 cm total length) and found that the olfactory epithelium is flat, except for a small V-shaped primary lamella on the surface below the incurrent nare (KEB unpublished data). The brain of *M. mola* has greatly reduced olfactory nerves as well as reduced olfactory bulbs (Burr 1928). Together this morphological data and the small number of OR genes suggest that olfaction is particularly regressed in this species.

We observed a 2.5-fold difference in OR number between the Zebrafish *Danio rerio* and the related cyprinid *Danionella dracula*. The miniaturized and transparent species belonging to *Danionella* have recently attracted the interest of neuroscientists because of the advantages for functional imaging, and of developmental biologists as a model for understanding the regulation of size and developmental heterochronies and truncations (Schulze et al. 2018; Conway et al. 2021). *Danionella dracula* and *D. translucida* have small, skull-less brains but rich behavioral repertoires, including acoustic communication. A reduction in social olfactory communication associated with OR gene losses could be related or compensated for by the emergence of a sound-producing apparatus and acoustic communications skills in these species. Such sensory trade-offs could also explain the relatively small OR gene repertoire of electric fishes *Electrophorus electricus* and *Paramormyrops kingsleyae* compared to their closest relatives, because their social interactions may primarily involve electric communication (Gracheva and Bagriantsev 2018). However, additional analyses are required to investigate the importance of sensory trade-offs in the evolution of the size of OR gene repertoires in teleosts.

Perspectives

What kinds of ecological transitions drove in parallel the loss or reappearance of a multilamellar olfactory rosette and contraction or re-expansion of the OR gene repertoire in different teleost lineages? How can Sea-Horses and Ocean Sunfish thrive with such poor OR repertoires and why do several species of *Takifugu* have a multilamellar olfactory rosette and a large OR gene repertoire relative to other Tetraodontiformes that have a flat olfactory epithelium and a limited OR gene repertoire? Further studies examining social communication, ecology, life history traits, and sensory physiology are needed to answer these fascinating questions.

New high-quality genome sequences are becoming available at an ever-increasing rate, but we found that anatomical data on the olfactory organ is rare and that refined descriptions of the olfactory epithelium shape and organization are available only for a handful of model species. In particular, it would be very interesting to analyze the evolution of the number, distribution, and density of olfactory sensory neurons in relation with the olfactory epithelium shape and the number of OR genes. Future large-scale comparative analyses should include such data for a large sample of teleost species. At the molecular level, an interesting complementary analysis would concern the TAAR gene repertoire. It will be particularly important to examine if the correlation we observed between OR repertoire size and olfactory epithelium shape also holds for the TAAR repertoire size.

Materials and Methods

Assessment of Genome Completeness and OR Gene Mining

The genomes of 307 teleost species and those of two outgroups (*Acipenser ruthenus* and *Lepisosteus oculatus*) were downloaded from NCBI and ENSEMBL databases on 04/10/2020. The names of the species and the genome assembly versions can be found in supplementary Data S1, Supplementary Material online.

For each species, the genome completeness was assessed with BUSCO 3.1.0 (Waterhouse et al. 2018) using the Actinopterygii odb9 database and the expected genome size extracted from the Animal Genome Size Database (Gregory 2020).

We developed a pipeline (https://github.com/MaximePolicarpo/Olfactory_receptor_genes) to mine olfactory receptor genes in each genome following the procedure of Niimura and Nei (2005). Briefly, tblastn (Ye et al. 2006) was used with a set of query sequences that are teleost OR genes identified by Niimura (2009) and an e-value threshold of 1e−10 to find genomic regions containing OR genes. Nonoverlapping best-hit regions were extracted using samtools (Li et al. 2009). These sequences were extended 1,000 bp upstream and 1,000 bp downstream and open reading frames (ORFs) larger than 750 bp were extracted using EMBOSS getorf (Rice et al. 2000). Only ORFs that best matched to a known olfactory receptor against the UniProt database (The UniProt Consortium 2019) were retained. These ORFs were aligned with the teleost OR gene set identified by Niimura

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab145/6265480 by guest on 29 June 2021

(2009) and several outgroup sequences (ora1: A0A0R4IM31.1, taar13: XP_021328050.1, adra2a: NP_919345.2, adrb1: AAI62819.1, galr2b: XP_001339169.1, hrh1: XP_009304002.1 and htr1aa: NP_001116793.1) using MAFFT v7.310 (Katoh et al. 2005), and a maximum likelihood tree was computed with IQ-TREE (Nguyen et al. 2015). We only kept genes that clustered with known OR genes by visual inspection of trees and removed 100% identical nucleotide sequences using CD-HIT (Fu et al. 2012). Retained sequences were classified as functional ORs.

This set of functional ORs was then used for a new round of tblastn with a more stringent e-value of 1e−20. For each genome, best-hit regions were extracted and discarded if overlapping with a functional OR gene region. Otherwise, the best tblastn match was extracted and sequences were classified as follows: 1) "edge" if the blast target sequence was less than 30 bp away from the scaffold end; 2) "pseudogene" if the blast target sequence contained at least a stop codon or frameshift; 3) "truncated" if the sequence was incomplete but without internal stop codons or frameshifts.

Functional, Truncated and Edge genes were aligned with the teleost OR gene set identified by Niimura (2009) using MAFFT and classified into OR gene families based on their position on a maximum likelihood tree computed with IQ-TREE. Pseudogenes were classified based on their best blastx match.

Phylogenetic Framework and Morphological/Ecological Trait Assembly

A time-calibrated phylogeny of ray-finned fishes was downloaded from https://fishtreeoflife.org (Rabosky et al. 2018). The tree was pruned using the R package ape (Paradis and Schliep 2019) to keep only the 165 species whose genomes we used in our study. Ecological information on each species was extracted from FishBase (https://www.fishbase.se) using the R package rfishbase (supplementary Data S1, Supplementary Material online).

We gathered information about the olfactory epithelium shape and the number of lamellae per olfactory rosette in 220 actinopterygians from an extensive literature survey (supplementary Data S1, Supplementary Material online). Among these species, genomic data were available for 41 species; for an additional 38 species genomic data were available for at least one species belonging to the same genus. For each species, the olfactory epithelium was classified as multilamellar olfactory rosette present (3 or more lamellae), or multilamellar olfactory rosette absent if the olfactory epithelium was flat (0 lamellae) or only a few lamellae (1 or 2 lamellae) were present. Specimens referenced in the literature were adults. We did not take into account sexual dimorphism or individual age, which have been shown to impact number of olfactory lamellae (Kasumyan 2004; Abrahão et al. 2019).

The time-calibrated "fishtreeoflife" was pruned again to retain 205 of the 220 species for which information about the olfactory epithelium was available (15 species were not in the "fishtreeoflife"). The evolution of the olfactory epithelium shape was inferred by maximum likelihood (MPPA with F81 model) and parsimony (with DOWNPASS, DELTRAN, or ACCTRAN methods) using PastML (Ishikawa et al. 2019). Trees were rooted and visualized using iTOL (Letunic and Bork 2007). Species names were validated using Eschmeyer's Catalog of Fishes (Fricke et al. 2021) (supplementary Data S1, Supplementary Material online).

Small Scale OR Genes Phylogenies

To compare the results of our implementation of a method for identifying OR genes (Niimura 2009) with the results obtained in previous studies, sequences of OR genes from eight teleost species (*Danio rerio*, *Dichotomyctere nigroviridis*, *Oryzias latipes*, *Gasterosteus aculeatus*, *Takifugu rubripes*, *Siniperca chuatsi*, *Pseudoliparis swirei*, and *Ictalurus punctatus*) were retrieved from previous studies (Niimura 2009; Gao et al. 2017; Jiang et al. 2019; Lv et al. 2019). The sequences were aligned with those found in the present study using MAFFT (on amino acids) and a ML tree was computed with IQ-TREE. The optimal model was inferred with ModelFinder, while the robustness of the nodes was evaluated with 1,000 ultrafast bootstraps. Trees were visualized, rooted, and annotated using iTOL.

To obtain an overall picture of the diversity of OR gene families in teleosts, we also inferred a phylogenetic tree of all functional OR sequences from 38 species, each belonging to a different order (species and order names can be found in supplementary Data S2, Supplementary Material online). Protein sequences were aligned using MAFFT and a near-ML tree was computed with FastTree (Price et al. 2010), and the robustness of the nodes was evaluated with 1,000 fast-global bootstraps.

OR Birth and Death Rates

We inferred the number of gene gains and losses along each branch of the species tree using a gene tree—species tree reconciliation method. For each OR gene family, an alignment was performed with MAFFT and a maximum likelihood tree was computed with IQ-TREE. Nodes with low bootstrap values (<90%) were collapsed into polytomies using the R package ape. We then used NOTUNG 2.9.1.5 (Chen et al. 2000) to root and reconcile gene trees with the species tree using the phylogenomics option.

Birth and death rates along each branch as well as the mean birth and death rate were computed using the equations of Niimura et al. (2014). We excluded branches with lengths < 2 million years because differences in gene retrieval and genome qualities had a major impact on inferred birth and death rates (supplementary Data S1, Supplementary Material online).

Phylogenetic Comparative Analyses

The phylogenetic signal (Pagel's λ) of each trait was estimated with the function phylosig in the R package phytools with the option test = TRUE (Revell 2012).

The R package phylolm (Tung Ho and Ané 2014) was used to perform phylogenetic linear regressions between continuous traits using the phylolm function with various evolutionary models, and phylogenetic logistic regression between binary and continuous traits using the phyloglm function.

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab145/6265480 by guest on 29 June 2021

Correlation between the number of gene losses and the number of pseudogenes in terminal branches

For each terminal branch of the species tree, the number of gene losses was inferred using the gene family trees of intact OR genes and the program NOTUNG 2.9.1.5 (Chen et al. 2000) as described above. We also estimated the numbers of OR pseudogenes (OR sequences with at least one LoF mutations), truncated genes (most likely pseudogenes resulting from large genomic deletions) and edge genes (truncated genes most likely resulting from assembly artifacts).

The correlation (Pearson's r) between the number of gene losses and the number of pseudogenes and truncated genes, together or separately, and the number of edge genes was computed for all terminal branches or using only a subset of branches shorter than a given threshold ranging from 2 Myr to 200 Myr.

To estimate the time necessary for a pseudogene to become undetectable due to the accumulation of substitutions and small indels, we ran simulations in which an intact OR sequence accumulated nucleotide substitutions (at a rate of 10-8 per generation) and indels at a rate of 0.05×10^{-8} per generation as estimated previously (Policarpo et al. 2021). We also took into account the transition/transversion ratio and the relative frequency of indels ranging from 1 and 9 bp in length (more details on the numerical values used for the simulations are given in the Python script available at https://github.com/MaximePolicarpo/Olfactory_receptor_genes).

Each time a new mutation was produced in the evolving sequence, a tblastn analysis was performed using the evolving DNA sequence and the translated original OR sequence. The evolving sequence was classified as unrecognized if the e-value was greater than 10^{-20}, which was the threshold used to mine OR pseudogenes, and the number of generations to get an unrecognized sequence registered. Running this simulation 10,000 times allowed the estimate of the distribution of the number of generations necessary to make a pseudogene undetectable.

Data Availability

All data and results of the analyses performed in this study are available for download at figshare (https://figshare.com/articles/dataset/Supplementary_files_-_ORs_in_teleost/14180090): nucleotide sequences of the olfactory receptor genes in three FASTA format files (functional genes, incomplete genes, and pseudogenes); version of genomes used for this study, assembly sizes, expected genome sizes, BUSCO assessment results, olfactory receptor gene count, olfactory epithelium shape, ecological data, and NOTUNG results in Supplementary_Table.xlsx; BUSCO assessment results in graphic format and phylogenetic trees (Newick format).

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank Matthieu Haudiquet for his help with the R package phylolm.

References

Abrahão VP, Pastana M, Marinho M. 2019. On a remarkable sexual dimorphic trait in the Characiformes related to the olfactory organ and description of a new miniature species of *Tyttobrycon Géry* (Characiformes: characidae). *PLoS One* 14(12):e0226130.

Arcila D, Tyler JC. 2017. Mass extinction in tetraodontiform fishes linked to the Palaeocene–Eocene thermal maximum. *Proc R Soc B.* 284(1866):20171771.

Berglund A, Rosenqvist G. 2001. Male pipefish prefer ornamented females. *Anim Behav.* 61(2):345–350.

Biechl D, Tietje K, Ryu S, Grothe B, Gerlach G, Wullimann MF. 2017. Identification of accessory olfactory system and medial amygdala in the zebrafish. *Sci Rep.* 7:44295.

Bird DJ, Murphy WJ, Fox-Rosales L, Hamid I, Eagle RA, Van Valkenburgh B. 2018. Olfaction written in bone: cribriform plate size parallels olfactory receptor gene repertoires in Mammalia. *Proc R Soc B.* 285(1874):20180100.

Buck L, Axel R. 1991. A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell* 65(1):175–187.

Burne RH. 1909. The anatomy of the olfactory organ of teleostean fishes. *Proc Zool Soc Lond.* 2:610–663.

Burr HS. 1928. The central nervous system of *Orthagoriscus mola*. *J Comp Neurol.* 45(1):33–128.

Calvo-Ochoa E, Byrd-Jacobs CA. 2019. The olfactory system of zebrafish as a model for the study of neurotoxicity and injury: implications for neuroplasticity and disease. *Int J Mol Sci.* 20(7):1639.

Chen K, Durand D, Farach-Colton M. 2000. NOTUNG: a program for dating gene duplications and optimizing gene family trees. *J Comput Biol.* 7(3-4):429–447.

Conway KW, Kubicek, KM Britz, R. 2021. Extreme evolutionary shifts in developmental timing establish the miniature *Danionella* as a novel model in the neurosciences. *Dev Dyn.* 250(4):601–611.

Doyle WI, Meeks JP. 2018. Excreted steroids in vertebrate social communication. *J Neurosci.* 38(14):3377–3387.

Dymek J, Rosenqwist G, Kuciel M, Lauriano ER, Capillo G, Zaccone G, Żuwała K. 2021. Micro- and macro-morphology of the olfactory organ of *Syngnathus typhle* (Syngnathidae, Actinopterygii). *Acta Zool.* 102(2):206–214.

Fricke R, Eschmeyer WN, Van der Laan R. 2021. Eschmeyer's Catalog of Fishes: Genera, species, references. [cited 2021 Feb 19] Available from: http://researcharchive.calacademy.org/research/ichthyology/catalog/fishcatmain.asp.

Friedrich RW, Korsching SI. 1997. Combinatorial and chemotopic odorant coding in the Zebrafish olfactory bulb visualized by optical imaging. *Neuron* 18(5):737–752.

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28(23):3150–3152.

Gao S, Liu S, Yao J, Li N, Yuan Z, Zhou T, Li Q, Liu Z. 2017. Genomic organization and evolution of olfactory receptors and trace amine-associated receptors in Channel Catfish, *Ictalurus punctatus*. *Biochim Biophys Acta Gen Subj.* 1861(3):644–651.

Gracheva EO, Bagriantsev SN. 2018. Communication: potassium channels define the dialect. *Curr Biol.* 28(13):R744–R746.

Gregory TR. 2020. Animal Genome Size Database. Available from: http://www.genomesize.com.

Hansen A, Anderson KT, Finger TE. 2004. Differential distribution of olfactory receptor neurons in Goldfish: structural and molecular correlates. *J Comp Neurol.* 477(4):347–359.

Hansen A, Rolen SH, Anderson K, Morita Y, Caprio J, Finger TE. 2003. Correlation between olfactory receptor cell type and function in the Channel Catfish. *J Neurosci.* 23(28):9328–9339.

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab145/6265480 by guest on 29 June 2021

Hansen A, Zielinski BS. 2005. Diversity in the olfactory epithelium of bony fishes: development, lamellar arrangement, sensory neuron cell types and transduction components. *J Neurocytol*. 34(3-5):183–208.
Hara TJ. 1975. Olfaction in fish. *Prog Neurobiol*. 5(4):271–335.
Hughes GM, Boston ESM, Finarelli JA, Murphy WJ, Higgins DG, Teeling EC. 2018. The birth and death of olfactory receptor gene families in mammalian niche adaptation. *Mol Biol Evol*. 35(6):1390–1406.
Ishikawa SA, Zhukova A, Iwasaki W, Gascuel O. 2019. A fast likelihood method to reconstruct and visualize ancestral scenarios. *Mol Biol Evol*. 36(9):2069–2085.
Jiang H, Du K, Gan X, Yang L, He S. 2019. Massive loss of Olfactory Receptors but not Trace Amine-Associated Receptors in the world's deepest-living fish (*Pseudoliparis swirei*). *Genes* 10(11):910.
Kasumyan A. 2004. The olfactory system in fish: structure, function, and role in behavior. *J Ichthyol*. 44:S180–S223.
Katoh K, Kuma K-I, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res*. 33(2):511–518.
Khan I, Yang Z, Maldonado E, Li C, Zhang G, Gilbert MTP, Jarvis ED, O'Brien SJ, Johnson WE, Antunes A. 2015. Olfactory receptor subgenomes linked with broad ecological adaptations in Sauropsida. *Mol Biol Evol*. 32(11):2832–2843.
Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: a resource for timelines, timetrees, and divergence times. *Mol Biol Evol*. 34(7):1812–1819.
Letunic I, Bork P. 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* 23(1):127–128.
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079.
Lindqvist C, Sundin J, Berglund A, Rosenqvist G. 2011. Male broad-nosed Pipefish *Syngnathus typhle* do not locate females by smell. *J Fish Biol*. 78(6):1861–1867.
Liu A, He F, Shen L, Liu R, Wang Z, Zhou J. 2019. Convergent degeneration of olfactory receptor gene repertoires in marine mammals. *BMC Genomics*. 20(1):977.
Lv L-Y, Liang X-F, He S. 2019. Genome-Wide Identification and Characterization of Olfactory Receptor Genes in Chinese Perch, *Siniperca chuatsi*. *Genes*. 10(2):178.
Michel WC, Lubomudrov LM. 1995. Specificity and sensitivity of the olfactory organ of the Zebrafish, *Danio rerio*. *J Comp Physiol A*. 177(2):191–199.
Musilova Z, Cortesi F, Matschiner M, Davies WIL, Patel JS, Stieb SM, de Busserolles F, Malmstrøm M, Tørresen OK, Brown CJ, et al. 2019. Vision using multiple distinct rod opsins in deep-sea fishes. *Science* 364(6440):588–592.
Nei M, Niimura Y, Nozawa M. 2008. The evolution of animal chemosensory receptor gene repertoires: roles of chance and necessity. *Nat Rev Genet*. 9(12):951–963.
Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*. 32(1):268–274.
Niimura Y. 2009. On the origin and evolution of vertebrate olfactory receptor genes: comparative genome analysis among 23 Chordate species. *Genome Biol Evol*. 1:34–44.
Niimura Y, Matsui A, Touhara K. 2018. Acceleration of olfactory receptor gene loss in primate evolution: possible link to anatomical change in sensory systems and dietary transition. *Mol Biol Evol*. 35(6):1437–1450.
Niimura Y, Matsui A, Touhara K. 2014. Extreme expansion of the olfactory receptor gene repertoire in African elephants and evolutionary dynamics of orthologous gene groups in 13 placental mammals. *Genome Res*. 24(9):1485–1496.
Niimura Y, Nei M. 2005. Evolutionary dynamics of olfactory receptor genes in fishes and tetrapods. *Proc Natl Acad Sci U S A*. 102(17):6039–6044.
Paradis E, Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35(3):526–528.
Policarpo M, Fumey J, Lafargeas P, Naquin D, Thermes C, Naville M, Dechaud C, Volff J-N, Cabau C, Klopp C, et al. 2021. Contrasting gene decay in subterranean vertebrates: insights from cavefishes and fossorial mammals. *Mol Biol Evol*. 38(2):589–605.
Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS One* 5(3):e9490.
Rabosky DL, Chang J, Title PO, Cowman PF, Sallan L, Friedman M, Kaschner K, Garilao C, Near TJ, Coll M, et al. 2018. An inverse latitudinal gradient in speciation rate for marine fishes. *Nature* 559(7714):392–395.
Revell LJ. 2012. phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol Evol*. 3(2):217–223.
Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*. 16(6):276–277.
Sato Y, Miyasaka N, Yoshihara Y. 2007. Hierarchical regulation of odorant receptor gene choice and subsequent axonal projection of olfactory sensory neurons in Zebrafish. *J Neurosci*. 27(7):1606–1615.
Schulze L, Henninger J, Kadobianskyi M, Chaigne T, Faustino AI, Hakiy N, Albadri S, Schuelke M, Maler L, Del Bene F, et al. 2018. Transparent *Danionella translucida* as a genetically tractable vertebrate brain model. *Nat Methods*. 15(11):977–983.
Sharma K, Syed AS, Ferrando S, Mazan S, Korsching SI. 2019. The chemosensory receptor Repertoire of a true shark is dominated by a single olfactory receptor family. *Genome Biol Evol*. 11(2):398–405.
The UniProt Consortium. 2019. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res*. 47:D506–D515.
Tung Ho LS, Ané C. 2014. A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. *Syst Biol*. 63(3):397–408.
Tyler JC. 1980. Osteology, phylogeny, and higher classification of the fishes of the order Plectognathi (Tetraodontiformes). U.S. Department of Commerce. NOAA Technical Report, NMFS Circular, vol. 434. p. 1–422.
Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM. 2018. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol Biol Evol*. 35(3):543–548.
Ye J, McGinnis S, Madden TL. 2006. BLAST: improvements for better sequence analysis. *Nucleic Acids Res*. 34(Web Server issue):W6–W9.
Yoshihara Y. 2014. Zebrafish olfactory system. In: Mori K, editor. The olfactory system: from odor molecules to motivational behaviors. Tokyo: Springer Japan. p. 71–96.
Zhu G, Wang L, Tang W, Wang X, Wang C. 2017. Identification of olfactory receptor genes in the Japanese Grenadier Anchovy *Coilia nasus*. *Genes Genomics*. 39(5):521–532.

Downloaded from https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msab145/6265480 by guest on 29 June 2021

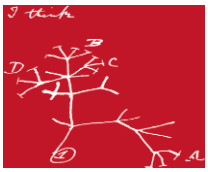
B- Articles publiés en coauteur

1) Pourquoi le taux de mutation n'est-il jamais égal à zéro ?

médecine/sciences 2019 ; 35 : 245-51

Pourquoi le taux de mutation n'est-il jamais égal à zéro ?

Didier Casane[1,2], Maxime Policarpo[1,2], Patrick Laurenti[2,3]



[1]Laboratoire Évolution, Génomes, Comportement, Écologie, UMR9191, CNRS, IRD, Univ Paris-Sud, Université Paris-Saclay, 91198 Gif-sur-Yvette, France.
[2]Université Paris-Diderot, UFR des sciences du vivant, Sorbonne Paris Cité, Paris, France.
[3]Laboratoire Interdisciplinaire des Énergies de Demain, UMR 8236 CNRS, Bâtiment Lamarck, 35, rue Hélène Brion 75013 Paris, France.
patrick.laurenti@univ-paris-diderot.fr

> Alfred H. Sturtevant fut le premier à s'en étonner : le taux de mutation est faible mais n'atteint jamais zéro. Pourtant, la plupart des mutations qui modifient le phénotype ont un effet délétère, les individus qui produisent le moins de mutants génèrent donc plus de descendants viables et fertiles. La sélection naturelle devrait ainsi progressivement faire tendre le taux de mutation vers zéro au cours des générations. Des analyses récentes suggèrent que ce taux dépend principalement de la taille efficace des génomes et de l'effectif efficace des populations. Le maintien de taux de mutation plus élevés que nécessaire illustrerait les limites de la sélection naturelle dans un monde vivant constitué de populations de taille finie. <

« *Why does the mutation rate not become reduced to zero? No answer seems possible at present, other than the surmise that the nature of genes does not permit such a reduction. In short, mutations are accidents, and accidents will happen* ».

Alfred Henry Sturtevant (1937)

« *There seems to be enough evidence to show that mutation rates are under genetic control. Thus, the mutation rate characteristic of each species must be a product of past evolution like any other morphological or physiological character* ».

Motoo Kimura (1967)

« *A potentially revealing hypothesis for mutation-rate evolution is that natural selection primarily operates to improve replication fidelity, with the ultimate limits to what can be achieved set by the power of random genetic drift. This drift-barrier hypothesis is consistent with comparative measures of mutation rates, provides a simple explanation for the existence of error-prone polymerases and yields a formal counter-argument to the view that selection fine-tunes gene-specific mutation rates* ».

Michael Lynch (2016)

En 1937, quand Alfred Henry Sturtevant[1] pose cette question simple et pourtant extraordinairement importante : « Pourquoi le taux de mutation n'atteint-il jamais zéro ? » [1], la génétique, en particulier grâce à l'utilisation intensive des drosophiles comme modèle d'étude, a fait des progrès considérables depuis sa fondation quelques décennies auparavant. Quelques faits sont déjà bien établis : des individus mutants apparaissent de façon récurrente dans les populations, comme des mouches aux yeux blancs dans une population de mouches avec les yeux rouges ; ces individus mutants sont les descendants de parents porteurs de modifications (des mutations) dans des cellules de leur lignée germinale ; enfin, la plupart des mutations qui modifient le phénotype ont des effets délétères car ces mutants ont souvent une viabilité et/ou une fécondité réduite, voire nulle (si la mutation est létale ou rend le porteur stérile). Il est par ailleurs établi que le nombre de mutations qui se produisent par génération, ou taux de mutation, est variable dans les différentes lignées de drosophiles [2-4]. Le taux de mutation étant un caractère héréditaire et variable, il peut donc évoluer. Dans ce cadre, un taux de mutation réduit représente un avantage qui devrait conduire à la sélection des individus *antimutateurs*. En effet, les individus qui produisent le moins de descendants mutants produiront plus de descendants viables et fertiles à la génération suivante. Notons que la sélection ne se fait pas sur le phénotype de l'individu mutateur mais sur le phénotype de ses descendants, une forme de sélection appelée parfois sélection de second ordre. Progressivement, la sélection récurrente des individus les moins mutateurs, qui apparaissent au cours du temps, devrait faire tendre le taux de mutation vers zéro. Comme l'existence des mutations, à l'origine de la diversité génétique, est la condition *sine qua non* de l'évolution des espèces, l'évolution devrait donc, en théorie, s'arrêter faute de matière première ; or ce n'est manifestement pas le cas. Sturtevant en conclut que la nature

[1] Alfred Henry Sturtevant (1891-1970) était un généticien américain drosophiliste.

m/s n° 3, vol. 35, mars 2019
https://doi.org/10.1051/medsci/2019038

des gènes, inconnue en 1937, « ne permet pas une telle réduction » et que « les mutations sont des accidents et que les accidents finissent toujours par survenir » [1]. Pour mieux comprendre le contrôle du flux de mutations et son évolution au cours du temps, il faudra attendre le dernier tiers du XXe siècle. En effet, jusque-là, les mutations n'étaient qu'indirectement observables par le biais de leurs effets phénotypiques associés : modifications de couleur, de forme, de viabilité, de fertilité. Les mutations n'étaient donc définies, au mieux, que par une association entre une position sur un chromosome et un phénotype. Avec la découverte de méthodes de séquençage de l'ADN, il devient possible d'observer des différences entre génomes indépendamment de tout effet observable sur les organismes. L'accroissement récent des capacités de séquençage rend possible la quantification des différences à l'échelle des génomes entiers. Toutefois, si les mutations à petite échelle, comme le remplacement d'un nucléotide par un autre, une insertion ou une délétion de quelques nucléotides, sont facilement identifiables, les remaniements à grande échelle restent difficiles à étudier. C'est pour cela que les analyses du taux de mutation et de son évolution sont aujourd'hui essentiellement fondées sur l'identification de mutations ponctuelles (changement d'un nucléotide en un autre ou polymorphisme nucléotidique – SNP en anglais). Dans la suite de cet article, nous utiliserons le mot mutation dans le sens étroit de mutation ponctuelle car la plupart des travaux sont focalisés sur ce type de mutations, mais cela ne devrait pas modifier significativement la validité et la portée générale des mécanismes évolutifs mis en évidence.

Le nombre de mutations qui se produisent par unité de temps dépend en particulier des complexes protéiques impliqués dans la réplication et la réparation de l'ADN. Des mutations peuvent se produire dans les gènes qui codent ces protéines, il peut donc exister une variabilité génétique interindividuelle des mécanismes de réplication et de réparation de l'ADN, c'est-à-dire une variation interindividuelle du taux de mutation. Dans les cas les plus extrêmes, des mutations dans les gènes de réparation de l'ADN sont responsables de maladies génétiques extrêmement graves, comme le *xeroderma pigmentosum*[2]. Ces mutations ne sont pas maintenues au cours des générations car leurs porteurs ont une espérance de vie et des capacités de reproduction réduites. En théorie, à l'échelle de l'évolution, seules des mutations de ces gènes avec des effets plus réduits sur le taux de mutation des autres gènes devraient se maintenir dans les populations, mais nous verrons plus loin qu'il peut exister des exceptions à cette règle.

Les taux de mutation mesurés chez des organismes aussi divers que des virus à ARN, des virus à ADN, des bactéries, des eucaryotes unicellulaires et pluricellulaires sont relativement bas, mais ne sont jamais égaux à zéro et varient dans une gamme extrêmement étendue : de 10^{-3} à 10^{-11} mutations par site par réplication [3-7]. En d'autres termes, la probabilité qu'un nucléotide soit remplacé par un autre peut être proche de 1 sur mille chez des virus à ARN [4], mais être proche de 1 sur cent milliards chez des paramécies [8]. Chez l'homme, le taux de mutation par génération est d'environ 10^{-8}, soit moins de 10^{-10} par réplication [9].

Les mutations qui n'ont pas d'effets sur le nombre de descendants sont invisibles du point de vue de la sélection naturelle : ce sont des mutations neutres. Les effets des mutations ponctuelles ont été caractérisés à de multiples reprises chez des organismes aussi différents que des virus et l'homme, en passant par des bactéries et des levures [10]. Toutes ces analyses montrent que la plupart des mutations qui apparaissent sont neutres, qu'une petite fraction est délétère et que les mutations avantageuses sont extrêmement rares. Très récemment, cette observation classique a pu être analysée très finement grâce à une technologie très sophistiquée qui permet l'analyse d'un très grand nombre de mutants chez la bactérie *Escherichia coli* [11]. Les mutations qui modifient le phénotype ont un effet le plus souvent légèrement délétère, beaucoup moins souvent franchement délétère, rarement létal et quasi jamais avantageux. Cette observation, qui peut paraître contre-intuitive, se comprend facilement si l'on considère que les systèmes génétiques sont les résultats de processus d'optimisation qui se sont déroulés pendant des millions, voire parfois des milliards d'années, et qu'il est donc aujourd'hui beaucoup plus facile de les détériorer que de les améliorer [12] (➜).

(➜) Voir le Forum de D. Casane et P. Laurenti, *m/s* n° 12, décembre 2014, page 1177

En 1967, Motoo Kimura, un des principaux théoriciens de la biologie évolutive, suggère qu'il doit y avoir une forte sélection contre les mutations qui augmentent le taux de mutation. Par ailleurs, il doit aussi exister une sélection contre un taux de mutation trop bas, car dans un environnement à long terme instable, il est nécessaire qu'il existe une diversité génétique (polymorphisme génétique) pour que l'adaptation puisse se réaliser. Sans ce polymorphisme, l'extinction devient inévitable. La baisse du taux de mutation à petite échelle évolutive impliquerait une sélection interindividuelle (intragroupe), alors que son maintien serait favorisé à long terme par une sélection intergroupe (entre populations ou espèces), de façon analogue aux mécanismes qu'on suppose à l'œuvre pour le maintien de la reproduction sexuée qui présente aussi un désavantage à court terme à l'échelle des individus, mais un avantage à long terme à l'échelle des populations [13]. Le taux de mutation observé correspondrait à un état d'équilibre entre deux forces de sélection qui agissent à des niveaux d'organisation différents. Le point d'équilibre dépend de l'environnement dans lequel vit une espèce. Si cet environnement est très stable, la pression pour maintenir de la variabilité sera faible et le taux de mutation tendra vers une valeur très basse sous l'effet de la sélection interindividuelle, alors que pour une espèce vivant

[2] Une sensibilité excessive de la peau aux rayons ultraviolets, des troubles oculaires et un risque fortement accru de développer un cancer de la peau ou des yeux.

dans un environnement très instable, un taux de mutation élevé sera maintenu. Si Kimura ne doute pas de l'existence de la sélection contre les allèles mutateurs qui augmente le taux de mutation des génomes, il doute fortement de l'existence d'une sélection contre les allèles antimutateurs qui mène vers un taux de mutation trop bas. Il propose donc une autre explication aux taux de mutation observés : ils correspondent aux taux de mutation les plus bas possibles, en tenant compte du fait qu'il y a peut-être une limite biochimique à l'amélioration des systèmes de réplication et de réparation qui ne peut être franchie. On peut aussi supposer un coût, en terme énergétique par exemple, de la réduction du taux de mutation et que ce coût peut devenir trop important pour pouvoir être supporté par un organisme. Au début des années 1970, il n'était pas possible de tester toutes ces hypothèses.

Depuis une dizaine d'années, Michael Lynch[3] propose de tenir compte d'un paramètre négligé jusqu'alors : la taille finie des populations d'organismes vivants. En effet, les fréquences des allèles varient au cours du temps du fait de leur valeur sélective, mais aussi du seul fait du hasard de l'échantillonnage des gamètes à l'origine des individus qui constituent une nouvelle génération. Par exemple, un allèle peut être présent avec une fréquence de 0,5 à une génération donnée, puis être présent à une fréquence de 0,6 à la génération suivante, du seul fait de ce processus d'échantillonnage. Cette variation aléatoire des fréquences, qu'on appelle la dérive génétique, sera d'autant plus grande que la population est de petite taille. Dans une très petite population, un allèle peut ainsi se fixer ou disparaître en quelques générations sans qu'il ait un effet sur la *fitness*[4]. Dans les grandes populations, ces fluctuations stochastiques sont de faibles amplitudes et le polymorphisme génétique peut se maintenir pendant un grand nombre de générations. Les généticiens s'intéressent ainsi à ce qu'ils appellent l'effectif efficace de la population (N_e) plutôt qu'au nombre d'individus qui la constituent. N_e peut être considéré, en première approximation, comme le nombre d'individus qui se reproduisent à chaque génération. L'effectif efficace est donc souvent beaucoup plus petit que l'effectif réel d'individus. Une forte variance du nombre de descendants parmi les individus qui se reproduisent et de fortes fluctuations de la taille de la population au cours du temps réduisent aussi beaucoup la valeur de N_e. Pour fixer les idées, on estime à 10 000 l'ordre de grandeur de l'effectif efficace à long terme de l'espèce humaine dans son ensemble. L'effectif efficace est un paramètre difficile à estimer, et on utilise souvent le polymorphisme génétique neutre pour l'évaluer. En effet, les grandes variations de fréquences alléliques dans les petites populations conduisent à une élimination plus rapide du polymorphisme. En conséquence, s'il y a peu de polymorphisme, on estimera que N_e est petit. À l'inverse, s'il y a beaucoup de polymorphisme, on estimera N_e grand. Pour un allèle soumis à sélection, s'il y a peu de dérive génétique (N_e grand), il se fixera (sa fréquence atteindra 1) presque sûrement s'il est avantageux, et il sera presque sûrement éliminé s'il est délétère. Mais ce même allèle peut se fixer alors qu'il est délétère ou être perdu alors qu'il est avantageux, s'il y a de fortes fluctuations stochastiques des fréquences alléliques (N_e petit) qui perturbent les effets de la sélection directionnelle.

En résumé, les effets de la sélection naturelle peuvent être réduits à néant par la dérive génétique d'autant plus que cette dernière est forte et que l'avantage (ou désavantage) sélectif est faible (pour une explication plus détaillée voir [13], et une analogie footballistique présentée dans l'*Encadré* « le football évolutionnaire »). La dérive génétique est donc une barrière à la sélection naturelle quand cette dernière est de faible ampleur. Lynch propose que c'est cette barrière qui s'oppose à la réduction continue vers zéro du taux de mutation et il rejette l'hypothèse que le taux de mutation soit le résultat de la sélection pour une valeur optimale. Il propose également que les espèces qui ont le taux de mutation le plus faible sont celles qui ont les effectifs efficaces les plus grands et donc chez lesquelles la sélection pour un taux de mutation bas est la plus efficace [14].

Nous allons maintenant examiner un ensemble de résultats récents et très divers qui permettent de tester quelques hypothèses et des attendus associés à ces différents modèles pour expliquer l'existence des mutations et l'extrême variabilité de leur fréquence d'apparition chez différents organismes.

Le taux de mutation a-t-il atteint la plus basse valeur possible ?

Question à laquelle il est très difficile de répondre pour les organismes multicellulaires, mais qui est plus facilement testable chez des organismes unicellulaires en faisant des expériences de mutagenèse pour générer des allèles antimutateurs qui réduisent le taux de mutation. C'est ainsi qu'il est possible d'abaisser le taux de mutation chez la bactérie *E. coli* pourtant à l'origine déjà très bas ($\sim 10^{-10}$) [15, 16]. Bien qu'il soit souhaitable de tester d'autres espèces, ce résultat suggère néanmoins que le taux de mutation observé dans les populations naturelles n'est pas le plus bas possible.

Le coût de la fidélité limite-t-il la baisse du taux de mutation ?

Question à nouveau difficile à aborder expérimentalement, mais qui a pu être testée chez un virus à ARN [17]. Les virus à ARN sont connus pour avoir un taux de mutation extrêmement élevé, mais ce n'est pas toujours le cas [17]. Dans le cadre du modèle d'équilibre entre le fardeau de mutations (les nouvelles mutations délé-

[3] Michael Lynch est professeur à la *School of Life Sciences* et directeur du *Biodesign Center for Mechanisms of Evolution*, Tempe, États-Unis.

[4] En génétique des populations, la valeur sélective (ou *fitness*) d'un génotype est une mesure du succès reproducteur associé à ce génotype.

Le football évolutionnaire

Dans une population, les gènes existent sous différentes formes qu'on appelle allèles. Si les individus porteurs d'un allèle donné produisent, en moyenne, plus de descendants viables et fertiles que les porteurs d'autres allèles, cet allèle aura une meilleure valeur adaptative (ou meilleure *fitness*). En toute logique, un tel allèle finira toujours par se fixer, c'est-à-dire qu'il sera le seul à se maintenir dans la population, tandis que les autres allèles disparaitront. Pourtant, dans les populations naturelles, ce destin n'est pas écrit, en particulier à cause de la dérive génétique. Risquons une analogie footballistique pour comprendre comment la dérive génétique, un processus stochastique, perturbe un processus déterministe comme la sélection naturelle. Si un joueur pose un ballon sur le point central d'un terrain de foot et qu'il le pousse d'un mètre dans la direction du but adverse, et cela de façon répétée, le ballon finira inévitablement sa course dans ce but. Supposons maintenant qu'avant de pouvoir pousser à nouveau le ballon dans la bonne direction, un joueur adverse contre le ballon et l'envoie dans n'importe quelle direction (imaginons que le défenseur arrive parfois par les côtés, parfois par derrière et parfois en face de l'attaquant). Si le contre ne pousse ainsi le ballon que d'un centimètre avant qu'il ne soit poussé à nouveau d'un mètre dans la bonne direction, le ballon finira tout de même sa course, et rapidement, dans le but adverse. Mais, si le contre envoie le ballon dans une direction quelconque et à une distance de quelques mètres avant qu'il ne soit poussé à nouveau d'un mètre dans la bonne direction, il y aura alors de fortes chances que ce ballon sorte du terrain sans atteindre le but de l'adversaire. Il est même possible qu'il finisse sa course dans le but du joueur attaquant !

Taper au hasard dans le ballon correspond à la dérive génétique, c'est-à-dire aux variations aléatoires des fréquences alléliques entre deux générations. Le déplacement directionnel correspond à la sélection. Si c'est un petit déplacement aléatoire qui est réalisé par rapport au déplacement directionnel, cela a peu d'effet et c'est l'effet de la sélection qui domine ; si le déplacement aléatoire est grand, l'évolution des fréquences alléliques n'est plus prévisible et seule une probabilité peut être attribuée à la fixation d'un allèle.

tères à éliminer) et la nécessité de générer des mutations adaptatives, un taux de mutation élevé est supposé être la conséquence d'un besoin d'adaptation fréquent [18]. Ainsi, le taux très élevé de mutation d'un virus comme le VIH (virus de l'immunodéficience humaine), le menant à la limite du fardeau de mutation supportable, serait nécessaire à son adaptation rapide et récurrente au système immunitaire de son hôte et aux traitements médicaux [17]. L'équipe de Furio *et al.* [17] montre que l'augmentation de la fidélité de la réplication par mutation de l'ARN polymérase se fait au prix d'un ralentissement du taux de réplication, induisant une réduction de la valeur sélective. Les lignées les plus mutatrices ne présentent pas une meilleure adaptabilité. Les auteurs en concluent que le taux de mutation est un équilibre résultant de deux forces sélectives opposées, le fardeau de mutations favorisant les taux de mutation faibles et la vitesse de réplication favorisant les taux de mutation élevés.

Quel est le taux de mutation le plus élevé possible ?

Des taux de mutation très élevés ne sont observés que chez certains virus, mais jamais chez les procaryotes, ni chez les eucaryotes. On peut faire l'hypothèse que ces virus vivent dans des environnements beaucoup plus instables que les autres virus et les organismes cellulaires. Non seulement c'est assez peu convaincant, mais une autre explication est suggérée par l'observation d'une corrélation négative entre la taille du génome et le taux de mutation [11]. Cette corrélation est encore plus évidente en considérant la taille efficace du génome. La taille efficace du génome est la partie du génome soumise à sélection, c'est-à-dire la part du génome où les mutations ne sont pas neutres ; soit pratiquement la totalité du génome chez la plupart des virus et des procaryotes, mais souvent seulement une petite partie du génome des eucaryotes qui est composé essentiellement de séquences intergéniques. La taille efficace d'un génome est souvent estimée, en première approximation, en additionnant la longueur des séquences codant les protéines. Il est attendu que plus un génome contient de séquences codantes, plus le nombre de sites soumis à sélection est important, c'est-à-dire plus de sites qui, s'ils mutent, auront un effet négatif sur la valeur sélective. Focalisons-nous maintenant sur les mutations létales pour faire un petit calcul simple. Si un génome contient 1 000 nucléotides invariables, qui ne peuvent être mutés sans conséquence létale pour le porteur, alors la probabilité de répliquer ce génome sans qu'il y ait une seule mutation à ces mille positions est égal à $(1-\mu)^{1000}$ (μ étant le taux de mutation). Cette probabilité peut être calculée en fonction du nombre de sites invariables et pour différents taux de mutation (*Figure 1*). Pour un taux de mutation égal à 10^{-3}, un génome contenant moins de mille sites invariables (soit seulement quelques gènes) peut produire des copies sans mutation, donc fonctionnelles. Pour ce même taux de mutation, un génome contenant plus de 10 000 sites invariables ne pourra pas produire de copies fonctionnelles. Ainsi, un virus ne contenant que quelques gènes peut avoir un fort taux de mutation sans que cela empêche la production de copies fonctionnelles du génome. Les procaryotes et les eucaryotes contiennent entre quelques centaines et quelques dizaines de milliers de gènes, ce qui implique nécessairement bien plus de 10 000 sites invariables et donc l'impossibilité d'observer un tel taux de mutation. Par exemple, le génome humain contient un peu plus de 20 000 gènes dont la longueur moyenne de la partie codante est d'un peu plus de 1 000 paires de bases, soit au total plus de 20 millions de nucléotides constituant

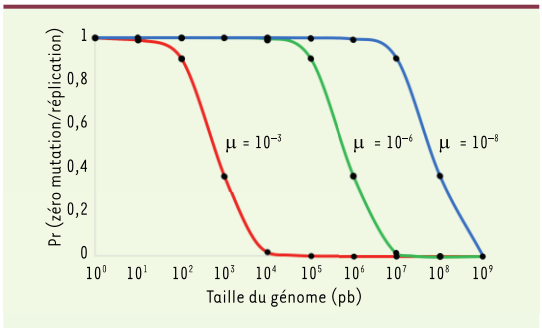


Figure 1. ***Probabilité d'une réplication sans mutation en fonction de la taille du génome et du taux de mutation.*** μ : taux de mutation ; pb : paire de bases ; Pr : probabilité.

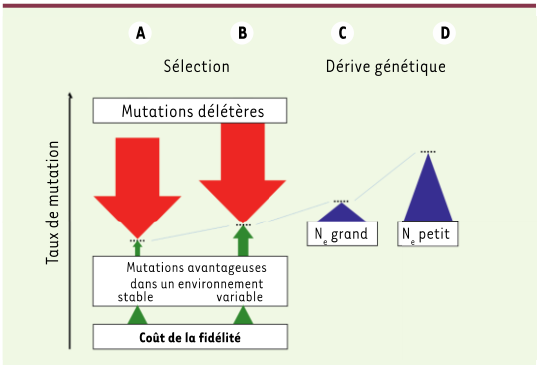


Figure 2. ***Effets de la sélection et de la dérive génétique sur le taux de mutation.*** ***A, B.*** sans dérive génétique, le taux de mutation est un équilibre entre des forces sélectives opposées favorisant les organismes mutateurs et antimutateurs. ***C, D.*** la dérive génétique pousse l'équilibre vers des valeurs plus grandes que celles attendues du seul fait de la sélection. N_e : effectif efficace de la population.

les séquences codant les protéines. Sans même considérer les séquences non-codantes ultra-conservées qui sont aussi soumises à la sélection, et en tenant compte que seule une fraction des séquences codantes ne peut pas être modifiée sans conséquence létale, il existe plusieurs millions de positions invariables parmi les 3,2 milliards de nucléotides par génome haploïde. Dans une telle situation, seul un taux de mutation inférieur à 10^{-8} permet de produire une forte proportion de génomes fonctionnels (*Figure 1*). Le taux de mutation maximal étant limité par la taille efficace du génome, on peut en déduire qu'au cours de l'évolution, l'augmentation de la taille efficace d'un génome ne peut se faire qu'en association avec la réduction du taux de mutation. S'il s'avérait qu'il est très difficile d'améliorer les taux de mutation les plus bas observés, alors ceci pourrait expliquer que le nombre de gènes dans des génomes plafonne toujours à quelques dizaines de milliers.

Quel est le taux de mutation le plus bas possible ?

Les analyses théoriques décrites précédemment convergent pour nous amener à conclure que chez des organismes dont le génome est constitué de plusieurs milliers de gènes et dans un environnement relativement stable, le taux de mutation doit tendre vers zéro mais rester au-dessus d'une limite imposée par les propriétés physico-chimiques des complexes moléculaires qui répliquent et réparent l'ADN et par la nécessité de produire des mutations à valeur adaptative. Lynch a proposé que s'il ne tend pas vers zéro, c'est plutôt parce que la sélection pour des taux de mutation toujours plus bas devient de plus en plus difficile car les différences de valeurs sélectives deviennent de plus en plus faibles. Pour comprendre ce point, imaginons un taux de mutation de 10^{-1} dans une population dans laquelle apparaît un individu avec un taux de mutation dix fois plus faible, soit 10^{-2}. La différence des nombres de descendants viables produits par ce mutant par rapport aux autres individus sera importante. Mais si le taux de mutation est de 10^{-8} dans une population dans laquelle apparaît un individu avec un taux de mutation dix fois plus faible, soit 10^{-9}, cette différence sera très faible car dans cette population, la plupart des descendants sont viables. En d'autres termes, plus un système est optimisé, plus son amélioration par le mécanisme de sélection de nouveaux allèles devient difficile dans une population de taille finie, car son action est de plus en plus faible et donc d'autant plus perturbée par la dérive génétique qui augmente ou réduit de façon aléatoire la fréquence des allèles anti-mutateurs. Dans une population de taille efficace très grande, la dérive génétique étant plus faible, la sélection devrait amener le taux de mutation vers des valeurs plus petites (*Figure 2*). Cet attendu est largement vérifié, aussi bien pour les mutations ponctuelles [9, 14, 19] que pour le taux d'insertion/délétion de courtes séquences [20]. Ainsi, le taux de mutation est corrélé négativement à la taille efficace des populations (*Figure 3*) : les bactéries ont des tailles efficaces souvent plus grandes que celles des eucaryotes et des taux de mutation plus faibles ; les eucaryotes unicellulaires ont des tailles efficaces souvent plus grandes que celles des pluricellulaires et des taux de mutation plus faibles.

Le taux de mutation est-il égal dans la lignée germinale et les lignées somatiques ?

Nous avons discuté des contraintes agissant sur l'évolution du taux de mutation, en particulier la sélection naturelle qui favorise les individus antimutateurs au travers d'une meilleure valeur adaptative de leurs

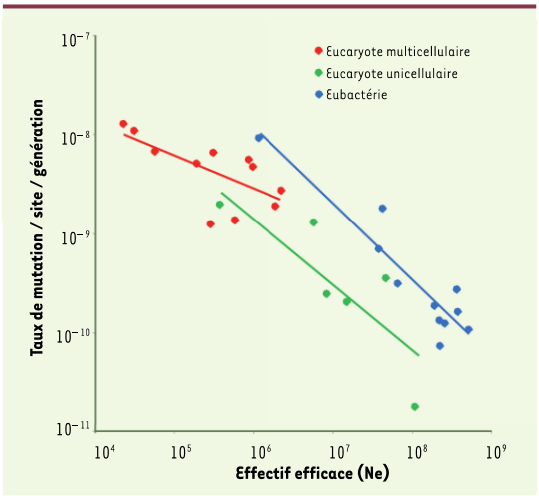


Figure 3. ***Corrélation entre le taux de mutation et l'effectif efficace*** (modifié d'après Lynch *et al.* [14]).

descendants. C'est donc le taux de mutation dans la lignée germinale qui est importante ici. Même si beaucoup de mutations se produisent dans les lignées somatiques, c'est finalement peu important car ces mutations disparaîtront toutes à la mort de l'individu. On peut donc supposer que les lignées somatiques sont soumises à une sélection moins forte pour réduire le taux de mutation, et c'est effectivement ce qui est observé [13].

Quel est le rôle du mode de reproduction sur l'évolvabilité[5] du taux de mutation ?

Nous venons de décrire des arguments théoriques et expérimentaux qui suggèrent que le taux de mutation dans une espèce donnée dépend principalement de 1) la taille efficace de son génome qui définit le taux de mutation maximum et 2) l'effectif efficace qui contrôle le taux de mutation minimum accessible, c'est-à-dire au-dessus de la barrière de la dérive génétique. Il est attendu que le taux de mutation pourra donc fluctuer autour du taux de mutation minimum accessible au gré des allèles mutateurs et antimutateurs qui apparaîtront au cours du temps. Nous pouvons introduire maintenant un niveau de complexité supplémentaire en tenant compte du fait que des espèces possèdent une reproduction sexuée (la plupart des eucaryotes) et d'autres non (les procaryotes). Nous allons voir que cette différence a des conséquences sur l'évolvabilité du taux de mutation. Chez les procaryotes, si une cellule a un fort taux de mutation alors, dans des conditions où une mutation est nécessaire pour s'adapter à l'environnement, il est

[5] Dans le contexte de cet article, ce terme désigne la possibilité de grandes variations au cours d'une période évolutive réduite.

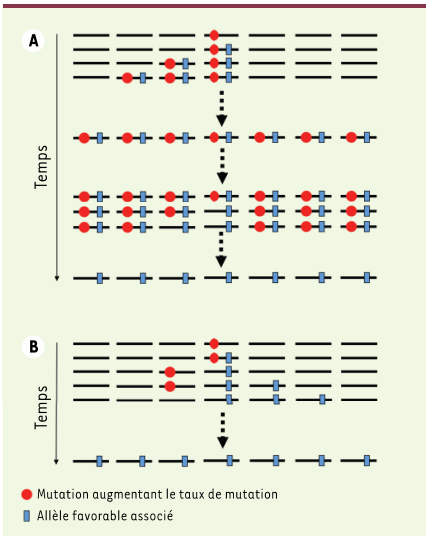


Figure 4. ***Effet du mode de reproduction sur le taux de mutation. A.*** Dans une population à reproduction clonale, du fait de la liaison totale des gènes, par effet auto-stop, l'allèle « mutateur » se fixe dans la population avec l'allèle favorable. L'élimination de l'allèle mutateur implique une réversion ou le transfert horizontal d'un allèle non-mutateur. ***B.*** Dans une population à reproduction sexuée, la recombinaison dissocie rapidement l'allèle mutateur de l'allèle favorable si la liaison n'est pas très forte, diminuant ainsi l'effet auto-stop et la probabilité de fixation de l'allèle mutateur en même temps que l'allèle favorable (modifié d'après Sniegowski *et al.* [11]).

probable que cette mutation utile apparaisse chez un descendant de la cellule mutatrice. Sans reproduction sexuée, il y a peu de recombinaisons des génomes et l'allèle muté se fixera dans la population en association avec l'allèle mutateur (c'est l'effet auto-stop, Figure 4). En parallèle de son adaptation, la population aura acquis un fort taux de mutation. Si une cellule, tout en gardant l'allèle utile, perd l'allèle mutateur, elle sera avantagée car elle restera adaptée à l'environnement et elle réduira le fardeau de mutation dû à l'allèle mutateur. Ainsi, dans des milieux instables, il peut se mettre en place des cycles dans lesquels des populations de bactéries voient leur taux de mutation augmenter fortement puis revenir vers des valeurs beaucoup plus basses. Bien décrit sur le plan théorique, ce processus a été observé aussi bien dans des population expérimentales que naturelles de bactéries *E. coli* [23-27]. Chez une espèce à reproduction sexuée, si une mutation utile apparaît associée à un allèle muta-

teur, la recombinaison sépare rapidement les deux allèles, il n'y a pas d'effet auto-stop et il est donc peu probable de pouvoir identifier des populations mutatrices apparues par ce mécanisme [123] (Figure 4).

Conclusion

La question « Pourquoi le taux de mutation n'est-il jamais égal à zéro » pourrait sembler triviale au premier abord : s'il n'y avait pas de mutations, il n'y aurait pas eu d'évolution et puisqu'il y a évolution, il y a donc bien eu des mutations, fermez le ban ! Elle a pourtant été à l'origine d'une suite de réflexions théoriques et de travaux expérimentaux d'une grande subtilité. Il reste beaucoup de questions ouvertes concernant ce point très important de la biologie évolutive, mais le voile se lève peu à peu. L'ensemble des approches disponibles aujourd'hui montre que la production de mutations, matériel de base de l'évolution, est elle-même un processus soumis aux mécanismes évolutifs qui dépendent de façon complexe de l'architecture des génomes des différentes espèces ainsi que de la stabilité des environnements dans lesquels elles vivent, et de leur démographie. ◊

SUMMARY

Why the mutation rate never reaches zero?

Alfred H. Sturtevant was the first to raise the question: why does the mutation rate not become reduced to zero? Indeed, most new mutations with a phenotypic effect are deleterious. Therefore, individuals who produce less mutants produce more viable and fertile offspring. Consequently, natural selection should increase the frequency of antimutator genotypes and progressively reduce the mutation rate to zero. However, no species has ever been found with a mutation rate equal to zero. Recent analyses suggest that setting the mutation rate above zero depends mainly on the effective size of the genome and the effective population size. The mutation rate is a trade-off between natural selection that operates to improve replication fidelity and the random genetic drift that sets the ultimate lower limit. This trade off illustrates the limitation of the power of natural selection in a world where natural populations have a finite size. ◊

REMERCIEMENTS

Nous remercions chaleureusement nos collègues Jean-Luc Da Lage et Julien Fumey pour leurs fort pertinents commentaires et suggestions au cours de l'élaboration de cet article.

LIENS D'INTÉRÊT

Les auteurs déclarent n'avoir aucun lien d'intérêt concernant les données publiées dans cet article.

RÉFÉRENCES

1. Sturtevant AH. Essays on evolution. I. On the effects of selection on mutation rate. *Quart Rev Biol* 1937 ; 12 : 464-7.
2. Demerec M. Genetic factors stimulating mutability of the miniature gamma wing character of *Drosophila virilis*. *Proc Natl Acad Sci USA* 1929 ; 15 : 834-8.
3. Demerec M. Frequency of spontaneous mutations in certain stocks of *Drosophila melanogaster*. *Genetics* 1937 ; 22 : 469-78.
4. Muller HJ. The measurement of gene mutation rate in Drosophila, its high variability, and its dependence upon temperature. *Genetics* 1928 ; 13 : 279-357.
5. Sniegowski PD, Gerrish PJ, Johnson T, Shaver A. The evolution of mutation rates: separating causes from consequences. *Bioessays* 2000 ; 22 : 1057-66.
6. Drake JW. A constant rate of spontaneous mutation in DNA-based microbes. *Proc Natl Acad Sci USA* 1991 ; 88 : 7160-4.
7. Drake JW. Rates of spontaneous mutation among RNA viruses. *Proc Natl Acad Sci USA* 1993 ; 90 : 4171-5.
8. Sung W, Tucker AE, Doak TG, *et al.* Extraordinary genome stability in the ciliate *Paramecium tetraurelia*. *Proc Natl Acad Sci USA* 2012 ; 109 : 19339-44.
9. Lynch M. Evolution of the mutation rate. *Trends Genet* 2010 ; 26 : 345-52.
10. Eyre-Walker A, Keightley PD. The distribution of fitness effects of new mutations. *Nat Rev Genet* 2007 ; 8 : 610-8.
11. Robert L, Ollion J, Robert J, *et al.* Mutation dynamics and fitness effects followed in single cells. *Science* 2018 ; 359 : 1283-6.
12. Casane D, Laurenti P. Syllogomanie moléculaire: l'ADN non codant enrichit le jeu des possibles. *Med Sci (Paris)* 2014 ; 30 : 1177-83.
13. Charlesworth B. Effective population size and patterns of molecular evolution and variation. *Nat Rev Genet* 2009 ; 10 : 195.
14. Lynch M, Ackerman MS, Gout JF, *et al.* Genetic drift, selection and the evolution of the mutation rate. *Nat Rev Genet* 2016 ; 17 : 704-14.
15. Quinones A, Piechocki R. Isolation and characterization of *Escherichia coli* antimutators. A new strategy to study the nature and origin of spontaneous mutations. *Mol Gen Genet* 1985 ; 201 : 315-22.
16. Loh E, Salk JJ, Loeb LA. Optimization of DNA polymerase mutation rates during bacterial evolution. *Proc Natl Acad Sci USA* 2010 ; 107 : 1154-9.
17. Furio V, Moya A, Sanjuan R. The cost of replication fidelity in an RNA virus. *Proc Natl Acad Sci USA* 2005 ; 102 : 10233-7.
18. Lloyd SB, Kent SJ, Winnall WR. The High Cost of Fidelity. *AIDS Res Hum Retrovir* 2014 ; 30 : 8-16.
19. Lynch M. The lower bound to the evolution of mutation rates. *Genome Biol Evol* 2011 ; 3 : 1107-18.
20. Sung W, Ackerman MS, Dillon MM, *et al.* Evolution of the insertion-deletion mutation rate across the tree of life. *G3 (Bethesda)* 2016 ; 6 : 2583-91.
21. Giraud A, Matic I, Tenaillon O, *et al.* Costs and benefits of high mutation rates: adaptive evolution of bacteria in the mouse gut. *Science* 2001 ; 291 : 2606-8.
22. Denamur E, Matic I. Evolution of mutation rates in bacteria. *Mol Microbiol* 2006 ; 60 : 820-7.
23. Denamur E, Lecointre G, Darlu P, *et al.* Evolutionary implications of the frequent horizontal transfer of mismatch repair genes. *Cell* 2000 ; 103 : 711-21.
24. Labat F, Pradillon O, Garry L, *et al.* Mutator phenotype confers advantage in *Escherichia coli* chronic urinary tract infection pathogenesis. *FEMS Immunol Med Microbiol* 2005 ; 44 : 317-21.
25. Taddei F, Radman M, MaynardSmith J, *et al.* Role of mutator alleles in adaptive evolution. *Nature* 1997 ; 387 : 700-2.
26. Tenaillon O, Toupance B, Le Nagard H, *et al.* Mutators, population size, adaptive landscape and the adaptation of asexual populations of bacteria. *Genetics* 1999 ; 152 : 485-93.
27. Tanaka MM, Bergstrom CT, Levin BR. The evolution of mutator genes in bacterial populations: the roles of environmental change and timing. *Genetics* 2003 ; 164 : 843-54.
28. Johnson T. Beneficial mutations, hitchhiking and the evolution of mutation rates in sexual populations. *Genetics* 1999 ; 151 : 1621-31.

TIRÉS À PART

P. Laurenti

2) Diversity of Olfactory Responses and Skills in Astyanax Mexicanus Cavefish Populations Inhabiting different Caves

Article

Diversity of Olfactory Responses and Skills in *Astyanax Mexicanus* Cavefish Populations Inhabiting different Caves

Maryline Blin [1], Julien Fumey [2,#], Camille Lejeune [1,#], Maxime Policarpo [1,3], Julien Leclercq [1], Stéphane Père [1], Jorge Torres-Paz [1], Constance Pierre [1], Boudjema Imarazene [1,4] and Sylvie Rétaux [1,*]

[1] Paris-Saclay Institute of Neuroscience, CNRS and University Paris-Saclay, 91190, Gif-sur-Yvette, France; blin@inaf.cnrs-gif.fr (M.B.); camille.sabine.lejeune@orange.fr (C.L.); Maxime.Policarpo@egce.cnrs-gif.fr (M.P.); julien.leclercq@u-psud.fr (J.L.); Stephane.pere@inaf.cnrs-gif.fr (S.P.); jorge.torres-paz@cnrs.fr (J.T.-P.); constance.pierre@free.fr (C.P.); boudjema.imarazene@gmail.com (B.I.)

[2] Human Genetics and Cognitive Functions, Institut Pasteur, UMR3571 CNRS, Université de Paris, 75015, Paris, France; julien.fumey@pasteur.fr

[3] Évolution, Génomes, Comportement et Écologie, IRD, CNRS and Université Paris-Saclay, 91190, Gif-sur-Yvette, France

[4] INRAE, Laboratoire de Physiologie et Génomique des Poissons, 35000, Rennes, France

* Correspondence: retaux@inaf.cnrs-gif.fr; Tel.: +33(0)-1-69-82-34-52

[#] Equal contribution to this work

Received: 15 September 2020; Accepted: 08 October 2020; Published: 13 October 2020

Abstract: Animals in many phyla are adapted to and thrive in the constant darkness of subterranean environments. To do so, cave animals have presumably evolved mechano- and chemosensory compensations to the loss of vision, as is the case for the blind characiform cavefish, *Astyanax mexicanus*. Here, we systematically assessed the olfactory capacities of cavefish and surface fish of this species in the lab as well as in the wild, in five different caves in northeastern Mexico, using an olfactory setup specially developed to test and record olfactory responses during fieldwork. Overall cavefish showed lower (i.e., better) olfactory detection thresholds than surface fish. However, wild adult cavefish from the Pachón, Sabinos, Tinaja, Chica and Subterráneo caves showed highly variable responses to the three different odorant molecules they were exposed to. Pachón and Subterráneo cavefish showed the highest olfactory capacities, and Chica cavefish showed no response to the odors presented. We discuss these data with regard to the environmental conditions in which these different cavefish populations live. Our experiments in natural settings document the diversity of cave environments inhabited by a single species of cavefish, *A. mexicanus,* and highlight the complexity of the plastic and genetic mechanisms that underlie cave adaptation.

Keywords: fieldwork; wild fish; comparative biology; behavior; troglomorphism; olfactory test; infrared movies; amino acids; chondroitin; plasticity

1. Introduction

A very broad diversity of fauna (micro-organisms, insects, vertebrates) lives in underground environments in a more or less permanent manner. Among various niches in the subterranean milieu, caves are emblematic and attractive to human exploration. Species living there permanently display striking phenotypic convergences in their morphology, physiology, or behaviors, with the hallmarks of troglomorphism being the loss of eyes and pigmentation [1]. Caves are often considered as an extreme environment. In the absence of photoautotrophic production, the quantity of food available is limited or

irregular, the space available is finite, and reproduction seems difficult. Finding food and mates in the absence of vision are the two main challenges faced by cave animals and must limit cave colonization and survival. The evolutionary forces at play during cave adaptation and the respective contributions of natural selection and genetic drift, along with the evolutionary mechanisms, are still a matter of debate [2]. Biologists currently aim at disentangling the roles of genetic mutations and phenotypic plasticity, or epigenetics, in the process. Finally, the observation that some species or lineages have repeatedly adapted to the cave environment while some others never did may support the questioned idea of "pre-adaptive traits" that might favor adaptation to permanent darkness (e.g., [3,4]).

During evolution, most epigean representatives of species that became cave-adapted have become extinct, leaving the underground lineages the only representatives of their taxon, which hampers comparative or genetic studies. The teleost fish *Astyanax mexicanus* is one of the few exceptions to this rule [5,6]. Therefore, the surface-dwelling and cave-dwelling morphs of this species are increasingly used in evolutionary studies to address the developmental, genetic, or genomic mechanisms of morphological evolution and behavioral adaptation [2,7–9]. The surface form (SF) lives in the rivers of the southern United States and Central America, while the blind and depigmented cave form (CF) is endemic to caves in a karst region located in the states of San Luis Potosi and Tamaulipas in Mexico. There, 30 identified caves host *Astyanax mexicanus* cavefish populations. They are distributed into three geographically distant groups located, respectively, in the Sierra de El Abra, the Sierra de Guatemala, and the Sierra Colmena [5,10,11]. All populations of cavefish and surface fish are interfertile, indicating that they are conspecific [12,13]. In nature, the hybridization phenomenon has also been observed and documented [14]. In addition, crosses between geographically distant populations of cavefish can lead to eyed F1 offspring [13], indicating that different mutations are involved in ocular regression in different cave populations and suggesting that some of these populations have evolved independently. However, the evolutionary history of *A. mexicanus* cave populations is still poorly understood because of the geographic dispersion of caves, the lack of knowledge on the underground aquifer network, and the possibility of surface fish introgressions into caves as well as cavefish migrations between caves using underground flows. Recent studies have indicated that initial cave colonization by *A. mexicanus* surface-like ancestors occurred very recently, less than 20,000 years ago [15,16], prompting evolutionary biologists to revise some views about the (rapid) mechanisms of cave adaptation. Like most cave-adapted animals, *A. mexicanus* cavefish present sensory specializations to life without vision. The brains and sensory systems of surface fish and cavefish differ (reviewed by the authors of [7]), along with their sensory systems: Cavefish have more taste buds [17], more neuromasts [18], and larger olfactory epithelia [14,19,20]. From a behavioral point of view, these mechano- and chemosensory specializations are associated with vibratory attraction behavior to locate moving objects [18] and to an excellent sense of smell to detect low concentrations of food-related odors [20], respectively.

Most of the results described above were obtained in the laboratory, often on one or two lab-raised cave population(s). It is therefore important to extend the studies to other populations, and to validate the results on wild animals in order to avoid possible misinterpretations. Going to the field to observe the natural environment of fish, taking samples, and filming behaviors can help answer questions or revise preconceived ideas [21,22]. For instance, it is often stated that caves are a food-poor environment. However, analysis of stomach contents of wild individuals from the Pachón cave has shown that juveniles feed on small arthropods, and adults on decaying materials and bat guano. Overall and contrary to common belief, Pachón cavefish seemed relatively well fed [23]. Another study revealed that growth curves and age/size relationship are comparable in wild surface fish and wild cavefish, again indicating that cave environment is probably not as food-poor as it may seem [24]. In fact, depending on the location of the cave, its topography, and the hydraulic regime, the amounts of carbon flux can sometimes be of the same order as those reported for surface rivers [1] and, most importantly, the carbon content in the mud sampled from different *Astyanax* caves can show up to three-fold variation [24]. Energy sources can come from percolating water, animals entering caves and depositing their excrement or dead bodies, or rivers that overflow during the rainy season and carry organic matter. These energy sources are both spatially and temporally variable [1]. During several cave expeditions, our team noticed the diversity of local environments between caves —

and also between pools within a single cave—reinforcing the idea that field comparisons on fish biology between different caves can be as interesting as the comparison between cavefish and surface fish.

During a field trip in 2013, we carried out, for the first time, experiments of olfactory behavior in situ, in the Subterráneo cave [14]. We had the idea to use a small, light, compact, inflatable children's plastic pool, which was easy to bring on the field. With a rudimentary perfusion system and an infrared camera, we showed that only fish with eyeless phenotype and large olfactory epithelia swam toward an odor source consisting of a food extract. Thanks to this experience and after visits to many other caves, we set up a more complex behavioral experiment, with the aim of systematically assessing olfactory skills and responses of *A. mexicanus* cavefishes inhabiting different Mexican caves. During expeditions in 2016, 2017, and 2019, we performed olfactory tests in the Pachón, Sabinos, Tinaja, Chica, and Subterraneo caves. We found that wild adult cavefish from these five caves showed very variable responses to the three different odorant molecules they were exposed to, with Pachón and Subterraneo cavefish showing the highest olfactory capacities, and Chica cavefish showing no olfactory responses to the odors presented. We discuss these data with regard to the environmental conditions in which these different cavefish populations live.

2. Materials and Methods

2.1. Field Experimentation in Five Different Caves: Constraints and Criteria for Choice

Caves chosen to carry out olfaction experiments had to fulfill several criteria: (1) Reasonably easy access and climbing challenges, as the total weight of experimental equipment was approximately 25 kg carried in backpacks and each location had to be visited twice on two consecutive days. (2) Sufficient space inside the cave to install three experimental plastic pools near the water. Thanks to our field experience, we excluded some caves. For example, Curva's ceiling is too low, Chiquitita's entrance is too narrow, as it is located in a big tree's root [10], Toro is a fault in the rock, and Molino and other Guatemala caves are too challenging in terms of climbing ([11] and team observations). (3) Caves already known and visited by the team in the past were preferred to plan the precise place where to install plastic pools and to anticipate troubleshooting. (4) Good representative sampling of diverse local environments (e.g., rocky and muddy caves). (5) Good representative sampling of diverse cavefish population histories (e.g., with or without surface gene flow, or mountain range in which the cave is located). (6) Possibility to compare with our lab studies performed on Pachón cavefish.

Consequently, we decided to perform experiments in four caves: Pachón and Sabinos (fully troglomorphic fish morphotypes in muddy caves), Chica (introgressed fish population in "dirty" cave), and Subterráneo (introgressed population in rocky cave). We also performed some preliminary experiments in Tinaja (fully troglomorphic fish morphotypes in rocky or muddy cave ponds).

2.2. Cavefishes in the Wild

Olfactory behavior tests were carried out during three field trips to San Luis Potosi and Tamaulipas States, Mexico, in March 2016, March 2017, and March 2019, in five cave localities. Fieldwork Mexican permits 02438/16, 05389/17 and 1893/19 (to SR and Patricia Ornelas-Garcia) were delivered by the Secretaria de Medio Ambiente y Recursos Naturales. The history of the discovery and precise descriptions of *Astyanax* caves are given by the authors of [5,11].

The Pachón cave is located in altitude near the village of Praxedis Guerrero (22°37′ N latitude and 99°01′ W longitude, about 16 km SW of Ciudad Mante), in the north Sierra de El Abra, and is easy to access [11]. The cave is small, and the water is stagnant on a muddy bottom. Fish from the Pachón cave (named here CF-Pachón) present a fully troglomorphic type.

The Sabinos cave is located near the village of El Sabino, in the central Sierra de El Abra (22°06′ N latitude and 89° 56′ W longitude, about 13 km NNE of Ciudad Valles) [11]. Villagers installed a padlock grid to exploit this cave and the access is chargeable. The entrance is majestic and the succession of two pits involves bringing harnesses and ropes to abseil down. Fish from the Sabinos cave (CF-Sabinos) are also fully troglomorphic.

The Chica cave is located at the south Sierra de El Abra, on the property of a farmer, at about 21°52′ N latitude and 89° 56′ W longitude, near the village of El Pujal. It is easy to access, and the entrance and the first cavity are wide. Fish from the Chica cave (CF-Chica) are phenotypically diverse because of surface fish introgression and hybridization. We worked on cavefish from the Chica superficial pool, which are the most troglomorphic/least introgressed among the three natural pools of this cave.

The Subterráneo cave is located in the Micos region, in the Sierra de Colmena (22°03′ N latitude and 99°14′ W longitude, about 10 km SSW of Micos). Access to this cave is not difficult and the entrance is easy, at the level of the polje or sugar cane field. Fish from the Subterráneo cave (CF-SubT) also occasionally hybridize with surface fish. Hence, they show variable levels of eye regression and pigmentation.

The Tinaja cave is very close to the Sabinos cave, also located in the territory of El Sabino (entrance is free). The cave is located at 22°05′ N latitude and 89°57′ W longitude, about 10.5 km NE of Ciudad Valles on the Rancho de La Tinaja. Access is via a sugar cane field and permission must be obtained from the owner before crossing it. The cave entrance is accessible after a 2-h hike through a thorny tree forest and a dry canyon covered with jungle. According to Eliott, 2016 [6], the underground hydraulic systems of Tinaja, Sabinos and Sótano de Soyate caves are connected. It has been suggested indeed that the cavefish populations of Tinaja and Sabinos are genetically close [25,26]. The cave entrance is majestic. Climbing equipment is not needed but hiking is difficult due to a very slippery mud covering a stony soil. The air is charged with CO_2 and renders physical effort somewhat difficult. Fish from the Tinaja cave (CF-Tinaja) show a full cave morphotype.

2.3. Fishes from the Lab Facility

Laboratory *Astyanax mexicanus* surface fish (origin: San Salomon spring, Reeves County, TX, USA) and cavefish (Pachón population) were obtained in 2004 from the Jeffery laboratory at the University of Maryland, College Park, Prince George's County, MD, USA. Here, we also used F1 hybrids, which were the progeny of a cross between a Pachón female and a surface fish male. Colonies were maintained at 22 °C (cavefish and F1 hybrids) or 26 °C (surface fish) on a 12:12 h light:dark cycle. In the present paper, lab-raised fish are named Lab-Pachón, Lab-SF, and Lab-Hyb, respectively. SR's authorization for use of *Astyanax mexicanus* in research is 91–116 and the Paris Centre-Sud Ethic Committee protocol authorization number related to this work is 2017-04#~8545. The animal facility of the Institute received authorization 91272105 from the Veterinary Services of Essonne, France, in 2015.

2.4. Sampling and Photography

Wild and lab fish were caught with a net (hand net or seine). In order to record their phenotypes after the behavioral tests, wild cavefish were photographed individually in a small aquarium or a plastic support with a graduated ruler and immediately returned to their pond of origin. Total body lengths were measured from these pictures using the ImageJ software. In the Pachón and Sabinos caves, we also weighed the fish using a portable balance. Taring of the balance was carried out with a glass of water, and each fish was weighed inside the glass of water.

2.5. Odor Choice

The general principle of the experiment was to place eight fish in an experimental square plastic pool, and to successively perfuse three odorant compounds, each in a different corner of the pool (Figure 1). Each odor perfusion was preceded and followed by a perfusion of water, and a water counterflow flowed permanently from the corner opposite to the perfusion of odor. The odors used were L-alanine and L-serine amino acids, and chondroitin (all from Sigma). They correspond to degradation products of organic compounds assimilated to food odors and serve as attractants to fish.

The rationale for choosing alanine and serine as odorant molecules were: (i) To compare them with the results already obtained in the laboratory [19,20]); (ii) due to the fact that attraction and food-searching behavior are easier to identify than repulsion behavior; (iii) because finding food is a matter of survival for these fish and the sense of smell probably plays a major role in this quest. Indeed, previous experiments

performed in the lab had shown that responses to alanine and serine are olfactory-mediated in *Astyanax* larvae [20].

We chose working concentrations of 10^{-7} M for alanine and serine because it is intermediate between the detection threshold for larvae of surface fish (10^{-5} M) and Pachón cavefish (10^{-10} M) in the lab. We also reasoned that it would be "risky" to use a very low concentration close to the cavefish detection threshold in non-controlled cave/field experimental conditions (in particular, water cleanliness, more or less charged with natural odors).

As chondroitin induces freezing in zebrafish [27], we tested this molecule in the lab assuming that *A. mexicanus* would also adopt a freezing behavior [28,29]. Surprisingly, both surface fish and Pachón cavefish instead showed a pronounced and persisting foraging behavior (see Video S1). Preliminary tests in the lab showed that Pachón cavefish have a chondroitin detection threshold at 10^{-4} M.

2.6. Setup Design

With the conditions for setting up and carrying out the experiments in situ being difficult, the setup was meant to be light, easy to assemble and dismantle, and not too bulky. Plastic pools (Intex, 85 cm × 85 cm × 23 cm; thoroughly rinsed several times in the lab before use to eliminate inorganic volatile odors) were inflated and placed on the ground after adjusting for horizontality with mud or rocks if necessary. They were filled with 40 L of local cave pond water where cavefish swim. Before pouring into the plastic pools, water was filtered on a coffee filter paper to remove large suspended particles and to ensure cleanliness and good video quality.

A tripod was placed near the corner #1 of the plastic pool. Two infrared lamps (IR Torch 850nm, (Maketheone, LA, USA) and the syringe holder were attached to the tripod at a height of 110 cm. The pool was enlightened from above to reduce shadows and filmed from the top to avoid blind spots. For the 2016–2017 campaigns, we built our own infrared camera using a Raspberry pi 3 model B and a Pi-NoIR V2.1 camera. We used a 5-V power bank to supply electricity to this camera, which was controlled from a computer connected in ethernet via a SSH connection. Using Python 2.7.9 and the PiCamera V1.12 package, we wrote two scripts (available at https://github.com/julienfumey/PiCaveRecord) to frame on the pool and to record the video. Videos were converted from h264 to mp4 with VLC software [30]. In 2019, to avoid having to use a computer in caves and to save time, we opted for an infrared hunting camera (nature camera Full HD WK-590, VisorTech, Nairobi, Kenya) which records films in AVI format.

For each plastic pool, we constructed a set of four tubing lines consisting of four 50-mL syringes connected to a 180-cm-long medical solution administration tubing (Infusion device, Intrafix® SafeSet; B. Braun; inside volume of 20 mL) and terminated by a 0.6-mm-diameter needle (protected by a plastic cap to avoid wounding the fish). The opening of the perfusion was controlled by a Luer stopper and the perfusion rate was regulated by the needle. The end of the tubing was attached to a metal guide (which did not touch the water) to hold the needle in the corner of the plastic pool. Each perfusion line was guided to its respective corner.

A set of three pools was installed in the visited cave, with eight cavefish in each (n = 24 total). The installation of three experimental setups was completed in ~4 h by four people. Fish were left for a 20-h habituation period (from approximately 5 p.m. day 1 to 1 p.m. day 2). The next day, behavioral tests were carried out in the dark (IR recordings) and in silence, in parallel, by three experimenters.

2.7. Procedure

Chemicals were weighed in the laboratory on the day before field trip departure. In a hermetically closed 50-mL tube, 10 mg of L-alanine (CAS 56-41-7, Sigma-Aldrich, Saint-Quentin Fallavier, France), 10 mg of L-serine (CAS 56-45-1, Sigma-Aldrich), and 5 mg or 110 mg of chondroitin (CAS 9082-07-9, Sigma-Aldrich) were each placed (a series of three tubes of powder per cave were thus prepared in advance). Solutions were prepared extemporaneously by adding 50 mL of filtered cave water to each tube. Chondroitin concentration was 10^{-4} M and 10^{-3} M, respectively. For alanine and serine working solutions, a second dilution (3 µL in 50 mL) was prepared to obtain 10^{-7} M solutions. For experiments performed in the

lab with surface fish, we prepared a different dilution (300 µL in 50 mL) to obtain 10^{-5} M alanine or serine solutions.

The experimenter seated near the pool took care not to move and not to speak during the whole experiment. To start the experiment, filtered water (= control) was perfused from tubing #1 and #3 (i.e., opposite corners), and then from tubing #2 and #4 (i.e., opposite corners), for 6 min each. The aim of this step was to accustom fish to flow and possible vibrations of perfusions and to reduce subsequent nonspecific responses. Indeed, thanks to their lateral line, fish perceive and are attracted by vibrations [18].

Then, water was perfused for 6 min from tubing #1 and #3. Solution flow from the two syringes was initiated simultaneously. When syringes emptied, 50 mL of alanine solution was added to syringe #1 and water was added again for counter-flow in syringe #3 (Figure 1, left panel, orange for alanine). After completion of alanine perfusion, the test continued with a new water perfusion from corners #1 and #3. Experimenters took care to ascertain a continuous flow by filling syringes with water or odor solutions before they were completely empty to avoid the introduction of air bubbles into the system.

The same principle was then applied for the perfusion of water, serine or water, and water at corners #2 and #4 (Figure 1, middle panel, green for serine). Finally, water, chondroitin or water, and water, were perfused at corners #3 and #1 (Figure 1, right panel, blue for chondroitin).

With the exact same setup and procedure, we performed series of experiments in the lab, in a dark and soundproof room, using animals from our breeding facility.

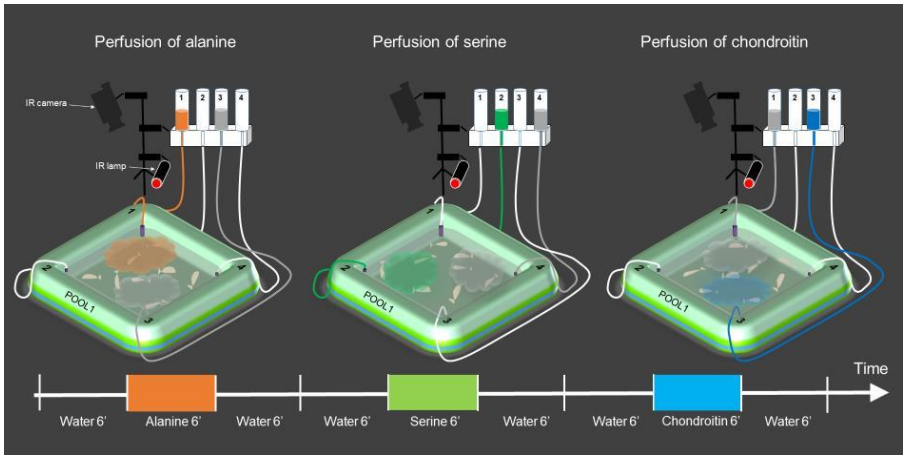


Figure 1. Establishment of a behavioral setup and protocol to test cavefish olfaction in the field. The setup allows testing olfactory responses of eight adult fish in inflatable plastic pools in the dark under infrared recordings. Three different odors were perfused sequentially by gravity flow at different corners of the plastic pool, according to the indicated timeline. See Methods and the first paragraph of Results for details.

2.8. Video Scoring

A total of ~60 h of infrared videos was recorded. We first attempted to analyze them with an automatic multiple tracking video software [31], which unfortunately did not detect the fish correctly. The reasons were probably numerous: There were eight fish in the pool, the film was in infrared, the contrast of fish on white background was not strong enough, the lamps sometimes produced lighting reflections on the water surface, and the pool edges generated blind spots. Thus, we could not satisfactorily extract the time spent by each fish in each zone and turned to manual scoring.

We first established that, after perfusion of 50 mL of blue-colored water (with a counterflow), the blue color occupied roughly half of the arena, i.e., a triangle formed by the corner from which the perfusion arrived and the diagonal of the plastic pool (Figure S1A). We therefore considered this half of the pool as

the "odorant area" and its opposite half as the "water/control area." Videos were tracked manually through frame-by-frame analysis using Windows Media Player. The number of fish present in each half of the pool was counted every 15 s during the whole experiment. The odor Preference Index (PI) was calculated for each odor using the formula: (Fish count (odor area) – Fish count (water area) / Total fish count (Figure S1B). When all fish were in the odorant area, PI = 1 (suggesting attraction effect); when all fish were in the opposite corner, PI = −1 (suggesting repulsion effect); and when fish were distributed evenly/randomly in the two parts of the pool, PI = 0 (suggesting no effect).

2.9. Data Analyses and Statistics

When the odorous solution was poured into the syringe, 20 mL (= dead volume contained inside the 180 cm long tubing) of water flew out before the actual odorant solution entered the pool. For each perfusion, we calculated the speed of the flow to determine odor (or water) perfusion duration. Although tubing were the same length and volume perfused by gravity flow was always 50 mL, the perfusion duration (theoretically 6 min) was appreciably variable both between the four perfusion lines, between the successive water/odor/water perfusions, and between the three plastic pools in a given cave. To overcome this problem, we normalized time and calculated PI means over periods corresponding to 25% of the total perfusion time for each sequence. Thus, we obtained four PI means for each perfusion duration. To determine the statistical significance of the behavioral responses, we used Friedman nonparametric test by ranks for repeated measures in order to compare PI variations along time, during water perfusion before odor, during odor perfusion, and during water perfusion after odor. We also used Wilcoxon Mann-Whitney tests to determine the statistical significance of each PI distribution against PI = 0 (no effect) for each time segment.

Fish sizes and weights were also compared using a nonparametric Wilcoxon–Mann–Whitney test.

Statistical analyses were performed using R 3.6.1 [32] in Rstudio environment [33] with rstatix package, version 0.5.0 [34]. Plots were generated using ggplot2 package [35] from tidyverse open-source R packages [36].

3. Results

3.1. Methodological Considerations: Testing Olfactory Responses in the Lab versus in the Wild

Our previous analyses of olfactory skills and behaviors in *A. mexicanus* were mostly performed on larvae, in small 9- × 13-cm U-shaped "olfaction boxes" containing 150 mL of water, and under laboratory-controlled conditions and standards [19,20]. We had also performed a preliminary experiment on adult fish in the Subterráneo cave but the insights were limited because, among other difficulties, a single "odor" consisting of crushed food pellets was tested as odorant cue [14]. Here, to reach our goal of testing several relevant odors on adult cavefishes inhabiting several caves, we developed a novel setup and a novel experimental procedure adapted to the field.

First, we carried out laboratory experiments to establish and validate the experimental setup (Figure 1 and see Methods) and to study adult fish olfactory behavioral responses. We reasoned that such results would also help us interpreting data obtained in caves. Cavefish from our animal facility originated from the Pachón cave but have been raised under markedly different conditions (food, light, water quality) for several generations [37], and they might have been affected by captivity and environment.

Moreover, laboratory experiments were the only option we had to study surface fish olfactory responses. In our experience, it has been impossible to test wild surface fish in the field. Besides the difficulty and time needed for catching wild fish in rivers, they are highly sensitive to manipulations and stress and do not behave "normally," displaying most signs of their stress repertoire [25]. In addition, due to daylight, setting up an experimental test near the river is complex, with a risk of predation on fish by wild animals during the habituation period or destruction/stealing of the equipment.

Finally, we also carried out tests on lab-generated F1 Hybrids. We previously showed that F1 Hybrids larvae have a relatively poor odor detection threshold, similar or even below surface fish skills [19]. In the

Subterráneo and Chica caves, wild fish present a wide variety of intermediate hybrid-like phenotypes. It was therefore interesting to perform tests on adult F1 Hybrids in the lab to compare with result obtained on larvae and on wild cavefish populations where hybridization occurs.

3.2. Responses to Odors in Laboratory-Raised A. Mexicanus

The results of laboratory experiments are presented in Figure 2 (n = 8 plastic pools averaged for each graph, hence 64 fish were tested for each morphotype).

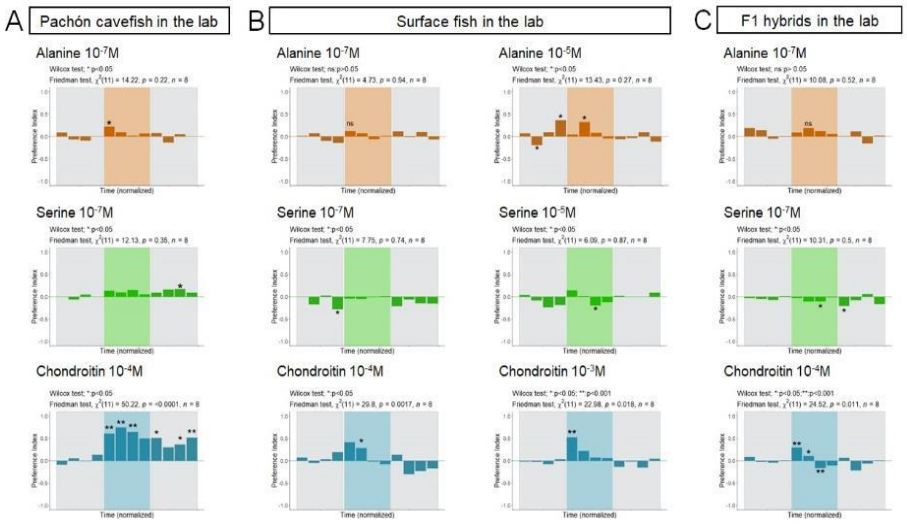


Figure 2. Testing olfactory skills in the laboratory; Olfactory responses of Pachón cavefish (**A**), Surface fish (**B**) and F1 hybrids (**C**) to the indicated odors at the indicated concentrations using the described olfactory setup in laboratory conditions (n = 8 for each); The preference index (PI; positive values suggest attraction) is shown as a function of time. Time intervals corresponding to odor perfusion are shaded. Asterisks on graphs indicate significance as compared to no response (i.e., PI = zero) for a given time interval (Wilcoxon Mann-Whitney test). The results of Wilcoxon and Friedman tests to probe the significance of the response across time with repeated measures are also indicated.

In agreement with previous experiments on larvae, only Lab-Pachón showed a positive attractive response to the low concentration of alanine 10^{-7} M (Figure 2A, orange). A modest but significant response was observed during the first quarter of alanine perfusion time (PI = 0.22; Wilcoxon test, p = 0.008; but Friedman test, NS), suggesting that Lab-Pachón detected the odor as soon as it arrived in the arena but were not attracted for more than a few minutes. On the other hand, Lab-SF did not respond to alanine 10^{-7}M but were significantly attracted by the higher concentration of 10^{-5} M alanine during the second quarter of perfusion time (Figure 2B, orange; PI = 0.32; Wilcoxon test, p = 0.008; but Friedman test, NS). These data suggest that, like larvae, adult Lab-Pachón have a better olfactory detection threshold than Lab-SF for alanine. Finally, Lab-Hyb did not respond to the low (10^{-7} M) alanine concentration, and were not further tested for a higher concentration (Figure 2C, orange).

Serine elicited very little, if any, response on surface and cave adult *Astyanax* (Figure 2A–C, green; Wilcoxon tests and Friedman tests, all NS). Of note, Lab-Pachón were present in the odor perfusion zone in a delayed manner, i.e., when water was subsequently perfused from that corner (PI = 0.16; Wilcoxon test, p = 0.008), suggesting that the response to serine may have different kinetics compared to the response to alanine. Surprisingly, transiently negative PIs were observed for both Lab-SF (10^{-5} M serine, PI = −0.2;

Wilcoxon test, $p = 0.008$) and Lab-Hyb (10^{-7} M serine, PI = −0.07 and −0.03; Wilcoxon test, $p = 0.008$ and 0.02), which was an unexpected result (see Discussion).

Chondroitin provoked strong positive responses in all lab-raised fish, as shown by statistical significance with both Wilcoxon tests and Friedman tests (Figure 2A–C, blue). Lab-Pachón showed intense response to 10^{-4} M chondroitin and were present in the perfusion zone as soon as the odor arrived, with very high PIs (PI = 0.6 / 0.7 / 0.6 / 0.5; Wilcoxon test, p = 0.0004 at all times). The attraction was persistent, since fish remained in the odor zone during the water perfusion that followed. Lab-SF showed a comparatively more modest response to 10^{-4} M chondroitin (PI = 0.27; Wilcoxon test, $p = 0.008$), but a strong attraction for the higher chondroitin concentration of 10^{-3} M (PI = 0.52; Wilcoxon test, $p = 0.0004$). Finally, the Lab-Hyb were present in the odor zone during the first half of the chondroitin perfusion (PI = 0.29 and 0.1; Wilcoxon test, $p = 0.0004$ and 0.008) but then seemed to avoid the area (PI = −0.17; Wilcoxon test, $p = 0.0004$). Thus, all morphotypes/genotypes were attracted by chondroitin, but persistence and intensity of the response was particularly spectacular with Lab-Pachón (Video S1). Overall, these laboratory experiments showed that our setup allowed us to measure olfactory responses on adult *A. mexicanus* in a reliable manner.

3.3. Responses to Odors in Caves, in Wild A. Mexicanus Cavefish Populations

Figure 3. Sampling cavefish olfactory skills in their natural environment; (**A**) Simplified map of the region of Ciudad Valles, Mexico. Mountain ranges are in grey. The locations of visited caves are indicated by colored circles. The color code indicates the geographical group where they belong; (**B**) Fieldwork in the Pachón cave. The main pool was muddy and the water level was low. Note that in March 2017 and March 2019, the three plastic pools were installed and processed in the best possible reproductive manner, on the "beach" along the main pool. Close-ups show the Luer-lock perfusion system and the eight fish installed in one of the experimental plastic pools; (**C**) Work in pool 2 of the Subterráneo cave, located just after the 3-m pit. This pool offers a rocky substrate. There, we repeatedly observed the presence of crayfishes (predators on adult cavefish). The bottom right panel shows the infrared, hunting-type camera, used in our 2019 campaign; (**D**) Work in Los Sabinos cave. Pool 1 of this cave had crystal-clear water. Again, note that olfactory setups were installed in the same place in March 2017 and 2019, respectively, i.e., on a rocky plateau just above pool 1. A close-up shows the Raspberry Pi camera used during our 2017 campaign. (**E**) Testing olfaction on fish from

the Chica superficial pool. The top photos show the guano slope leading to the superficial pool, which was very rich in organic material (decomposing bat cadaver). On the bottom left panel, note the large size of the eight fish (compare with the equivalent picture in Pachón, and see Figure 4).

3.3.1. Pachón Cave

Despite the thick mud, it was easy to install three plastic pools near the natural main pool (Figure 3B). CF-Pachón were of relatively small size (mean: 4.6 cm) compared to other caves studied (Figure 4A, purple) and they were all eyeless and depigmented (Figure 4D). Fish used for olfaction experiments came from the small lateral pool in 2017 and from the main pool in 2019. These two pools communicate with each other when the water level is high and the fish can swim and mix between these two pieces of water. There was no difference in size (or phenotype) between the fish tested during the two expeditions (Figure 4B, purple), suggesting that the local conditions have been stable across the years in the cave. The fauna encountered there and the feeding habits of the CF-Pachón have been described by Espinasa et al., 2017 [23].

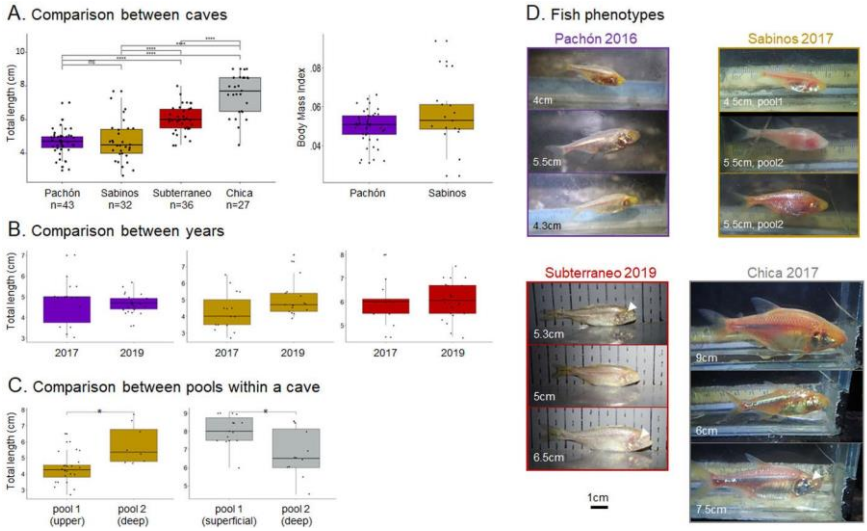


Figure 4. Cavefish sizes and phenotypes in different caves; (**A**) Sizes (left graph) and body mass indexes (BMI, right graph) of the cavefish individuals tested for olfactory responses in the Pachón (purple), Sabinos (brown), Subterráneo (red), and Chica (grey) caves. Asterisks indicate significant differences (Mann–Whitney test); (**B**) Comparison of sizes of the cavefish tested for olfactory responses in different field campaigns, in 2017 and 2019, showing that the condition of the fish tested did not vary; (**C**) Comparison of sizes of cavefish in different water ponds in a same cave locality, showing that fish condition varied, probably due to local trophic, environmental and/or genetic parameters; (**D**) Fish phenotypes are fully troglomorphic in the Pachón and Sabinos caves, whereas hybrid-type fish with small eyes (arrowheads) and some pigmentation can be seen in Subterraneo (pool 2) and Chica (superficial pool) caves. They result from hybridization with introgressed surface fish. Of note, in Subterraneo, surface fish enter by the cave entrance at the polje level, whereas, in Chica, surface fish enter the cave by the bottom. Hence, in both caves, the most troglomorphic fish are larger.

Despite fair conditions for installation of the olfactory setup, CF-Pachón from two experimental pools out of the six recorded presented an obvious place preference behavior (one in 2017 and one in 2019), for unknown reasons (Figure 5). These data were thus discarded from the analyses shown in Figure 6A.

Contrary to Lab-Pachón, CF-Pachón did not respond to alanine 10^{-7} M. Even more so, the preference index reached negative values during the perfusion. Conversely, CF-Pachón were strongly and significantly attracted by serine 10^{-7} M (PI = 0.41) and by chondroitin 10^{-4} M (PI = 0.55), as shown both by Friedman and Wilcoxon statistical tests (Figure 6A, green and blue). Moreover, the type of behavioral response elicited by chondroitin in the field was similar to the food-seeking behavior recorded in the lab (Video S2). Although slightly surprising because the alanine and serine responses seem divergent between the Lab-Pachón and the CF-Pachón, these data confirm, in the field, that CF-Pachón have excellent olfactory detection skills.

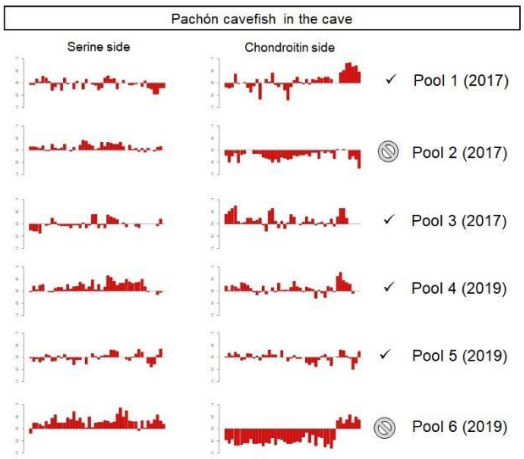


Figure 5. The place preference problem during cave experimentation. Graphs showing the position of the fish along the whole ~1 h protocol for the n = 6 experiments performed in the Pachón cave. The bars represent the preference of fish toward the serine perfusion side (left column) or the chondroitin perfusion side (right column). In theory, the eight fish should swim across the arena and distribute or explore randomly except for the response to the considered odor. However, in some instances (here, plastic pools 2 and 6), it was not the case: The fish remained in the same part of the plastic pool throughout the experiment, showing place preference for unknown reasons. These data were excluded from analysis.

3.3.2. Sabinos cave.

The first fish pool encountered, i.e., the upper pool, lies under a large arch and contains shallow water that disappears under the rock (Figure 3D). Exploration there revealed a large (20–30 m), continuous body of water under a low ceiling, hosting a significant cavefish population. The water was clear with a muddy substrate. Juvenile cavefish (size 1–2 cm) were observed, indicating that reproduction occurs. Large isopods and numerous mysid shrimps were present. Just above this natural pool, a natural rocky stage allowed us to install three plastic setups in the same place in 2017 and 2019 (Figure 3D). From there, a corridor led into a large and relatively low room where a second fish pool, or bottom pool, was encountered. Bats were numerous (observed in 2013, 2017, and 2019). The air was loaded with spores and the soil was covered with patches of microorganism covered with insects. CO_2 was not measured but was probably high.

All fish in the Sabinos cave had a fully troglomorphic phenotype (Figure 4D). Fish used in the experiments were sampled from both the upper and the bottom pools. Like in the Pachón cave, their size and body mass index was modest (Figure 4A, brown; mean: 4.7 cm) and did not vary between 2017 and 2019 samplings (Figure 4B, brown), suggesting that the two experimental series were performed on fish of similar condition. We noted that CF-Sabinos found in the deeper pool were significantly larger (Figure 4C, brown) and more corpulent (Body Mass Index, Wilcoxon test, $p = 0.006$) than those fished in the upper pool, suggesting that they were older [24] and/or better fed. However, this observed difference in condition could

not be correlated to their olfactory responses. As in the Pachón cave, we observed a strong place preference bias in two experiments out of six, hence these data were discarded (Figure S2). When tested for olfactory responses, CF-Sabinos were attracted neither by alanine nor by serine (Figure 6B). However, they did spend time in the odorous part of the arena after chondroitin perfusion (PI values between 0.46 and 0.28; Wilcoxon test, $p = 0.02$; but Friedman test, NS), suggesting that they were, although moderately, attracted to this molecule (Figure 6B).

3.3.3. Subterráneo Cave

A 20 min descent in a boulder tunnel filled with organic waste (including of very large size like trees) carted inside during the rainy season leads to a first small fish pool, where almost all individuals are surface-like (*A. mexicanus*, poecilids, some cichlids, presumably washed in during flooding) and where troglomorphic fish are rare (see Simon et al., 2017 [24]). Then, the tunnel continues and leads to a small 5-m pit which descends directly into the natural pool 2, with clear water on a rocky substrate, where CF-SubT swim. All fish used and measured in 2013 [14], 2016, and 2019 came from this room (Figure 3C). The ground of the cave at this level was relatively horizontal but it was covered with pebbles. As there was no mud to flatten the ground, we installed plastic pools on small flat mounds (Figure 3C).

As we previously described [14], CF-SubT presented mixed phenotypes in terms of eye size and pigmentation as a result of hybridization with introgressed surface fish, and most of them were not fully troglomorphic (Figure 4D). We found that CF-SubT were significantly larger than CF-Pachón and CF-Sabinos (Figure 4A, red; mean: 6 cm), suggesting that they may have been overall slightly older [24] and/or in better nutritional condition. In addition, the CF-SubT tested in 2017 and 2019 were of similar sizes (Figure 4B, red). During the olfaction tests, the CF-SubT were present in the 10^{-7} M alanine perfusion area approximately 2 min after the odor arrived and they remained in this zone during the entire perfusion time (PI = 0.25/0.26/0.19; Wilcoxon test, $p = 0.007/0.118/0.025$) (Figure 6D). Conversely, they did not significantly respond to serine. Finally, CF-SubT were attracted by chondroitin (but note that the concentration perfused was higher than in other caves: 10^{-3} M), since they stayed in the odorous area with high preference indexes (PI between 0.48 and 0.33; Wilcoxon test, p = 0.025/0.007). Thus, overall, CF-SubT showed significant olfactory skills and responses in their natural settings. Of note, in videos, it was impossible to make any correlation between olfactory responses and the degree of troglomorphism exhibited by individual CF-SubT.

3.3.4. Chica Cave

The Chica superficial pool (or pool 1) hosting cavefishes is a large body of water with a guano slope bank, probably very rich in organic content due to a large bat colony and influx of organic materials from the surface during the rainy season (decaying bats, vegetal debris were observed; Figure 3E). A corridor with a flat rocky floor allowed us to set up experiments just above this superficial pool (Figure 3E). Further down, Chica pool 2 could be reached by following the underground river, which cascaded after a pit [11], and also contained fish that were phenotypically less troglomorphic and more surface-like or hybrid-like than the fish in superficial pool 1 (not shown).

The superficial pool CF-Chica tested were not fully troglomorphic, i.e., some individuals showed some degree of pigmentation and had tiny to small eyes (Figure 4D), probably as a result of hybridization with surface fish. CF-Chica were, by far, the largest and most corpulent fish that we tested among the different caves visited (Figure 4A, grey and 4D). The individuals fished for the olfaction experiments originated from the superficial pool, where the most extreme sizes (mean: 8.1 cm, max: 9 cm) were encountered (compare with CF-Chica from pool 2; mean: 6.8 cm, Figure 4C, grey). In this cave, olfaction experiments were performed once, in 2017. Hence, only $n = 3$ plastic pools were recorded. For the CF-Chica, Friedman tests indicated that the Preference Indexes did not vary between the perfusions of water and amino acids, and the Wilcoxon tests indicated that they were not different from 0 (Figure 6C). A video problem at the end of one recording resulted in $n = 2$ for chondroitin, for which we could not perform statistical tests. However,

for chondroitin, the pattern was flat along the whole perfusion sequence, and there was no such behavioral foraging response as visually observed in other caves. These results showed that CF-Chica did not respond to the odors used in our tests, including chondroitin that elicited strong responses in all other cavefishes tested as well as in Lab-SF and Lab-Hyb.

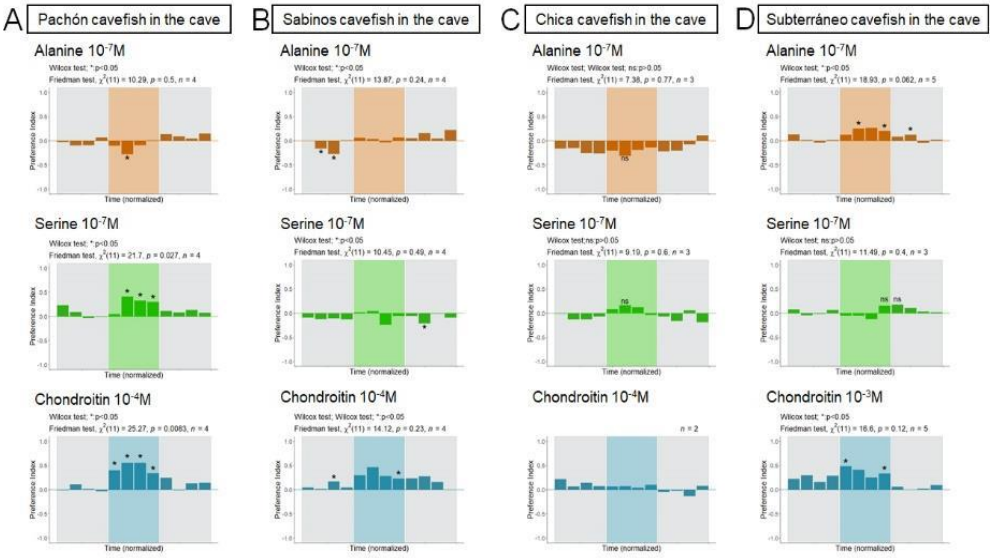


Figure 6. Olfactory responses of cavefishes in natural cave settings; (**A**), (**B**), (**C**), (**D**) Olfactory responses recorded in the Pachón (**A**), Sabinos (**B**), Chica (**C**), and Subterráneo (**D**) caves; Odors and concentrations are indicated, as well the number of replicates. The preference index (PI; positive values suggest attraction) is shown as a function of time. Time intervals corresponding to odor perfusion are shaded. Asterisks on graphs indicate significance as compared to no response (i.e., PI = zero) for a given time interval (Mann–Whitney test). The results of Wilcoxon and Friedman tests to probe the significance of the response across time with repeated measures are also indicated.

3.3.5. Tinaja Cave

The first, very small, Tinaja "perched pool" was encountered after a 20-min walk and consisted of a small piece of crystal-clear water retained between rocks (Figure 7A). We found fully troglomorphic CF-Tinaja there (approximately n = 20–30) in two consecutive years (2016 and 2017). After another 20 min of hiking down, the tunnel narrows before reaching a fault, at the bottom of which lies Traverse Lake (or Tinaja pool 1) (Figure 7B). A small natural beach allowed for experimental installation, but the space was small and inconvenient (~3 m wide). Drops of water constantly falling from the ceiling generated a very thick and sticky layer of mud. The water was cloudy. In such conditions, it was difficult to install the olfaction setup. This was partly achieved in March 2016, with only n = 2 plastic pools and with video recordings performed in the light (Figure 7B). The size of the fish was 6.1 cm on average (n = 12 measured). In 2017, we aimed to complete the Tinaja study and we performed n = 1 additional olfactory test directly in the small natural "perched pool 1." Although we cannot provide statistical support (n = 2 and n = 1, respectively), the video analyses suggest that that CF-Tinaja were not attracted by any of the three perfused odors (data not shown).

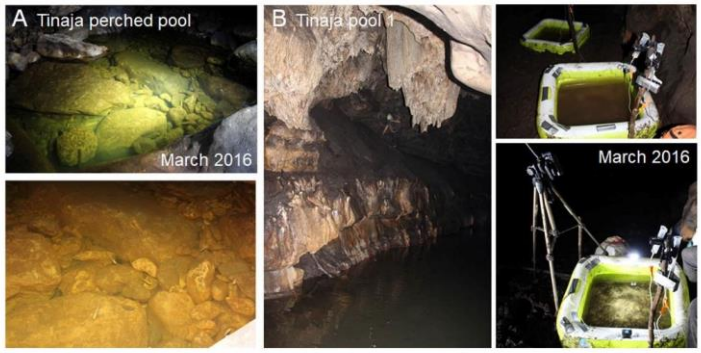


Figure 7. Experimentation in the Tinaja cave; (**A**) single test was "manually" performed in Tinaja perched pool, a very small (~3–4 m^2) and shallow natural pool where water is retained between rocks. There, over the years, we repeatedly observed 20–30 fully troglomorphic fish, presumably trapped there during the rainy season. The water is crystal clear; (**B**) Preliminary experiments were performed in March 2016 in Tinaja pool 1 (called Traverse Lake), a relatively large and muddy water reservoir. At the time, films were recorded in the light (instead of in the dark with infrared) and the perfusion system and the odorant molecules perfused were still under tuning. This type of expedition was of paramount importance to improve and establish the final experimental setup and design we used successfully in 2017 and 2019 in other caves.

4. Discussion

4.1. Testing Cavefish Olfaction in the Field: A Challenge

To our knowledge, this is the first study reporting systematic, precise, and quantified olfactory responses to specific odorant molecules in behavioral tests with subterranean animals in their natural environment (but see [14,22]). The trip organization, transport of equipment, and difficulties of access, together with the struggle of setting up experiments in "natural laboratories" which were different each time and handling the unexpected, means that few researchers risk of this kind of experimentation. The interpretation of the results must also be approached with caution because the experimental parameters are far from being fully mastered or even known. Of note, we found that establishing and pretesting of protocols in the lab are mandatory steps, and the comparison between lab and field results can help reach conclusions. Finally, this type of field behavior work needs to extend over the long term, with continuous improvements and feedbacks. For example, the difficulties we encountered in Tinaja in 2016 allowed us to improve the setup. From that expedition, we concluded that the plastic pools needed to be perfectly clean and the cave water needed to be filtered before being poured into the arena or else the videos would be difficult to interpret, that we should not use rechargeable batteries for infrared lamps because the Mexican electrical network sometimes fluctuates, that we should systematically include a water counterflow at the corner opposite to odor perfusion to remove biases of lateral line driven behaviors, and that we should not plan for 2-day-long experiments in a cave where access is difficult. Hence, our Tinaja 2016 experiments (n = 2 only) were not performed in the dark, did not have the same perfusion sequence nor the same odor concentrations, and were not included in the main results of the present paper, but they ended up being very instructive for improving the setup and protocols, and they allowed us to draw preliminary conclusions (see Figure 7).

4.2. Alanine, Serine and Chondroitin Elicit Variable Behavioral Responses in Wild and Lab-Raised Adult A. Mexicanus.

The choice of odors presented to fish was critical. Whereas alanine and serine amino acids are "classically" used in fish olfaction experiments because they correspond to food degradation products and therefore serve as food-related attractive cues (e.g., [38–40]), the use of chondroitin is more novel.

Chondroitin sulfate is a glycosaminoglycan composed of a chain of alternating sugars and a component of the cartilage matrix. It is also present in zebrafish skin mucus, where it was recently discovered to serve for the long-searched active molecule of the alarm substance [27,41]. Indeed, zebrafish strongly react to chondroitin with a typical alarm behavior, including erratic swimming and freezing, and the activated brain regions are the same as those activated after presentation of skin extracts [27]. In *A. mexicanus*, it was therefore unexpected that chondroitin would elicit (1) attraction and (2) foraging, and even more so, (3) the same response in the two morphs. Along with the idea that the alarm substance is species-specific [42], our findings suggest that its composition in zebrafish and *Astyanax* must be markedly different. Moreover, early studies have suggested that alarm reaction is indeed present in surface fish but mostly lost in cavefishes, at least those originating from the Pachón and Chica caves [28,29,43]. Here, surface fish and Pachón cavefish strongly reacted to chondroitin, but Chica cavefish did not, further suggesting that this molecule is not part of the alarm substance in the species. Rather, it seems to correspond to a foraging cue, which is in line with *Astyanax* being carnivorous, eating carcasses and responding to cartilage odors. In fact, we ended up considering chondroitin as a sort of positive control: All types of fish tested, in the lab and in the field (except CF-Chica, see below), adopted an intense food search behavior, almost leaking or sucking the extremity of the tube where chondroitin flew in (Supplementary videos). Of note, chondroitin was the third and last odor presented in the protocol because we reasoned that the more modest responses elicited by amino acids may be lost after the strong stimulation caused by chondroitin. Finally, response to chondroitin also validated, in a way, the whole test, and we were confident that absence of response to amino acids sometimes observed was true because fish reliably and repeatedly responded to chondroitin at the end of the behavioral assay (Table 1).

Table 1. A summary of the results obtained in all the cave and laboratory olfaction experiments. Attractive response (+) or no response (0) is summarized for each cavefish or surface fish population, when recorded in the field or in the lab (as indicated in the different columns) for the 3 different odors studied (as indicated in the different rows; orange/alanine, green/serine, blue/chondroitin). Light and dark colors indicate response to the high concentration (light color, lower detection capacities) or the low concentration (dark color, better detection capacities) of the odor considered. The last line (grey shades) indicates to how many odors, out of the 3 tested, the fish have responded.

Fishes / Attraction : (+) / No response : (0)	Cave fishes					Surface fishes		
	Lab-Pachón	CF-Pachón	CF-SubT	CF-Sabinos	CF-Chica	Lab-SF	Lab-SF	Lab-Hyb
Alanine								
10^{-5}M	+	0	+	0	0	+	0	0
10^{-7}M								
Serine								
10^{-5}M	+	++	0	0	0	0	0	0
10^{-7}M								
Chondroïtin								
10^{-3}M	+++	+++	++	+	0	++	+	+
10^{-4}M								
Olfactory performance summary	3/3	2/3	2/3	1/3	0/3	2/3	1/3	1/3

Regarding amino acids, the results we obtained with adult fish in the lab globally corroborate what has already been observed with 1-month-old larvae [20]: The adult Lab-SF had a lower olfactory detection threshold than the Lab-Pachón (Table 1). The results confirm that older and larger adult fish do not have a better detection threshold than 1 month-old-larvae [19], reinforcing the idea that olfactory detection threshold is not directly or exclusively linked to the size of the olfactory organs, or to the intrinsic properties of the olfactory system. The results also confirm that the olfactory system of the 1-month-old larvae tested in our previous studies was functionally mature.

For reasons explained earlier, adult surface fish (and F1 hybrids) could only be tested in the lab. We considered their olfactory performance to be poor (Table 1). They did respond to chondroitin-albeit less intensely than Lab-Pachón, but barely responded to amino acids, even at high concentrations. An explanation to this mitigated result might be that the olfactory response is slightly blurred by schooling behavior, even in the dark [25,44]. Indeed, for these fish, we observed transient statistically significant PI scores (positive or negative) before or after odor perfusion (see Figure 2B–C) that cannot correspond to true and specific olfactory responses, and that probably resulted from the grouping of all eight fish in the same area of the arena at some time points of some experiments. These observations further confirm the appropriateness of using 1-month-old surface fish—which do not yet school intensely—for olfactory tests with these morphotypes. Of note, such an interference of schooling in olfactory responses is irrelevant in the case of cavefish who have lost schooling behavior, as first observed by Parzefall in the field [22].

Finally, the comparative responses of Pachón cavefish from the lab and from the field to amino acids is also worth discussing. Lab-Pachón responded positively (= attraction, like larvae) to alanine, whereas CF-Pachón responded negatively (= repulsion). Moreover, Lab-Pachón response to serine was weak or delayed (if any), whereas CF-Pachón response was strong and immediate. Thus, both lab-raised and wild animals originating from the Pachón cave seem able to detect very low concentrations of amino acids (here, 10^{-7} M), thereby confirming their excellent olfactory skills. However, unexpectedly, their responses to a given odor can differ in nature or intensity. The first hypothesis we can draw relates to the distinct environment in which the lab and wild fish were grown and live, e.g., the water parameters or diet could be at the origin of their difference in reaction. Indeed, the water used in the lab tests (tap water bubbled for 24 h) is "cleaner" than the water used in the field, which, although filtered, is loaded with dissolved organic compounds. It has also been proposed that fish move toward or away from a given amino acid depending on the concentration as well as the age and the species of fish. For example, cysteine acts as a repellant for 1-month-old juvenile zebrafish (3×10^{-5} M; [45]) but as an attractant for rainbow trout (10^{-6} M; [39]). Moreover, olfactory conditioning or learning seems to play an important role in the expression of behavioral responses [46,47]. In the Pachón cave, juveniles feed on micro-arthropods and adults feed on decomposing organic debris and mud [23]. Conversely, in the lab, larvae are fed with micro-worms and artemia nauplii, and juveniles and adults are fed with granules that contain by-products of fish, cereals, vegetables, and crustaceans. Lab-raised fish eat very little decomposing food since they finish their meals in a few minutes. Such differences in feeding experience and odor exposure might explain the differential responses of Lab-Pachón and CF-Pachón. The repulsive behavior after alanine perfusion in the natural cave may suggest that CF-Pachón are exposed to an unpalatable organic matter that gives off alanine. We do not know the amino acids compositions of different organic materials found in this cave. It is therefore difficult to determine which one might be the source of this repulsion. Overall, the sometimes-contrasting responses that we observed in the lab and in the wild reinforce the interest and necessity to assess behavioral repertoires in the natural environment to discuss ecological or evolutionary relevance.

4.3. Not All Cavefish Respond the Same

Our field observations in five different caves lead to one main conclusion: Cavefish from different caves showed very distinctive responses when presented with exogenous odor cues at low concentrations (Table 1). For those that did respond (CF-Pachón, CF-SubT), we can propose that they have augmented olfactory skills when compared to their surface conspecifics, which confirms our previous studies [14,20]. For those which did not respond as well (CF-Sabinos) or did not respond at all (CF-Chica, CF-Tinaja), several questions arise. Is the lack of response due to local environment, previous experience and/or fish condition and nutritional status? Or is it due to experimental conditions?

(1) Let us start with the latter—perhaps less biologically interesting—hypothesis. We cannot rule out that CF-Chica did not respond because the cave water used (although filtered) was so loaded with debris and dissolved organic materials and endogenous olfactory cues that their olfactory

system was saturated and could not detect the low concentrations of amino acids or even chondroitin perfused (see Figure 3E). The same holds true in the Tinaja cave, where carbon content in the mud is high (32%, to be compared to 9.2% in the main pool of Pachón) [24], and, maybe to a lesser extent, in the Sabinos cave (see description in Results). A way to test directly this hypothesis would be to transport "clean" tap water into the caves to perform the experiments. However, we feel that the interest of such experiments would be rather limited, as we anticipate that they would be flawed with other problems and quite demanding in terms of logistics. Nevertheless, future studies examining the physicochemical parameters and the exact nature of compounds present in the water will be important to answer these questions.

(2) The environment-dependent and experience-dependent hypothesis is more scientifically exciting to discuss adaptation to the environment. It implies that the expression of olfactory skills depends on local environment, such as the configuration and ecological parameters of the cave and previous olfactory experiences of the fish related to these local environmental parameters, including possible interactions with other sensory systems and fish nutritional status and motivation to find food.

Indirect support for this hypothesis comes from our finding that CF-Pachón and CF-SubT, which do not belong to the same geographical group of caves and correspond to independently evolved cavefish populations [5,11,48], both show significantly augmented olfactory skills. Thus, this trait most probably corresponds to an evolutionary convergence. This result further suggests that the evolution of food-related (and maybe pheromonal, not addressed here) odor sensing is of paramount importance for cavefish life, adaptation, and survival in the dark. On the other hand, population genetic studies have shown that El Abra cave populations (Pachón, Sabinos, Tinaja, Chica) are genetically close and share many polymorphisms [25,48,49], suggesting that CF-Sab or CF-Tinaja probably carry all or part of the mutations in "olfaction genes" that genetically determine CF-Pachón olfactory skills. However, they do not necessarily express these skills as a consequence of plasticity, with expression depending on the environmental conditions at the precise time and location where we measured olfactory responses. The most striking pieces of evidences in favor of this possibility are discussed below.

CF-Pachón were the smallest among all fish assayed in this study, in agreement with previous measurements [24]. They live in a low-carbon cave (9.2% in the mud), with limited influx of organic matter from the surface. There, foraging must be challenging and strongly olfactory-driven. Conversely, CF-Chica were the largest and biggest among the cavefish that we tested and measured. The very large bat colonies in this "dirty" cave probably render foraging easy, together with fast growth and a long lifespan. CF-Tinaja are also large and old [24] and live in a carbon-rich (31.9%) environment [23]. Strikingly, and in line with these very different trophic environmental conditions, CF-Pachón and CF-Chica (and CF-Tinaja, according to preliminary results) were at the two ends of the spectrum of olfactory responses that we observed.

Vibration attraction behavior (VAB) is thought to help cavefish locate food droppings or vibrating objects at the surface of the water [18,50]. VAB is mediated by neuromasts, and therefore corresponds to a mechanosensory modality. In the field, VAB responses are variable in different caves and even in different ponds of a single cave [51]. In cave ponds where insects are numerous or where percolating water drops are abundant, food search is probably strongly guided by VAB, and less so by the chemosensory olfactory modality. Our results fit well with this idea. Indeed, in the Tinaja cave, where it literally "rains," or in the Sabinos cave, where we personally witnessed abundant water drops over pool 1 to which the CF-Sab were systematically attracted (SR and A. Alié, pers.obs.), the olfactory sense and skills may not be solicited as much as in the Pachón or Subterráneo caves, where vibrating objects are much rarer and food must be found by the nose. Thus, we propose the possibility of a balanced use of different sensory modalities and expression of sensory skills depending on local conditions. Of note, fish may preferentially use one or the other (or both) sense across their lifespan to find a proper diet and to adapt to different locations where they can swim in a cave, or even across seasonal fluctuations of local ecological conditions.

Finally, cavefish genetics and the hybridization with introgressed surface fish is also a parameter to consider. Whereas Pachón, Sabinos, and Tinaja cavefish are fully troglomorphic (see Figure 4), CF-SubT and CF-Chica correspond to "hybrid-like" populations and carry surface fish alleles (see [14]). Interestingly, the entrance of surface individuals into these two caves does not proceed by the same end: In Subterráneo, surface fish are washed inside the cave from the surface by flooding, whereas, in Chica, surface fish enter the cave by a bottom resurgence. This probably explains why, in Chica, the pool 2 (deeper) fish, which carry more surface alleles, are smaller than the pool 1 (upper) fish, which are less hybridized and more troglomorphic and cave-adapted. In any case, we cannot exclude that the important gene flow from the surface that exists in Chica [25,48] might counteract the effects of cave alleles, favoring the evolution of augmented olfactory skills. Hence, the absence of olfactory responses in CF-Chica might correspond to truly modest olfactory capacities. Deciphering the effects of genetics (presence of surface alleles) from the effects of environmental parameters on plasticity (food-rich cave) or from experimental conditions (odor-saturated water during test) is impossible at this stage. However, it must be noted that the CF-SubT, which also carry surface alleles, responded much better than CF-Chica to the odors we presented: They strongly responded to both alanine and chondroitin. Therefore, in this later case, the effects of hybridization and introgression of surface alleles appears very limited on the behavioral phenotype.

5. Conclusion

Almost 80 years have passed since the first descriptions of ecological conditions in which Chica cavefish live by Charles Breder [52]. Since then, fieldwork in *Astyanax* caves has continuously brought novel information on caves topography, population genetics, and cavefish biology in general. Recently, important examples of insights on the evolution of cavefish behaviors or the evolution of their immune system started from field observations [53,54]. Here, by bringing the "behavior room" into the field, we highlighted the diversity and complexity of the mechanisms that underlie cave adaptation and documented the diversity of cave environments inhabited by a single species of cavefish, *A. mexicanus*. Our data confirm the classical proposal of sensory compensations to the absence of vision in the dark. Indeed, overall cavefish showed lower (i.e., better) olfactory detection thresholds than surface fish. However, the picture appears more complicated than simply "cavefish smell better." The next challenge will be to disentangle the effects of genetics, plasticity, environment, and their interactions in the evolution of cavefish olfactory system. Then, of course, more fieldwork will be needed to refine behavioral observations and to further describe the diverse natural ecological conditions in which cavefish live.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, Figure S1: Rationale for olfactory scoring. Figure S2: Place preference problem in the Sabinos cave. Video S1: Behavioral response of Pachón cavefish to chondroitin in laboratory settings. Video S2: Behavioral response of Pachón cavefish to chondroitin in natural settings.

Author Contributions: Conceptualization, M.B. and S.R.; methodology, M.B., J.F. and S.R.; validation, M.B. and S.R.; formal analysis, M.B. and S.R.; investigation, all authors; data curation, M.B.; writing—original draft preparation, S.R. and M.B.; writing—review and editing, S.R.; visualization, M.B. and S.R.; supervision, S.R.; project administration, S.R.; funding acquisition, S.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by CNRS, ANR (Agence Nationale pour la Recherche) grant [BLINDTEST], Fondation pour la Recherche Médicale Equipe FRM grant [DEQ20150331745] to SR, and a French-Mexican Ecos-Nord Exchange Program grant [M15A03] to SR and Patricia-Ornelas Garcia (UNAM, Mexico). The APC was funded by CNRS.

Acknowledgements: We thank Laurent Legendre, Luis Espinasa, Didier Casane, Carole Hyacinthe, Victor Simon, Eric Queinnec, Karen Pottin, Stéphane Rode, Joel Attia and all other past and present members of the Rétaux's lab for their invaluable help and team spirit during fieldwork over the years. We are grateful to Krystel Saroul for taking care of our lab Astyanax colony.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Culver, D.C.; Pipan, T. *The Biology of Caves and Other Subterranean Habitats*; Biology of Habitats Series; 2nd ed.; Oxford University Press: Oxford, UK, 2019. ISBN 978-0-19-882077-2.
2. Rétaux, S.; Casane, D. Evolution of eye development in the darkness of caves: Adaptation, drift, or both? *EvoDevo* **2013**, *4*, 26, doi:10.1186/2041-9139-4-26.
3. Romero, A. Cave Biology: Life in Darkness. *Integr. Comp. Biol.* **2010**, *50*, 689–691, doi:10.1093/icb/icq067.
4. Simon, V.; Hyacinthe, C.; Rétaux, S. Breeding behavior in the blind Mexican cavefish and its river-dwelling conspecific. *PLoS ONE* **2019**, *14*, e0212591, doi:10.1371/journal.pone.0212591.
5. Mitchell, R.; Russell, W.; Elliott, W. Mexican eyeless characin fishes, genus *Astyanax*: Environment, distribution, and evolution. *Spec. Publ. Mus. Texas Tech. Univ.* **1977**, *12*, 1–89.
6. Elliott, W.R. Cave Biodiversity and Ecology of the Sierra de El Abra Region. In *Biology and Evolution of the Mexican Cavefish*; Elsevier: Amsterdam, The Netherlands, 2016; pp. 59–76. ISBN 978-0-12-802148-4.
7. Casane, D.; Rétaux, S. Evolutionary Genetics of the Cavefish *Astyanax mexicanus*. *Adv. Genet.* **2016**, *95*, 117–159, doi:10.1016/bs.adgen.2016.03.001.
8. Jeffery, W.R. Chapter 8 Evolution and Development in the Cavefish *Astyanax*. In *Current Topics in Developmental Biology*; Elsevier: Amsterdam, The Netherlands, 2009; Volume 86, pp. 191–221. ISBN 978-0-12-374455-5.
9. Keene, A.; Yoshizawa, M.; Mcgaugh, S.E. *Biology and Evolution of the Mexican Cavefish*; Elsevier Science: Amsterdam, The Netherlands, 2015. ISBN 978-0-12-802148-4.
10. Espinasa, L.; Legendre, L.; Fumey, J.; Blin, M.; Rétaux, S.; Espinasa, M. A new cave locality for Astyanax cavefish in Sierra de El Abra, Mexico. *Subterr. Biol.* **2018**, *26*, 39–53, doi:10.3897/subtbiol.26.26643.
11. Elliott, W.R. *The Astyanax Caves of Mexico Cavefishes of Tamaulipas, San Luis Potosí, and Guerrero*; Association for Mexican Cave Studies, PO Box 7672, Austin, Texas 78713, USA: 2018.
12. Şadoğlu, P. A mendelian gene for albinism in natural cavefish. *Experientia* **1957**, *13*, 394, doi:10.1007/BF02161111.
13. Borowsky, R. Restoring sight in blind cavefish. *Curr. Biol. CB* **2008**, *18*, R23–R24, doi:10.1016/j.cub.2007.11.023.
14. Bibliowicz, J.; Alié, A.; Espinasa, L.; Yoshizawa, M.; Blin, M.; Hinaux, H.; Legendre, L.; Père, S.; Rétaux, S. Differences in chemosensory response between eyed and eyeless Astyanax mexicanus of the Rio Subterráneo cave. *EvoDevo* **2013**, *4*, 25, doi:10.1186/2041-9139-4-25.
15. Fumey, J.; Hinaux, H.; Noirot, C.; Thermes, C.; Rétaux, S.; Casane, D. Evidence for late Pleistocene origin of Astyanax mexicanus cavefish. *BMC Evol. Biol.* **2018**, *18*, 43, doi:10.1186/s12862-018-1156-7.
16. Policarpo, M.; Fumey, J.; Lafargeas, P.; Naquin, D.; Thermes, C.; Naville, M.; Dechaud, C.; Volff, J.-N.; Cabau, C.; Klopp, C.; et al. Contrasting gene decay in subterranean vertebrates: Insights from cavefishes and fossorial mammals. *Mol. Biol. Evol.* **2020**, doi:10.1093/molbev/msaa249.
17. Yamamoto, Y.; Byerly, M.S.; Jackman, W.R.; Jeffery, W.R. Pleiotropic functions of embryonic sonic hedgehog expression link jaw and taste bud amplification with eye loss during cavefish evolution. *Dev. Biol.* **2009**, *330*, 200–211, doi:10.1016/j.ydbio.2009.03.003.
18. Yoshizawa, M.; Gorički, Š.; Soares, D.; Jeffery, W.R. Evolution of a Behavioral Shift Mediated by Superficial Neuromasts Helps Cavefish Find Food in Darkness. *Curr. Biol.* **2010**, *20*, 1631–1636, doi:10.1016/j.cub.2010.07.017.
19. Blin, M.; Tine, E.; Meister, L.; Elipot, Y.; Bibliowicz, J.; Espinasa, L.; Rétaux, S. Developmental evolution and developmental plasticity of the olfactory epithelium and olfactory skills in Mexican cavefish. *Dev. Biol.* **2018**, *441*, 242–251, doi:10.1016/j.ydbio.2018.04.019.
20. Hinaux, H.; Devos, L.; Blin, M.; Elipot, Y.; Bibliowicz, J.; Alié, A.; Rétaux, S. Sensory evolution in blind cavefish is driven by early embryonic events during gastrulation and neurulation. *Development.* **2016**, *143*, 4521–4532, doi:10.1242/dev.141291.
21. Torres-Paz, J.; Hyacinthe, C.; Pierre, C.; Rétaux, S. Towards an integrated approach to understand Mexican cavefish evolution. *Biol. Lett.* **2018**, *14*, doi:10.1098/rsbl.2018.0101.
22. Parzefall, J. Field observation in epigean and cave populations of the Mexican characid *Astyanax mexicanus* (Pisces, Characidae). *Mémoires de biospéologie* **1983**, *10*, 171–176.
23. Espinasa, L.; Bonaroti, N.; Wong, J.; Pottin, K.; Queinnec, E.; Rétaux, S. Contrasting feeding habits of post-larval and adult Astyanax cavefish. *Subterr. Biol.* **2017**, *21*, 1–17, doi:10.3897/subtbiol.21.11046.

24. Simon, V.; Elleboode, R.; Mahé, K.; Legendre, L.; Ornelas-Garcia, P.; Espinasa, L.; Rétaux, S. Comparing growth in surface and cave morphs of the species *Astyanax mexicanus*: Insights from scales. *EvoDevo* **2017**, *8*, doi:10.1186/s13227-017-0086-6.
25. Pierre, C.; Pradère, N.; Froc, C.; Ornelas-García, P.; Callebert, J.; Rétaux, S. A mutation in monoamine oxidase (MAO) affects the evolution of stress behavior in the blind cavefish *Astyanax mexicanus*. *J. Exp. Biol.* **2020**, doi:10.1242/jeb.226092.
26. Strecker, U.; Bernatchez, L.; Wilkens, H. Genetic divergence between cave and surface populations of *Astyanax* in Mexico (Characidae, Teleostei). *Mol. Ecol.* **2003**, *12*, 699–710, doi:10.1046/j.1365-294x.2003.01753.x.
27. Mathuru, A.S.; Kibat, C.; Cheong, W.F.; Shui, G.; Wenk, M.R.; Friedrich, R.W.; Jesuthasan, S. Chondroitin fragments are odorants that trigger fear behavior in fish. *Curr. Biol. CB* **2012**, *22*, 538–544, doi:10.1016/j.cub.2012.01.061.
28. Pfeiffer, W. [On the heredity of the fear reaction in *Astyanax* (Characidae, pisces)]. *Zeitschrift fur Vererbungslehre* **1966**, *98*, 97–105.
29. Parzefall, J.; Fricke, D. Alarm reaction and schooling in population hybrids of *Astyanax* fasciatus (Pisces, Characidae). *Memoires e Biospeologie* **1991**, *18*, 29–32.
30. VideoLan. VLC Media Player. 2006. Available online: https://www.videolan.org/vlc/index.html (accessed on 1st February 2010).
31. Pérez-Escudero, A.; Vicente-Page, J.; Hinz, R.C.; Arganda, S.; de Polavieja, G.G. idTracker: Tracking individuals in a group by automatic identification of unmarked animals. *Nat. Methods* **2014**, *11*, 743–748, doi:10.1038/nmeth.2994.
32. R Core Team. R: *A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria. 2017. Available online: https://www.R-project.org/ (accessed on 1 February 2018).
33. RStudio Team. *RStudio: Integrated Development for R*; RStudio, PBC: Boston, MA, USA. Springer-Verlag: New York, NY, USA, 2020. Available online: http://www.rstudio.com/ (accessed on 1 February 2018).
34. Alboukadel Kassambara rstatix: Pipe-Friendly Framework for Basic Statistical Tests. R package version 0.5.0. 2020. Available online: https://CRAN.R-project.org/package=rstatix (accessed on 1 February 2020).
35. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*; Springer-Verlag New York, NY, USA, 2016. ISBN 978-3-319-24277-4. 2016. Available online: https://ggplot2.tidyverse.org (accessed on 1 February 2020).
36. Wickham, H.; Averick, M.; Bryan, J.; Chang, W.; McGowan, L.D.A.; François, R.; Grolemund, G.; Hayes, A.; Henry, L.; Hester, J.; et al. Welcome to the tidyverse. *J. Open Sour. Softw.* **2019**, *4*, 1686.
37. Elipot, Y.; Legendre, L.; Père, S.; Sohm, F.; Rétaux, S. Astyanax transgenesis and husbandry: How cavefish enters the laboratory. *Zebrafish* **2014**, *11*, 291–299, doi:10.1089/zeb.2014.1005.
38. Hara, T.J. Olfaction and gustation in fish: An overview. *Acta Physiol. Scand.* **1994**, *152*, 207–217, doi:10.1111/j.1748-1716.1994.tb09800.x.
39. Hara, T.J. Feeding behaviour in some teleosts is triggered by single amino acids primarily through olfaction. *J. Fish. Biol.* **2006**, *68*, 810–825, doi:10.1111/j.0022-1112.2006.00967.x.
40. Friedrich, R.W.; Korsching, S.I. Combinatorial and chemotopic odorant coding in the zebrafish olfactory bulb visualized by optical imaging. *Neuron* **1997**, *18*, 737–752, doi: 10.1016/s0896-6273(00)80314-1.
41. Frisch, K. Die Bedeutung des Geruchsinnes im Leben der Fische. *Süffert F Ed. Naturwissenschaften* **1941**, 321–333, doi:10.1007/BF01481736
42. Døving, K.B.; Lastein, S. The alarm reaction in fishes-odorants, modulations of responses, neural pathways. *Ann. N. Y. Acad. Sci.* **2009**, *1170*, 413–423, doi:10.1111/j.1749-6632.2009.04111.x.
43. Fricke, D. Reaction to Alarm Substance in Cave Populations of *Astyanax fasciatus* (Characidae, Pisces). *Ethology* **1987**, *76*, 305–308, doi:10.1111/j.1439-0310.1987.tb00691.x.
44. Kowalko, J. Utilizing the blind cavefish *Astyanax mexicanus* to understand the genetic basis of behavioral evolution. *J. Exp. Biol.* **2020**, *223*, doi:10.1242/jeb.208835.
45. Vitebsky, A.; Reyes, R.; Sanderson, M.J.; Michel, W.C.; Whitlock, K.E. Isolation and characterization of the laure olfactory behavioral mutant in the zebrafish, *Danio rerio*. *Dev. Dyn.* **2005**, *234*, 229–242, doi:10.1002/dvdy.20530.
46. Braubach, O.R.; Wood, H.-D.; Gadbois, S.; Fine, A.; Croll, R.P. Olfactory conditioning in the zebrafish (*Danio rerio*). *Behav. Brain Res.* **2009**, *198*, 190–198, doi:10.1016/j.bbr.2008.10.044.

47. Valentinčič, T.; Wegert, S.; Caprio, J. Learned olfactory discrimination versus innate taste responses to amino acids in channel catfish (*Ictalurus punctatus*). *Physiol. Behav.* **1994**, *55*, 865–873, doi:10.1016/0031-9384(94)90072-8.
48. Bradic, M.; Beerli, P.; García-de León, F.J.; Esquivel-Bobadilla, S.; Borowsky, R.L. Gene flow and population structure in the Mexican blind cavefish complex (*Astyanax mexicanus*). *BMC Evol. Biol.* **2012**, *12*, 9, doi:10.1186/1471-2148-12-9.
49. Avise, J.C.; Selander, R.K. Evolutionary genetics of cave-dwelling fishes of the genus astyanax. *Evol. Int. J. Org. Evol.* **1972**, *26*, 1–19, doi:10.1111/j.1558-5646.1972.tb00170.x.
50. Yoshizawa, M.; Jeffery, W.R. Evolutionary tuning of an adaptive behavior requires enhancement of the neuromast sensory system. *Commun. Integr. Biol.* **2011**, *4*, 89–91, doi:10.4161/cib.4.1.14118.
51. Espinasa, L.; Heintz, C.; Rétaux, S.; Yoshisawa, M.; Agnès, F.; Ornelas-Garcia, P.; Balogh-Robinson, R. Vibration Attraction Response (VAB) is a plastic trait in Blind Mexican tetra (*Astyanax mexicanus*), variable within subpopulations inhabiting the same cave. *J Fish. Biol, in press.* doi: 10.1111/jfb.14586.
52. Breder, C.M. Descriptive ecology of La Cueva Chica, with especial reference to the blind fish, Anoptichthys. *Zoologica* **1942**, *27*, 7–15.
53. Hyacinthe, C.; Attia, J.; Rétaux, S. Evolution of acoustic communication in blind cavefish. *Nat. Commun.* **2019**, *10*, 4231, doi:10.1038/s41467-019-12078-9.
54. Peuß, R.; Box, A.C.; Chen, S.; Wang, Y.; Tsuchiya, D.; Persons, J.L.; Kenzior, A.; Maldonado, E.; Krishnan, J.; Scharsack, J.P.; et al. Adaptation to low parasite abundance affects immune investment and immunopathological responses of cavefish. *Nat. Ecol. Evol.* **2020**, doi:10.1038/s41559-020-1234-2.

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

C- Article en premier auteur uniquement disponible en version « pre-print »

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.02.446701; this version posted June 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Article

Genomic evidence that blind cavefishes are not wrecks of ancient life

Maxime Policarpo[1], Patrick Laurenti[2,7], Erik García-Machado[3,4], Cushla Metcalfe[5], Sylvie Rétaux*[,6] and Didier Casane*[,1,7]

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

[2] Université de Paris, Laboratoire Interdisciplinaire des Energies de Demain, Paris, France

[3] Department of Biology, Institut de Biologie Intégrative et des Systèmes, Université Laval, 1030 Avenue de la Médecine, Québec City, Québec G1V 0A6, Canada.

[4] Centro de Investigaciones Marinas, Universidad de La Habana, Calle 16, No. 114 entre 1ra y 3ra, Miramar, Playa, La Habana 11300, Cuba.

[5] Independent Researcher, PO Box 21, Nambour QLD 4560, Australia.

[6] Université Paris-Saclay, CNRS, Institut des Neurosciences Paris-Saclay, 91190, Gif-sur-Yvette, France.

[7] Université de Paris, UFR Sciences du Vivant, F-75013 Paris, France.

* Corresponding authors: e-mails: sylvie.retaux@cnrs.fr; didier.casane@egce.cnrs-gif.fr

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.02.446701; this version posted June 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

Cavefishes often have modified eyes, from small but otherwise functional, to highly degenerate structures embedded in a connective tissue and covered by skin. Darwin assumed that these animals are 'wrecks of ancient life', but several genomic studies suggests they are not 'ancient'. The most radical dating shift is for populations of a Mexican cavefish, *Astyanax mexicanus*, that have been recently estimated to be at the most a few tens of thousands years old. Despite having highly degenerate eyes, the eye-specific genes of *A. mexicanus* have low levels of decay. Other blind cavefishes we have examined so far are even older, but also can be dated to the Pleistocene. Here, we estimated the age of blindness of two additional fish species by the level of decay of eye-specific genes. Many pseudogenes were identified in the amblyopsid *Typhlichthys subterraneus*, suggesting that blindness evolved a few million years ago. In contrast, the blind cichlid *Lamprologus lethops* appears to be a new case of very recent and rapid eye regression, which occurred in deep river water, an environment similar to caves. Genome-wide analyses support the hypothesis that blindness in cavefishes is never very ancient, and ranges from the Early Pliocene to Late Pleistocene. Together with the description of hundreds of cavefish species, our results suggest that surface fishes were able to recurrently and rapidly adapt to caves and similar small dark ecosystems but the resulting highly specialized blind species with a limited distribution may be evolutionary dead-ends in a relatively short time.

Key words: cavefishes, eye regression, eye-specific genes, pseudogenization, molecular dating.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.02.446701; this version posted June 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Introduction

In *On the origin of species*, Darwin wrote: “Far from feeling any surprise that some of the cave-animals should be very anomalous, as Agassiz has remarked in regard to the blind fish, the Amblyopsis, and as is the case with the blind Proteus with reference to the reptiles of Europe, 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” (Darwin 1859). Since the middle of the 19th century, thousands of obligate subterranean dwelling animals (troglobionts) have been described, most of them small arthropods (Culver and Pipan 2009; Romero 2009). Among the vertebrates, in addition to the cavefish, *Amblyopsis spelaea*, and the cave salamander, *Proteus anguinus*, mentioned by Darwin, many other troglobionts, primarily approximately two hundred and fifty named species of cavefishes (https://cavefishes.org.uk/), and at least 11 salamander species (Romero 2009) have been reported.

In the text quoted above, anomalous animals refers to troglomorphic (cave-related) characters, often present in cave animals, the two most conspicuous being highly degenerate eyes and depigmentation. Caves are very challenging ecosystems and surface animals have to adapt after settling in order to thrive in the long term. Interspecific competition is probably lower because there are fewer predators and other species that are competitors for food resources, but intraspecific competition is high in such a nutrient-limited environment (Poulson and White 1969). The depiction of cavefishes as wrecks of ancient life suggests that regressed traits are the result of relaxed selection, but it does not acknowledge that direct and indirect selection could also be involved in the evolutionary process. For instance, eye degeneration depends on mutations disrupting their development, but both selection and genetic drift could have been implicated in their fixation. In *Astyanax mexicanus* cavefish, a QTL analysis (Protas, et al. 2007) and indirect arguments based on the energetic cost of eyes (Moran, et al. 2015) suggest that selection for smaller eyes may have been involved in eye degeneration. Other analyses suggest indirect selection on eye size as the consequence of direct selection on other traits genetically correlated (Yamamoto, et al. 2009; Yoshizawa, et al. 2012; Hinaux, et al. 2016). In addition, if we assume that the first mutations reducing eye size were advantageous in the darkness of caves, other mutations may have accumulated later through genetic drift in small-eyed or blind fish (Retaux and Casane 2013). It has proven very difficult to disentangle these processes because they are not mutually exclusive (Retaux and Casane 2013). In addition, the term “wreck” does not acknowledge that cave animals in general, and

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.02.446701; this version posted June 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

81 cavefishes in particular, may evolve constructive traits involved in their adaptation to cave
82 ecosystems (Hinaux, et al. 2016; Soares and Niemiller 2020).
83
84 Previous studies have used the level of eye regression to infer the relative age of cavefishes.
85 However, these studies were biased and unreliable because they assumed that eye
86 degeneration is a slow progressive process, lacked an explicit phylogenetic framework, or
87 used phylogenies based on troglomorphic characters (Eigenmann 1909; Poulson and White
88 1969; Poulson 2001). Our current knowledge of eye development suggests that the idea of a
89 slow, steady and progressive process is misleading because the impact of mutations on eye
90 development can vary enormously, from a minor effect on eye structure and function to the
91 most extreme case in which a single mutation can produce an eyeless phenotype (Loosli, et al.
92 2003). Today, phylogenetic inference and dating based on gene sequences have superseded
93 those relying on morphological characters. The main advantage of a gene-based phylogenetic
94 approach is that relationship and dating inferences are independent of the analysis of the
95 evolution of morphological traits which are mapped on *a priori* molecular phylogenies.
96 Assuming that genes used for phylogenetic analyses and troglomorphic traits evolved
97 independently, molecular phylogenies allow us to identify convergent evolution. Moreover,
98 the analysis of genes considered to be under relaxed purifying selection after cave settlement
99 allows us to date the environmental shift from surface to cave and to test hypotheses on the
100 mode and tempo of eye degeneration.
101
102 A recent study of the genomes of several cavefish species available at the time allowed us to
103 examine the decay of many eye-specific genes and non-visual opsins, and estimate the age of
104 the shift from surface to cave environment (Policarpo, et al. 2021). This genome-wide
105 analysis demonstrated a low level of gene decay suggesting that these cavefishes originated
106 recently, that is during the Pleistocene (from 2,580,000 to 11,700 years ago). Moreover, the
107 level of eye regression was not correlated with the proportion of eye-specific pseudogenes.
108 However, previous studies based on analyses of one or two genes suggest that other
109 cavefishes could be much more ancient, in particular some among the amblyopsid cavefishes
110 (Dillman, et al. 2011; Niemiller, et al. 2013). Here we sought to further investigate the time of
111 origin of a blind amblyopsid cavefish, *Typhlichthys subterraneus*, using our genome-wide
112 approach. We also examined the same set of eye-specific genes and non-visual opsins of a
113 blind cichlid, *Lamprologus lethops*, which inhabits the deep-water of the Congo river. The
114 level of gene decay in *Typhlichthys subterraneus* supports that blindness may be a few million

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.02.446701; this version posted June 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

115 years old in this lineage, indicating that this species may had settled in caves during the
116 Pliocene (from 5.333 million to 2.58 million years ago). In contrast, the very low level of
117 gene decay in *Lamprologus lethops* indicates that eye regression is very recent, that is it
118 evolved less than 100,000 years ago. These analyses strengthen the hypothesis that once
119 fishes have settled into a dark environment, eyes can degenerate rapidly, reducing in size,
120 sinking into their orbit and becoming covered by skin. In parallel, genes with eye-restricted
121 expression accumulate LoF mutations, but this process is comparatively much slower.
122

123 ## Results

124

125 ### Identification of LoF mutations and pseudogenes in vision, circadian clock and
126 pigmentation genes

127

128 In a previous study (Policarpo, et al. 2021), we defined three gene sets based on the zebrafish
129 genome: 95 vision genes, 42 circadian clock genes and 257 pigmentation genes (**fig. 1**). In the
130 present analysis, we looked for likely LoF mutations (i.e. premature STOP codons, losses of
131 START and STOP codons, losses of intron splice sites and small indels leading to
132 frameshifts) in homologs from the genomes of a blind cavefish from North America, the
133 Southern Cavefish *Typhlichthys subterraneus* (Amblyopsidae) and a blind fish living in deep
134 water of the Congo River in Central Africa, *Lamprologus lethops* (Cichlidae). For each
135 species, *T. subterraneus* and *L. lethops*, we also examined the genes in the closest related
136 surface fish for which a genome was available, *Percopsis transmontana* and *Lamprologus*
137 *tigripictilis*, respectively, and a more distantly related surface species, *Gadus morhua* and
138 *Neolamprologus pulcher*, respectively (**fig. 2)**.
139

140 A total of 55 LoF mutations were found in *T. subterraneus* and *L. lethops*. Most LoF
141 mutations (47) were found in vision genes, while only 2 LoF were found in circadian clock
142 genes and 6 in pigmentation genes. Most LoF mutations in vision genes (43/47) were found in
143 *T. subterraneus* (**supplementary fig. S1 and fig. S2**, **Supplementary Material** online).
144 In order to test if these LoF mutations were distributed randomly within the genes, that is,
145 they were not clustered at the 3' end of the genes where their deleterious effect would be
146 smaller, we computed the effective number of gene segments generated by LoF mutations and
147 compared this value with simulations of random distributions of mutations within genes

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.02.446701; this version posted June 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(Policarpo, et al. 2021). We found that premature STOP codons and frameshifts are distributed randomly throughout the coding sequences (**supplementary fig. S3**, **Supplementary Material** online).

We counted the number of pseudogenes in all six species, that is, the number of genes carrying at least one LoF mutation (**fig. 1** and **fig. 2**). Among the vision genes, we identified 27 pseudogenes in *T. subterraneus*, but only 3 pseudogenes in *L. lethops*. The LoF mutations found in *L. lethops* (**fig. 1**, **supplementary fig. S2**, **Supplementary Material** online) have been reported in a previous study (Aardema, et al. 2020). A gene, *opn3*, was also reported as pseudogenized in this publication, but it was considered a functional gene in our study. Indeed, the frameshift reported in *opn3* is probably an artefact as this gene is not where the deletion, at the origin of the frameshift, was found. We did not find any LoF mutations in this gene (for more details, see **supplementary fig. S4**, **Supplementary Material** online). The percentage of vision pseudogenes in *T. subterraneus* (44%) is the highest ever reported in cavefishes. In contrast, the percentage of vision pseudogenes is very low in *L. lethops* (4%). No circadian clock pseudogene was found in *L. lethops* and only one pseudogene (*per2*) in *T. subterraneus*. The loss of *per2* is recurrent in cavefishes, as pseudogenes of *per2* have been identified in *S. rhinocerous* and *P. andruzzii* (Yang, et al. 2016; Ceinos, et al. 2018; Policarpo, et al. 2021). A circadian clock pseudogene (*cry-dash*) was found in the surface fish *L. tigripictilis*, however, the specimen was heterozygous for this gene. Two pigmentation pseudogenes (*trpm1b* and *mreg*) were identified in *T. subterraneus* and one pseudogene (*oca2*) in *L. lethops*. This mutation has already been reported in a previous study (Aardema, et al. 2020). In the surface fish *L. tigripictilis*, a pseudogene (*pts*) was found, however, the specimen was also heterozygous for the gene (**fig. 2**, **Table S1**, **Supplementary Material** online).

Together, findings from the present study and a previous study (Policarpo, et al. 2021) has allowed us to estimate that among the set of 95 vision genes, 46 have been found pseudogenized in at least one cave species, that is, 48% of the gene set. In sharp contrast, only 4 pseudogenes were identified among 42 circadian clock genes, and 18 pseudogenes among the 257 pigmentation genes, that is, 10% and 7% respectively.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.02.446701; this version posted June 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Estimation of the number of neutral vision genes based on the distribution of LoF mutations per pseudogene

In *T. subterraneus*, among 62 vision genes, 43 LoF mutations were found in 27 pseudogenes. More precisely, there were 17 pseudogenes with 1 mutation, 7 with 2 mutations, 1 with 3 mutations, 1 with 4 mutations and 1 with 5 mutations. This distribution was compared with expected distributions obtained for different numbers of neutral genes ranging from 27 (44%) to 62 (100%). A better fit between the observed and expected distribution was found when at least 80% of genes were assumed to be neutral sequences (**fig. 3**). This result is congruent with the estimation based on the distribution of LoF mutations per pseudogene in *Lucifuga dentata* (Policarpo, et al. 2021) (**fig. 3**). Moreover, the higher percentage of pseudogenes with more than one LoF mutation in *T subterraneus* than in *L. dentata* is congruent with the higher percentage of pseudogenes (44% *vs* 25%) indicating that *T. subterraneus* is the oldest cavefish species.

Another approach to tackle this issue involved the analysis of the number of vision genes pseudogenized in both species. There are 27 pseudogenes among 62 vision genes in *T. subterraneus* (present study) and 19 pseudogenes among 76 vision genes in *L. dentata* (Policarpo, et al. 2021). Moreover, among 58 vision genes identified in both species, there are 10 genes - *opn7b*, *parapinopsin-1*, *parietopsin*, *cryaa*, *crybb1*, *crybgx*, *crygn2*, *grk1b*, *grk7a* and *gc2* - pseudogenized in both species. As expected for two distantly related blind species, no shared LoF mutation was found in orthologous pseudogenes (**supplementary Table S2**, **Supplementary Material** online). Assuming that all vision genes are dispensable and have the same LoF mutation rate, we can compute the expected number of genes pseudogenized in parallel in two species among a set of vision genes studied in both species (probability of pseudogenization in one species x probability of pseudogenization in the other species x number of genes), that is 27/62 x 19/76 x 58 = 6.3. A script was written to compute the distribution of the probabilities of the number of shared pseudogenes, in a range between 0 and 14, taking into account that 14 pseudogenes were found in *L. dentata* and 25 in *T. subterraneus* among the 58 vision genes found in both species (**supplementary fig. S5**, **Supplementary Material** online). The expected value (6.3) and the low probability (0.01) of finding 10 shared pseudogenes suggest a significant excess of shared pseudogenes that could be evidence that some vision genes cannot carry LoF mutations. However, with a more

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.02.446701; this version posted June 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

realistic model taking into account the length and the number of introns of each gene to estimate the relative LoF mutation rate of each gene, a probability higher than 5% was found for a number of shared pseudogenes in a range between 5 and 10 (**supplementary fig. S5**, **Supplementary Material** online). This result is another evidence that most, if not all, vision genes are dispensable in these blind species.

Dating the relaxation of purifying selection on vision genes in four blind fishes

To date the relaxation of purifying selection on vision genes in *T. subterranues*, *L. lethops* and *A. mexicanus,* for each species, we used the number of pseudogenes among vision genes and the LoF mutation rate per generation estimated as described in a previous publication (Policarpo, et al. 2021). The highest probability of finding 27 pseudogenes among 62 genes in *T. subterraneus* was obtained for relaxed selection starting 569,570 generations ago (probability > 5% between 447,390 and 712,830 generations). For *L. lethops* (3 pseudogenes among 79 genes), the time since relaxed selection was estimated to 28,970 generations ago (probability > 5% between 8,670 and 68,620 generations). For *A. mexicanus* (1 pseudogene among 85 genes), the time since relaxed selection was estimated to 8,560 generations ago (probability > 5% between 450 and 38,580 generations). For *L. dentata* (19 pseudogenes among 76 genes), the estimate was obtained in a previous study (Policarpo, et al. 2021), 367,780 generations ago (probability > 5% in a range between 273,990 and 480,980 generations)(**fig. 4**).

For each species, we also computed this probability using only a subset of vision genes found pseudogenized in at least one among the other cavefishes, increasing the likelihood that the genes examined are dispensable in caves. Excluding *T. subterraneus*, 33 genes were found with at least one LoF mutation in the other cave species. Twenty-one of these genes were identified in *T. subterraneus* of which 13 were pseudogenes. The highest probability of finding 13 pseudogenes among 21 genes was obtained for relaxed selection starting 910,940 generations ago (probability > 5% between 557,730 and 1,401,750 generations). Excluding *L. lethops*, 46 genes with a LoF were found in the other cave species. Forty-one of these were identified in *L. lethops*, of which 2 were pseudogenes. The time since relaxed selection was estimated to be 35,200 generations ago (probability > 5% between 6,640 and 103,250 generations). Excluding *A. mexicanus*, 47 genes with a LoF were found in the other cave species. Forty-four of these were identified in *A. mexicanus*, of which one was a pseudogene. The time since relaxed selection was estimated to be 15,010 generations ago (probability >

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.02.446701; this version posted June 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

5% between 790 and 67,740 generations). Excluding *L. dentata*, 41 genes with a LoF were found in the other cave species. Thirty-five of these were identified in *L. dentata*, of which 13 were pseudogenes. The time since relaxed selection was estimated to be 559,690 generations ago (probability > 5% between 365,220 and 814,710 generations).

Discussion

We discuss the patterns of gene decay among different gene sets in various cavefishes. Then, we show how our analyses of the genomes of a blind cavefish and a deep-water blind fish strengthen our hypotheses on the pace of eye loss and the evolutionary fate of cave-adapted fishes.

Patterns of gene decay

In a previous study, we defined a set of genes we called vision genes that are most likely to be dispensable in a dark environment and can carry LoF mutations without a large impact on overall fitness. They are primarily eye-specific genes involved in visual phototransduction, but we also included a set of non-visual opsin that showed similar gene decay (Policarpo, et al. 2021). In this publication, our analysis of gene decay in *L. dentata* suggested that most, if not all, vision genes are dispensable in blind cave fishes, however only 25% of *L. dentata* vision genes have LoF mutations. In the present study, we found that 44% of the vision genes of *T. subterraneus* have LoF mutations, an observation strongly strengthening our hypothesis. In *T. subterraneus*, the distribution of LoF mutations per genes is similar to the one in *L. dentata* and in accordance with the assumption that most vision genes are under relaxed purifying selection in the darkness of caves.

The analysis of seven cavefishes identified pseudogenes in 48% of the 95 vision genes. We have shown that this set of genes is a useful tool for estimating the time of origin of blindness and the extent of common origin and parallel evolution among closely related cavefishes. However, only the analysis of other cavefish genomes may allow us to unambiguously identify the subset of vision genes that are dispensable in the dark.

Decoupling of eye regression and vision gene decay

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.02.446701; this version posted June 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Assuming that eye regression is a progressive process and that cavefishes can be ordered on a ladder of degeneration, eye regression has been used to estimate the age of different cavefish lineages (Wilkens, et al. 1989). However, a well-established and independent phylogenetic framework and independent evidence of the age of eye regression are necessary to understand the dynamic of eye regression and the diversity of its modality in different cavefish lineages. The very low decay in vision genes in *Astyanax mexicanus* cavefish, despite their highly degenerate eyes, indicates that *A. mexicanus* has lost eyesight very recently (Fumey, et al. 2018; Policarpo, et al. 2021). The very similar pattern of gene decay observed in *Lamprologus lethops* suggests that eyesight was also lost very recently in this species, however, while the eyes are very small and sunk under the skin, their general organization is much better preserved. On the other hand, both cavefishes that have accumulated many vision pseudogenes, that is *Lucifuga dentata* and *Typhlichthys subterraneus*, have highly degenerate eyes. There are probably many evolutionary pathways leading to eye regression in cavefishes, some through intermediary stages involving first eye size reduction but normal development, while in other cases through changes in early development leading to highly degenerate eyes. Both chance and environmental constraints may be involved in this process. In particular, it is possible that some blind fishes need to retain the capacity to detect sources of light.

Conclusion

The analysis of two recently published genomes of blind fishes, one living in caves and the other in the deep-water of a river, strengthens our hypothesis that blind cavefishes are not very ancient and most of them lost eye function during the Pliocene and Pleistocene. The recent eye regression in *L. lethops* indicates that eyes can sink rapidly without other important morphological changes, or as in *A. mexicanus*, with extensive eye degeneration. In both cases, the number of eye-specific pseudogenes is very small, two and one respectively. In other words, eyes can sink rapidly but eye-specific genes rust slowly in cavefishes.

Among blind cavefishes for which the genome is available, several lines of evidence suggest that the oldest may be the amblyopsid *T. subterraneus*, in which 44% of vision genes are pseudogenized. Nevertheless, other North American amblyopsids may be older (Niemiller, et al. 2013). When the genomes of these species will be available, the analysis of the vision genes should clarify the age of these cavefishes and also to which extent cave amblyopsids evolved independently.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.02.446701; this version posted June 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The analysis of the genome of *T. subterraneus* allowed us to identify a blind cavefish that is probably older than *L. dentata,* which was previously considered the oldest cavefish among those with sequenced genomes, but this does not refute our current working hypothesis that blind cavefishes cannot thrive more than a few millions years in cave ecosystems.

Materials and Methods

Vision, circadian clock and pigmentation gene sets

The three sets of genes, vision, circadian clock and pigmentation, have been previously described (Policarpo, et al. 2021). Briefly, the set of vision genes was based on the expression patterns of 95 zebrafish genes coding for visual and non-visual opsins, crystallins and proteins involved in the phototransduction cascade. The set of circadian clock genes (42 genes in zebrafish) and the set of pigmentation genes (257 genes in zebrafish) were established in other studies (Li, et al. 2013; Lorin, et al. 2018). The complete list of genes with standard identifiers (https://zfin.org/) is given in **fig. 1**.

Construction of data sets

Danio rerio sequences were used as queries for Blast and tBlastn 2.6.0+ (Altschul, et al. 1990) to the genome of *Percopsis transmontana* (GCA_900302285.1), *Gadus morhua* (gadMor1), *Neolamprologus pulcher* (GCF_000239395.1) and *Typhlichthys subterraneus* (GCA_900302405.1). Regions corresponding to the best matches were extracted with SAMtools v1.9 faidx (Li 2011). Coding sequences and introns were predicted using EXONERATE v2.4.0 (Slater and Birney 2005) (model protein2genome, using *Danio rerio* protein sequences as query).

BAM files of *Lamprologus lethops* and *Lamprologus tigripictilis* reads aligned to the genome of *Neolamprologus pulcher* were downloaded from NCBI (SRR10811633, SRR10811634). SAMtools BCFtools v1.9 (Li 2011) was used to generate a consensus genome (BCFtools consensus) from which genes were predicted following the method described above. Coding sequences were classified as “complete” if the complete CDS was retrieved or “incomplete” otherwise. If less than 10% of a coding sequence was found, the gene was classified as “Not Found” and it was not further analyzed. A CDS was classified as a “pseudogene” if at least

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.02.446701; this version posted June 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

one LoF mutation was found (premature STOP codon, loss of the initiation codon, loss of the STOP codon, indel leading to a frameshift, mutations at intron splice sites).

Phylogenetic analyses

Orthologous and paralogous relationships between genes were inferred through phylogenetic analyses. Coding sequences were trimmed and aligned with MACSE v2.03 program trimNonHomologousFragments and alignSequences (Ranwez, et al. 2018). Frameshifts and internal stop codons were then removed with the MACSE v2.03 progam exportAlignment. For each alignment, DNA sequences were translated into protein sequences and a maximum likelihood phylogenetic tree was inferred using IQ-TREE v1.6.12 (Nguyen, et al. 2015) with the optimal model found by ModelFinder (Kalyaanamoorthy, et al. 2017). Robustness of the nodes was evaluated with 1,000 ultrafast bootstraps (Hoang, et al. 2018). Trees were rooted and visualized using iTOL v6 (Letunic and Bork 2007).

Estimation of the proportion of neutral vision genes in cavefishes

The distribution of LoF mutations per gene in *T. subterraneus* was used to estimate the proportion of neutral genes using the method described previously in (Policarpo, et al. 2021). In brief, 10,000 simulations of the distribution of 43 mutations in a random sample of V genes from the 62 *T. subterraneus* vision genes was generated. V ranged from the number of observed pseudogenes (27), which is the minimal number of neutral genes, to the total number of genes (62). The length of the coding sequence and the number of introns of each gene were taken into account to compute the relative mutation rate.

Then, we computed the expected number of vision genes pseudogenized in both *T. subterraneus* and *L. dentata*, assuming that all vision genes are dispensable and have the same LoF mutation rate. The following formula was used: $E = P_{Ts} \times P_{Ld} \times N$, where P_{Ts} is the probability that a vision gene is pseudogenized in *T. subterraneus*, P_{Ld} is the probability that a vision gene is pseudogenized in *L. dentata*, and N is the number of vision genes found in both species.

Given that of the 58 vision genes found in both species, 14 genes were pseudogenized in *L. dentata*, 25 in *T. subterraneus*, and 10 were shared, a custom R script was used to evaluate if this number of shared pseudogenes was likely under the hypothesis that all vision genes are

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.02.446701; this version posted June 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

dispensable. The script generated 10,000 simulations of random pseudogenization of the genes, taking into account the length and the number of introns of each gene to compute its relative mutation rate. For each simulation, the number of shared pseudogenes was recorded, allowing us to estimate the probability of obtaining a given number of shared pseudogenes in the range between 0 and 14 (**supplementary fig. S5**, **Supplementary Material** online).

Dating relaxation of purifying selection on vision genes in *L. lethops* and *T. subterraneus* from pseudogene numbers

We used the method previously described (Policarpo, et al. 2021). Briefly, we computed the probability than D genes among T genes accumulated a LoF mutation taking into account the mean gene length. This method requires an estimate of the LoF mutation rate per generation. Assuming a transition/transversion ratio in vision genes of 1.68 computed with PAML v4.9h (Yang 2007) and taking into account the codon usage in vision genes, we found that the probability that a single mutation led to a premature stop codon was 0.031. In *T. subterraneus*, 12 premature stop codons and 27 frameshifts were found in vision genes. Thus, the probability that a mutation is a frameshift was (27/12) x 0.031 = 0.069.
The probability of splice site mutation was 4 x (total intron number / total cds length) = 4 x (255/48703) = 0.021. The probability of start loss was equal to 3 x (number of genes / total cds length) = 3 x (62/48703) = 0.0038 and the probability of stop loss was equal to 0.852*(Start loss probability) = 0.0033. Thus μ_{LoF} was equal to 0.128μ with μ set to 10^{-8}. As a very low number of pseudogenes were found in *L. lethops*, the LoF mutation rate found in *T. subterraneus* was also used for dating relaxation of selection in *L. lethops and A. mexicanus*. Dating of *L. dentata* was estimated using the LoF mutation rate (0.0717μ) reported in a previous study (Policarpo, et al. 2021).

R analyses

R script files for reproducing analyses and figures is available in supplementary files (see Data Availability) and on GitHub (https://github.com/MaximePolicarpo/Genomic-evidence-that-cavefishes-are-not-wrecks-of-ancient-life). A list of R packages used is given below, with a reference to a publication if available or an URL otherwise: plotrix v3.7-7 (Lemon 2006), data.table v1.12.8 (https://cran.r-project.org/web/packages/data.table/index.html), dplyr v0.8.5 (https://cran.r-project.org/web/packages/dplyr/index.html), tidyverse v1.3.0 (Wickham, et al.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.02.446701; this version posted June 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2019), purr v0.3.3 https://cran.r-project.org/web/packages/purrr/index.html, ggplot2 v3.3.0 (Wickham 2016), ggpubr v0.2.5 (https://cran.r-project.org/web/packages/ggpubr/index.html), gtools v3.8.2 (https://cran.r-project.org/web/packages/gtools/index.html), forcats v0.5.0 (https://cran.r-project.org/web/packages/forcats/index.html), FSA v0.8.30 (Ogle, et al. 2020), patchwork v1.0.0 (https://cran.r-project.org/web/packages/patchwork/index.html), quantable 0.3.6 (https://cran.r-project.org/web/packages/quantable/index.html), reticulate v1.15 (https://cran.r-project.org/web/packages/reticulate/index.html), sjPlot v2.8.3 (Lüdecke 2020), gridExtra v2.3 (https://cran.r-project.org/web/packages/gridExtra/index.html), ape v5.3 (Paradis and Schliep 2019) and phytools v0.7-47 (Revell 2012).

Data Availability

All data and results of the analyses performed in this study are available for download at figshare (https://figshare.com/articles/dataset/Supplementary_files/14237279): three tables with all gene sequences and their genomic positions (one table per gene data set); three tables with a description of LoF mutations (one table per gene data set); Exonerate raw results; Gene alignments, IQ-TREE results and phylogenetic trees in PDF; a R script to reproduce the analyses.

Supplementary Material

Supplementary data are available at Genome Biology and Evolution.

Acknowledgments

This work was supported by a collaborative grant from Agence Nationale de la Recherche (BLINDTEST to S.R. and D.C.) and from Institut Diversité Ecologie et Evolution du Vivant (to S.R. and D.C).

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.02.446701; this version posted June 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure Legends

Fig 1: Gene sets and pseudogenes. (A) Vision genes. (B) Circadian clock genes. (C) Pigmentation genes. Pseudogenes are coloured according the species in which they were found. In candidate gene studies, only few genes were examined, whereas most genes were inspected in our genome-wide analyses. Gene copy number variations: there are four copies of *Rh2* in zebrafish whereas there is only one *Rh2* gene in *T. subterraneus*; there is one *opn6a* gene in zebrafish whereas there are two *opn6a* genes in *T. subterraneus*. Data for previously studied cavefishes (*A. mexicanus*, *Lucifuga* spp. and *Sinocyclocheilus* spp.) have been redrawn from Policarpo et al. (2021), for comparison.

Fig 2: Phylogeny and pseudogene mapping. For each gene set, the number of pseudogenes found and the number of gene examined in a species are given to the right of the species name. het: heterozygous. Data for previously studied cavefishes (*A. mexicanus*, *Lucifuga* spp. and *Sinocyclocheilus* spp.) have been redrawn from Policarpo et al. (2021), for comparison.

Fig 3: Estimation of the number of neutral genes in *T. subterraneus*. Red lines: observed distribution. Expected distributions were obtained for different numbers of neutral genes among 62 genes: between 27 and 62 (each expected distribution was obtained with 10,000 simulations). The distributions for *L. dentata* have been redrawn from Policarpo et al. (2021), for comparison.

Fig 4: Dating of the relaxation of purifying selection on vision genes. Solid lines: probabilities of finding 27 pseudogenes among 62 vision genes in *T. subterraneus,* 19 pseudogenes among 76 genes in *L. dentata,* three pseudogenes among 79 vision genes in *L. lethops* and one pseudogene among 85 vision genes in *A. mexicanus,* according to the time in caves (generations) and assuming relaxed selection on the whole vision gene set in this environment. Dashed lines: probabilities of finding pseudogenes using only genes found pseudogenized in other cave species, increasing the likelihood that the genes examined are dispensable in caves. The number of generations for which the highest probability was found is reported.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.02.446701; this version posted June 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

475 ## References

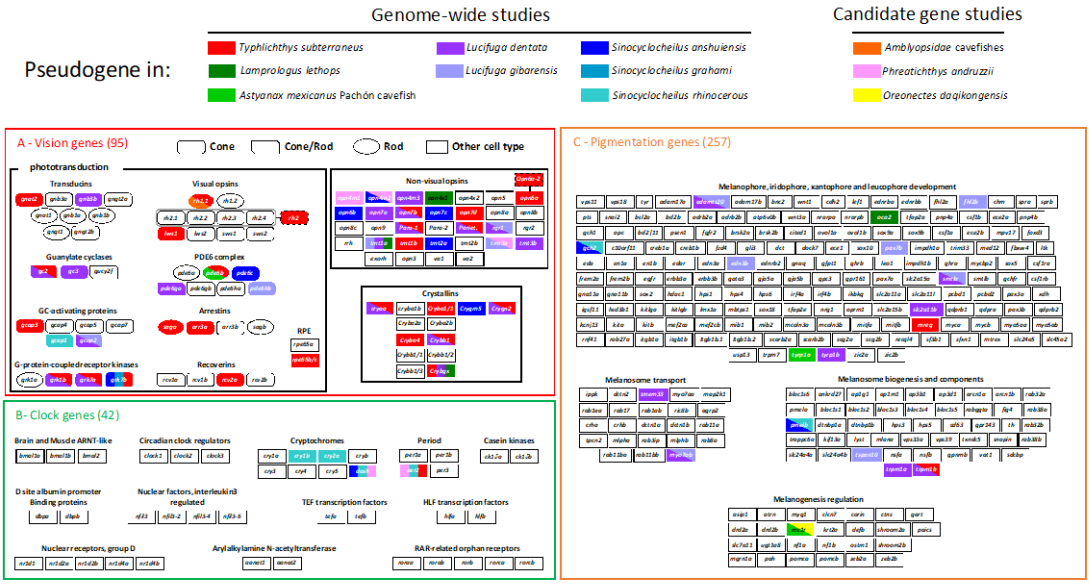
476

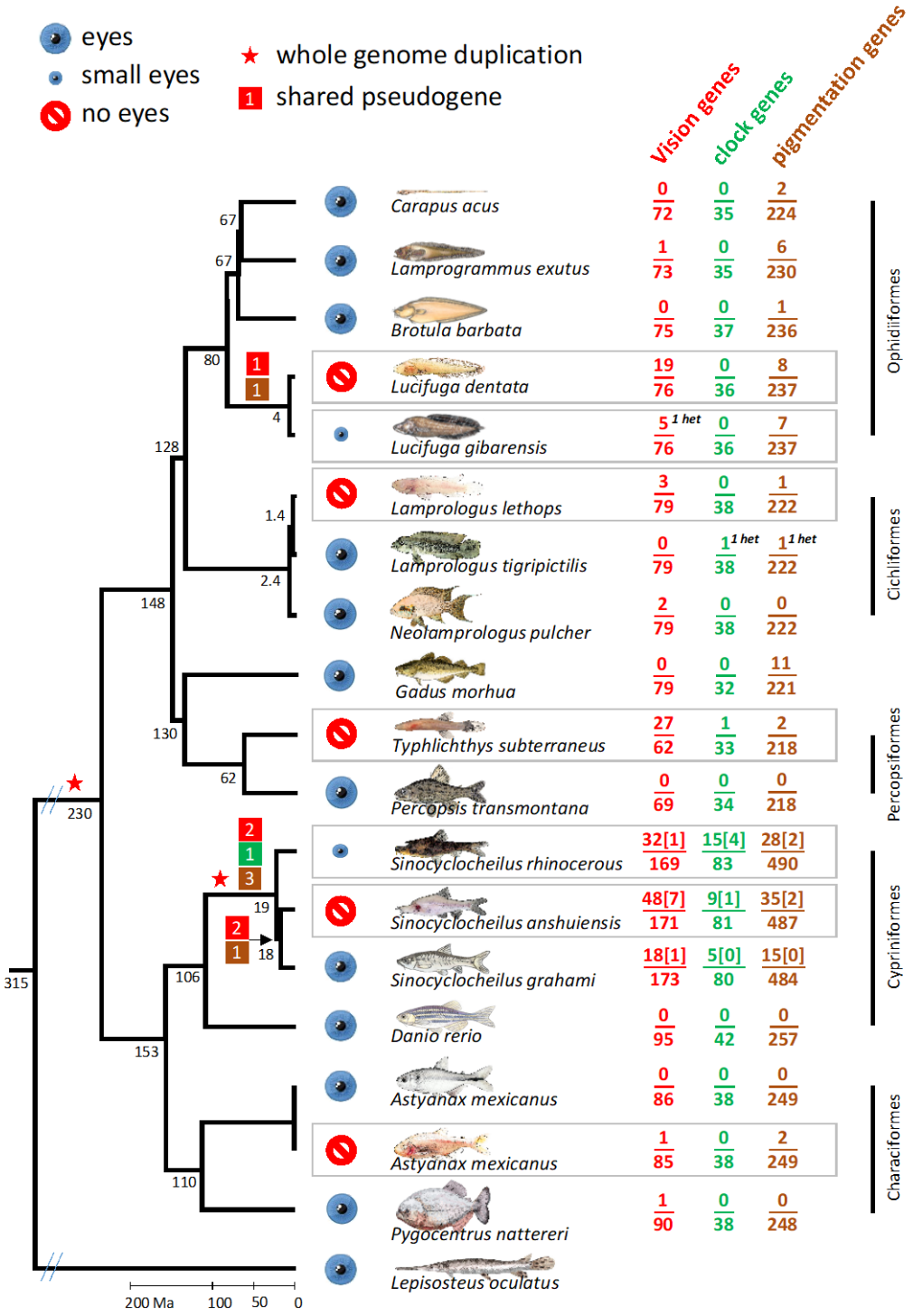
477 Aardema ML, Stiassny MLJ, Alter SE. 2020. Genomic Analysis of the Only Blind Cichlid
478 Reveals Extensive Inactivation in Eye and Pigment Formation Genes. Genome Biol Evol
479 12:1392-1406.
480 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search
481 tool. J Mol Biol 215:403-410.
482 Ceinos RM, Frigato E, Pagano C, Fröhlich N, Negrini P, Cavallari N, Vallone D, Fuselli S,
483 Bertolucci C, Foulkes NS. 2018. Mutations in blind cavefish target the light-regulated
484 circadian clock gene, *period 2*. Scientific Reports 8:8754.
485 Culver DC, Pipan T. 2009. The Biology of Caves and Other Subterranean Habitats. Oxford:
486 Oxford University Press.
487 Darwin CR. 1859. On the origin of species by means of natural selection, or the preservation
488 of favoured races in the struggle for life. London: John Murray.
489 Dillman CB, Bergstrom DE, Noltie DB, Holtsford TP, Mayden RL. 2011. Regressive
490 progression, progressive regression or neither? Phylogeny and evolution of the
491 Percopsiformes (Teleostei, Paracanthopterygii). Zoologica Scripta 40:45-60.
492 Eigenmann CH. 1909. Cave vertebrates of America : a study in degenerative evolution.
493 Washington, D.C.: Carnegie Institution of Washington.
494 Fumey J, Hinaux H, Noirot C, Thermes C, Rétaux S, Casane D. 2018. Evidence for late
495 Pleistocene origin of *Astyanax mexicanus* cavefish. Bmc Evolutionary Biology 18:43.
496 Hinaux H, Devos L, Blin M, Elipot Y, Bibliowicz J, Alié A, Rétaux S. 2016. Sensory
497 evolution in blind cavefish is driven by early embryonic events during gastrulation and
498 neurulation. Development 143:4521.
499 Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: Improving
500 the Ultrafast Bootstrap Approximation. Molecular Biology and Evolution 35:518-522.
501 Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder:
502 fast model selection for accurate phylogenetic estimates. Nature Methods 14:587-589.
503 Lemon J. 2006. Plotrix: A Package in the Red Light District of R. R-News 6.
504 Letunic I, Bork P. 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree
505 display and annotation. Bioinformatics 23:127-128.
506 Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping
507 and population genetical parameter estimation from sequencing data. Bioinformatics 27:2987-
508 2993.
509 Li Y, Li G, Wang H, Du J, Yan J. 2013. Analysis of a Gene Regulatory Cascade Mediating
510 Circadian Rhythm in Zebrafish. Plos Computational Biology 9:e1002940.
511 Loosli F, Staub W, Finger-Baier KC, Ober EA, Verkade H, Wittbrodt J, Baier H. 2003. Loss
512 of eyes in zebrafish caused by mutation of chokh/rx3. EMBO reports 4:894-899.
513 Lorin T, Brunet FG, Laudet V, Volff J-N. 2018. Teleost Fish-Specific Preferential Retention
514 of Pigmentation Gene-Containing Families After Whole Genome Duplications in Vertebrates.
515 G3: Genes|Genomes|Genetics 8:1795-1806.
516 Lüdecke D. 2020. sjPlot: Data Visualization for Statistics in Social Science. R package
517 version 2.8.5.
518 Moran D, Softley R, Warrant EJ. 2015. The energetic cost of vision and the evolution of
519 eyeless Mexican cavefish. Science Advances 1:e1500363.
520 Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: A fast and effective
521 stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular Biology and
522 Evolution 32:268-274.

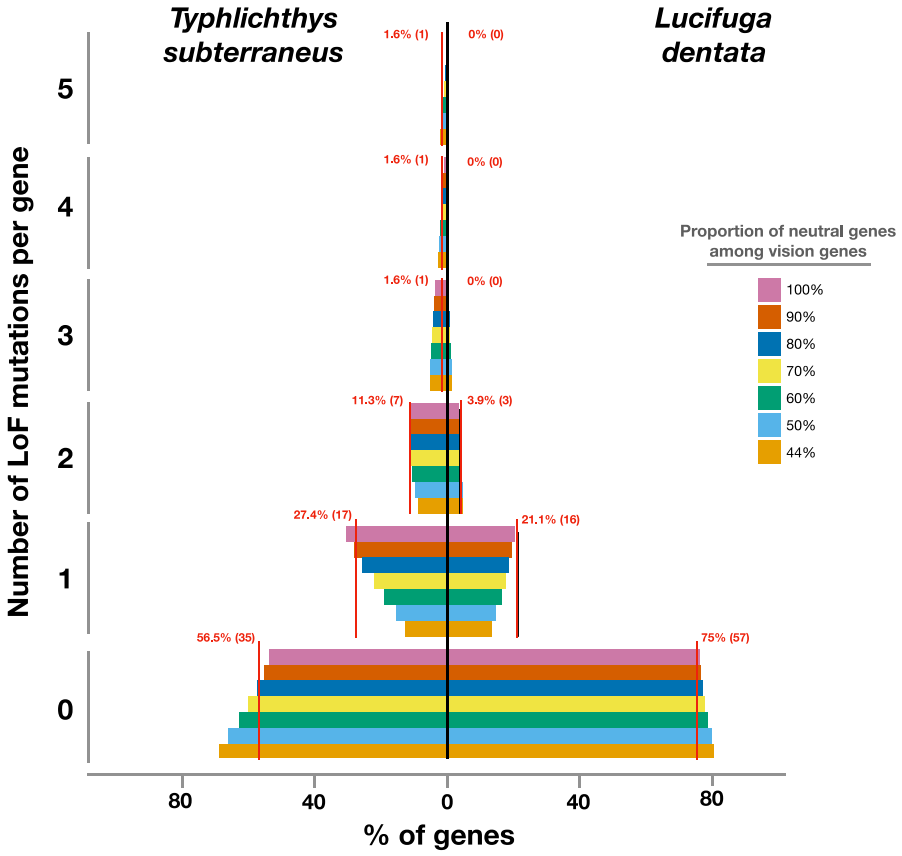
bioRxiv preprint doi: https://doi.org/10.1101/2021.06.02.446701; this version posted June 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

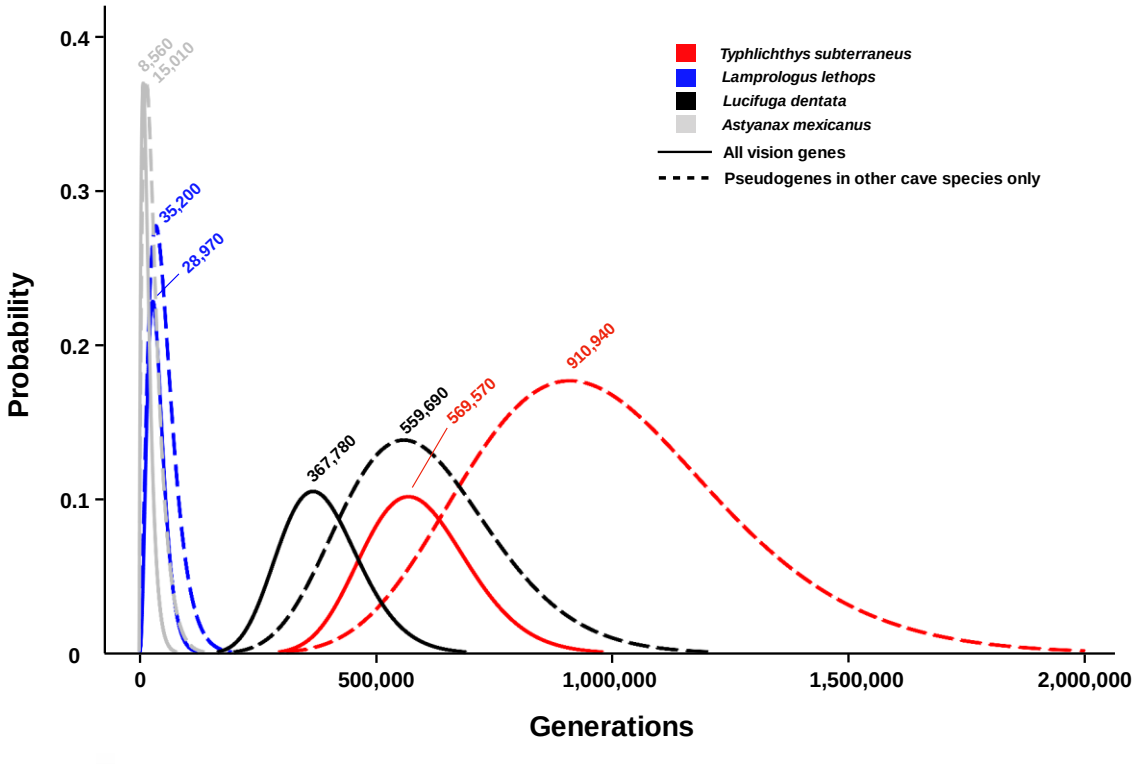
523 Niemiller ML, Fitzpatrick BM, Shah P, Schmitz L, Near TJ. 2013. Evidence for repeated loss
524 of selective constraint in rhodopsin of amblyopsid cavefishes (teleostei: amblyopsidae).
525 Evolution 67:732-748.
526 Ogle DH, Wheeler P, Dinno A. 2020. FSA: Fisheries Stock Analysis. R package version
527 0.8.30.
528 Paradis E, Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and
529 evolutionary analyses in R. Bioinformatics 35:526-528.
530 Policarpo M, Fumey J, Lafargeas P, Naquin D, Thermes C, Naville M, Dechaud C, Volff J-N,
531 Cabau C, Klopp C, et al. 2021. Contrasting gene decay in subterranean vertebrates: insights
532 from cavefishes and fossorial mammals. Molecular Biology and Evolution 38:589-605.
533 Poulson TL. 2001. Morphological and physiological correlates of evolutionary reduction of
534 metabolic rate among amblyopsid cave fishes. In: Romero A, editor. The biology of
535 hypogean fishes. Dordrecht: Springer Netherlands. p. 239-249.
536 Poulson TL, White WB. 1969. The Cave Environment. Science 165:971-981.
537 Protas M, Conrad M, Gross JB, Tabin C, Borowsky R. 2007. Regressive evolution in the
538 Mexican cave tetra, *Astyanax mexicanus*. Current Biology 17:452-454.
539 Ranwez V, Douzery EJP, Cambon C, Chantret N, Delsuc F. 2018. MACSE v2: Toolkit for
540 the Alignment of Coding Sequences Accounting for Frameshifts and Stop Codons. Molecular
541 Biology and Evolution 35:2582-2584.
542 Retaux S, Casane D. 2013. Evolution of eye development in the darkness of caves:
543 adaptation, drift, or both? Evodevo 4:26.
544 Revell LJ. 2012. phytools: an R package for phylogenetic comparative biology (and other
545 things). Methods in Ecology and Evolution 3:217-223.
546 Romero A. 2009. Cave Biology: Life in Darkness. Cambridge: Cambridge University Press.
547 Slater GSC, Birney E. 2005. Automated generation of heuristics for biological sequence
548 comparison. BMC Bioinformatics 6:31.
549 Soares D, Niemiller ML. 2020. Extreme Adaptation in Caves. The Anatomical Record
550 303:15-23.
551 Wickham H. 2016. ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag
552 Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, Grolemund G, Hayes
553 A, Henry L, Hester J, et al. 2019. Welcome to the Tidyverse. Journal of Open Source
554 Software 4:1686.
555 Wilkens H, Strecker U, Yager J. 1989. Eye reduction and phylogenetic age in ophidiiform
556 cave fish. Zeitschrift Fur Zoologische Systematik Und Evolutionsforschung 27:126-134.
557 Yamamoto Y, Byerly MS, Jackman WR, Jeffery WR. 2009. Pleiotropic functions of
558 embryonic sonic hedgehog expression link jaw and taste bud amplification with eye loss
559 during cavefish evolution. Developmental Biology 330:200-211.
560 Yang J, Chen X, Bai J, Fang D, Qiu Y, Jiang W, Yuan H, Bian C, Lu J, He S, et al. 2016. The
561 *Sinocyclocheilus* cavefish genome provides insights into cave adaptation. BMC Biol 14:1-13.
562 Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Molecular Biology
563 and Evolution 24:1586-1591.
564 Yoshizawa M, Yamamoto Y, O'Quin KE, Jeffery WR. 2012. Evolution of an adaptive
565 behavior and its sensory receptors promotes eye regression in blind cavefish. BMC Biol
566 10:108.

567


Genome-wide studies
Candidate gene studies
Pseudogene in:
Typhlichthys subterraneus
Lucifuga dentata
Sinocyclocheilus anshuiensis
Amblyopsidae cavefishes
Lamprologus lethops
Lucifuga gibarensis
Sinocyclocheilus grahami
Phreatichthys andruzzii
Astyanax mexicanus Pachón cavefish
Sinocyclocheilus rhinocerous
Oreonectes daqikongensis
A - Vision genes (95)
Cone
Cone/Rod
Rod
Other cell type
phototransduction
Transducins
Visual opsins
Non-visual opsins
Guanylate cyclases
PDE6 complex
Crystallins
GC activating proteins
Arrestins
RPE
G-protein-coupled receptor kinases
Recoverins
B- Clock genes (42)
Brain and Muscle ARNT-like
Circadian clock regulators
Cryptochromes
Period
Casein kinases
D site albumin promoter Binding proteins
Nuclear factors, interleukin3 regulated
TEF transcription factors
HLF transcription factors
Nuclear receptors, group D
Arylalkylamine N-acetyltransferase
RAR-related orphan receptors
C - Pigmentation genes (257)
Melanophore, iridophore, xantophore and leucophore development
Melanosome transport
Melanosome biogenesis and components
Melanogenesis regulation


eyes
small eyes
no eyes
whole genome duplication
shared pseudogene
Vision genes
clock genes
pigmentation genes
Carapus acus
0/72 0/35 2/224
Lamprogrammus exutus
1/73 0/35 6/230
Brotula barbata
0/75 0/37 1/236
Lucifuga dentata
19/76 0/36 8/237
Lucifuga gibarensis
5 1 het/76 0/36 7/237
Lamprologus lethops
3/79 0/38 1/222
Lamprologus tigripictilis
0/79 1 1 het/38 1 1 het/222
Neolamprologus pulcher
2/79 0/38 0/222
Gadus morhua
0/79 0/32 11/221
Typhlichthys subterraneus
27/62 1/33 2/218
Percopsis transmontana
0/69 0/34 0/218
Sinocyclocheilus rhinocerous
32[1]/169 15[4]/83 28[2]/490
Sinocyclocheilus anshuiensis
48[7]/171 9[1]/81 35[2]/487
Sinocyclocheilus grahami
18[1]/173 5[0]/80 15[0]/484
Danio rerio
0/95 0/42 0/257
Astyanax mexicanus
0/86 0/38 0/249
Astyanax mexicanus
1/85 0/38 2/249
Pygocentrus nattereri
1/90 0/38 0/248
Lepisosteus oculatus
Ophidiliformes
Cichliformes
Percopsiformes
Cypriniformes
Characiformes
67
67
80
4
128
1.4
2.4
148
130
62
230
19
18
106
315
153
110
200 Ma
100
50
0


Typhlichthys subterraneus
Lucifuga dentata
Number of LoF mutations per gene
5
4
3
2
1
0
1.6% (1)
0% (0)
1.6% (1)
0% (0)
1.6% (1)
0% (0)
11.3% (7)
3.9% (3)
27.4% (17)
21.1% (16)
56.5% (35)
75% (57)
Proportion of neutral genes among vision genes
100%
90%
80%
70%
60%
50%
44%
80
40
0
40
80
% of genes


Typhlichthys subterraneus
Lamprologus lethops
Lucifuga dentata
Astyanax mexicanus
All vision genes
Pseudogenes in other cave species only
8,560
15,010
35,200
28,970
367,780
559,690
569,570
910,940
Probability
0.4
0.3
0.2
0.1
0
0
500,000
1,000,000
1,500,000
2,000,000
Generations