UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE CIENCIAS BIOLÓGICAS

TESIS DOCTORAL

Molecular evolution and role of aquaporins in the water-to-land transitions of amphibious fishes

Evolución molecular y rol de las acuaporinas en las transiciones agua-tierra de los peces anfibios

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

Héctor Lorente Martínez

DIRECTORES

Ainhoa Agorreta Calvo
Diego San Mauro Martín

UNIVERSIDAD COMPLUTENSE DE MADRID
FACULTAD DE CIENCIAS BIOLÓGICAS

TESIS DOCTORAL

EVOLUCIÓN MOLECULAR Y ROL DE LAS ACUAPORINAS EN LAS TRANSICIONES AGUA-TIERRA DE LOS PECES ANFIBIOS

MOLECULAR EVOLUTION AND ROLE OF AQUAPORINS IN THE WATER-TO-LAND TRANSITIONS OF AMPHIBIOUS FISHES

MEMORIA PARA OPTAR AL GRADO DE DOCTOR EN BIOLOGÍA
PRESENTADA POR

HÉCTOR LORENTE MARTÍNEZ

DIRECTORES

AINHOA AGORRETA CALVO Y DIEGO SAN MAURO MARTÍN

MADRID, 2022

UNIVERSIDAD COMPLUTENSE DE MADRID
FACULTAD DE CIENCIAS BIOLÓGICAS
PROGRAMA EN BIOLOGÍA

TESIS DOCTORAL

EVOLUCIÓN MOLECULAR Y ROL DE LAS ACUAPORINAS EN LAS TRANSICIONES AGUA-TIERRA DE LOS PECES ANFIBIOS

MOLECULAR EVOLUTION AND ROLE OF AQUAPORINS IN THE WATER-TO-LAND TRANSITIONS OF AMPHIBIOUS FISHES

MEMORIA PARA OPTAR AL GRADO DE DOCTOR EN BIOLOGÍA
PRESENTADA POR

HÉCTOR LORENTE MARTÍNEZ

DIRECTORES

AINHOA AGORRETA CALVO Y DIEGO SAN MAURO MARTÍN

MADRID, 2022

La presente Tesis Doctoral ha sido financiada por la Universidad Complutense de Madrid mediante un contrato de personal investigador predoctoral en formación de la convocatoria CT63/19-CT64/19 y la estancia breve de investigación EB25/22; por el Ministerio de Ciencia e Innovación (MICINN) mediante el proyecto PID2020-115481GB-I00; por el Ministerio de Economía y Competitividad (MINECO) mediante los proyectos RYC-2011-09321, CGL2012-40082 y CGL2017-85637-P y por la Comisión Europea mediante una estancia breve Erasmus+ research mobility traineeship.

A mis padres y mi hermano

“When my information changes, I alter my conclusions. What do you do, sir?”

John Maynard Keynes

“In case I don’t see ya, good afternoon, good evening, and goodnight.”

Truman Burbank

Acknowledgements

First and foremost, I am deeply grateful to my supervisors, Ainhoa Agorreta Calvo and Diego San Mauro Martín, for their support and and guidance throughout the development of this thesis. Their tireless efforts and understanding nature have been instrumental in making this project a reality. I consider myself fortunate to have had not just two supervisors but also two friends (four including Ekai and Adei) who have been a constant source of inspiration and motivation

I would also like to express my gratitude to Iker Irisarri for warmly welcoming me to Germany and sharing his knowledge with utmost patience and enthusiasm. I am indebted to Armin and other members of his group for their support during my stay. A special thank you to Professor Scott Edwards and his team at Harvard University for providing me with an incredible opportunity to learn and grow. My heartfelt gratitude to Hevanna and Magdalena for being wonderful office mates and to Abraham and Jesica for making my stay at Harvard University a memorable experience.

To all my colleagues from the Complu who have been an integral part of my journey. To those who welcomed me with open arms, including Javi, the two Danis, Amparo, Elena, the two Álex. With special mention to María, for making me take my first steps in Bioinformatics and to Antón for sharing such good times. To my later colleagues, Lucía, Gabri, David, Pablo, and others, despite the challenges posed by the pandemic, thank you for your friendship and support. I would also like to extend my gratitude to Isaac and Carlos for their unwavering support and friendship over the years.

To Javi, Rafa and Rubén. Also to Irene, Sonia and Noelia. Thank you for the countless memories and for being a part of my life. I hope

Last but not least, I want to express my deepest gratitude to my parents and my brother,. who have always been my reference and supported me throughout my life. Without their support, none of this would have been possible.

Agradecimientos

En primer lugar, doy las gracias a mis directores, Ainhoa Agorreta Calvo y Diego San Mauro Martín, por su apoyo y supervisión, por su enorme esfuerzo y trabajo durante el desarrollo de esta tesis, pero sobre todo por su trato cercano y siempre comprensivo. Puedo decir que más que dos directores, me llevo dos amigos (cuatro contando a Ekai y Adei). Cuando empecé este proyecto no sabíamos dónde podríamos llegar pero gracias a vuestro entusiasmo esta tesis es una realidad.

Quiero hacer mención también a Iker Irisarri, por acogerme de una manera tan hospitalaria en Alemania y por enseñarme con pasión y paciencia. También a toda la gente de su grupo en especial a mi compañero de despacho Armin. También al profesor Scott Edwards y a todo su grupo por abrirme las puertas de la Universidad Harvard, donde nunca imaginé que podría estar. Y sobre todo a Hevanna y Magdalena mis compañeras de despacho, por todos los buenos momentos, aunque pasáramos mucho frío. Y también a Abraham y Jessica por acogerme en Boston como uno más.

A tod@s mis companeñer@s de la complu. A los que ya estaban cuando llegué, y me acogieron con los brazos abiertos, como Javi, los dos Danis, Amparo, Elena, los dos Álex… Con especial mención a María, por hacerme dar mis primeros pasos en la Bioinformática y a Antón por compartir tan buenos ratos. Y como no, a tod@s los que fueron llegando después, Lucía, Gabri, David, Pablo…, aunque la pandemia nos robara muchos ratos. También a Isaac y Carlos, por todos estos años de amistad y por su confianza y apoyo en esta tesis

A Javi, Rafa y Rubén, mis amigos de toda la vida, por todos los momentos vividos y en especial estos últimos años compartiendo y aprendiendo en el día a día. También a Irene, Sonia y Noelia. En definitiva, amistades que rozan lo familiar y personas de las que he aprendido y con las que he disfrutado los mejores años de vida ¡Gracias!

Por último, pero no por ello menos importante, quiero agradecer a mis padres y a mi hermano. Por ser mi referencia y aconsejarme en los momentos de duda, y por haber

estado siempre desde el primer momento apoyándome y confiando en mí durante toda esta tesis y durante toda mi vida !Sin vuestro apoyo esto no habría sido posible!

Index

List of abbreviations

This is a list of the abbreviations most frequently used throughout the thesis.

AQP: aquaporin
ar/R: aromatic/arginine
AVT: arginine vasotocin
BEB: bayes empirical bayes
BEGFE: bayesian estimation for gene family evolution
CDS: coding DNA sequences
EGT: endosymbiotic gene transfer
ESS: estimated sample size
FDR: false discovery rate
FECA: first eukaryotic common ancestor
GIP: glpf-like intrinsic protein
GLP: glyceroporins
HGT: horizontal gene transfer
HIP: hybrid intrinsic protein
LECA: last eukaryotic common ancestor
LIP: large intrinsic protein
LRT: likelihood ratio test
MDC: major intrinsic protein deep clade
MIP: major intrinsic protein
NIP: nodulin 26-like intrinsic protein
NPA: asparagine-proline-alanine
PIP: plasma membrane intrinsic protein
PSRF: potential scale reduction factor
ROS: reactive oxygen species
SAR: Stramenopiles and Alveolata
SFI: serine-phenylalanine-isoleucine
SIP: small basic intrinsic protein
TIP: tonoplast intrinsic protein
TSAR: Telonemia, Stramenopila, Alveolata, and Rhizaria

WGD: whole genome duplication

XIP: X or uncharacterised intrinsic protein

Abstract

Aquaporins (AQPs) or major intrinsic proteins (MIPs) form an ancient family of transporters for water and small solute across biological membranes. They constitute a highly diverse protein superfamily, and their evolutionary history and functions have been relatively well studied in vertebrates and plants. For instance, in vertebrates four well defined clusters are described: AQP1-like (water-selective classical aquaporins), AQP8-like (ammonia channels), AQP3-like (aquaglyceroporins), and AQP11-like (unorthodox or super aquaporins). In land vertebrates (Tetrapoda), an exclusive clade of MIPs/AQPs—which is clustered within the classical aquaporins and includes three different orthologues—appears to have been important for their process of colonisation of terrestrial environments. MIPs are broadly present across the eukaryotic tree of life suggesting both a more complex evolutionary history and a larger set of functions than previously thought. Here, we studied the diversity of MIP proteins in the entire eukaryotic lineage by setting a general phylogenetic context for understanding their evolution. Besides, considering the importance of aquaporins in the water-to-land transition of sarcopterygian vertebrates (emergence of tetrapods), we studied the molecular evolution of these proteins in several amphibious fishes of the actinopterygian branch in order to investigate possible new duplication events or adaptive modifications at the sequence level that could be related with their acquisition of an amphibious lifestyle. Finally, we provide a robust bioinformatic workflow and pipeline for gene isolation from large-scale genomic data and phylogenetic analyses that enabled the achievement of the molecular evolution goals.

Taking advantage of genomic and transcriptomic data from publicly available eukaryotic databases plus datasets compiled from previous studies, we managed to ascertain a comprehensive catalogue of MIPs within Eukaryotes. In agreement with previous studies, we reported a highly diverse repertoire of MIPs in unicellular eukaryotes suggesting a complex catalogue in the last eukaryotic common ancestor (LECA). In this sense, we reported three MIP clades that likely have deep evolutionary origins. However, whether the origin of these clades occurred during the transition from the first eukaryotic common ancestor (FECA) to LECA, or whether they were already present in FECA is still under investigation. Regarding AQPs/MIPs evolution in amphibious fishes, we obtained a final catalogue of 356 aquaporin sequences from 22 amphibious fish genomes

(plus four lungfish sequences). Unlike sarcopterygians, we found no evidence of the emergence of new AQP classes that can be related to the water-to-land transition in the studied species of actinopterygian amphibious fishes. Instead, we detected signatures of adaptive selection in 19 AQP branches (including AQP1, AQP3, AQP8, AQP10, AQP11, and AQP12 classes) in 13 different amphibious fish lineages, and spotted specific sequence changes in 13 of such branches. Some of these changes are located in, or close to, important motifs of the AQP sequence involved in pore formation or substrate selectivity (such as the NPA motif or the ar/R selectivity filter) suggesting a change in protein structure, function or regulation. Of all our results, those of AQP11 orthologs suggest these could constitute the most promising candidates for further research. Up to 15 positively-selected sequence positions (one shared among three lineages) correspond to six different AQP11 branches. Furthermore, it is remarkable the case of the adaptive selection detected in the AQP11b stem branch of the Gobiidae clade, which, in our dataset, is represented by the amphibious mudskippers and two fully-aquatic relatives. In these, we identified a change at the NPA motif where the canonical N (asparagine) is substituted by an S (serine). The sequence modification of this AQP11b occurred before the evolution of the mudskippers lineage and it could represent a possible case of exaptation in Gobiidae. We conclude that, due to the importance of aquaporins in osmoregulation in fishes, the sequence positions found under adaptive selection may have led to modifications in the structure or function of these proteins that could have played a role in the water-to-land transitions of the studied amphibious fish.

Resumen

Las acuaporinas (AQPs) o proteínas intrínsecas principales (MIPs) forman una antigua familia de transportadores de agua y pequeños solutos a través de las membranas biológicas. Constituyen una superfamilia de proteínas muy diversa, y su historia evolutiva y funciones han sido relativamente bien estudiadas en vertebrados y plantas. Por ejemplo, en los vertebrados se han descrito cuatro grupos bien definidos: tipo-AQP1 (acuaporinas clásicas o canales de agua), tipo-AQP8 (canales de amoníaco), tipo-AQP3 (acuagliceroporinas) y tipo-AQP11 (no ortodoxas o superacuaporinas). En los vertebrados terrestres (Tetrapoda), un clado exclusivo de MIPs/AQPs, que se agrupa dentro de las acuaporinas clásicas e incluye tres ortólogos diferentes, parece haber sido importante para su proceso de colonización de ambientes terrestres. Las MIPs están ampliamente presentes en todo el árbol de la vida eucariótico, lo que sugiere una historia evolutiva más compleja y un conjunto de funciones más amplio de lo que se pensaba inicialmente. En esta tesis, estudiamos la diversidad de proteínas MIP en todo el linaje eucariota estableciendo un contexto filogenético general para comprender su evolución. Además, considerando la importancia de las acuaporinas en la transición agua-tierra de los vertebrados sarcopterigios (surgimiento de los tetrápodos), estudiamos la evolución molecular de estas proteínas en varios peces anfibios de la rama de los actinopterigios con el fin de investigar posibles nuevos eventos de duplicación o modificaciones adaptativas a nivel de secuencia que podrían estar relacionadas con su adquisición de un estilo de vida anfibio. Finalmente, proporcionamos un protocolo bioinformático robusto para el aislamiento de genes a partir de datos genómicos a gran escala y análisis filogenético que ha permitido la consecución de los objetivos de evolución molecular.

Aprovechando los datos genómicos y transcriptómicos de las bases de datos eucariotas disponibles públicamente, así como datos recopilados de estudios previos, logramos determinar un catálogo completo de MIPs de los eucariotas. En corcondancia con estudios previos, proporcionamos un repertorio muy diverso de MIPs en eucariotas unicelulares que sugiere un catálogo complejo en el último ancestro común eucariota (LECA). En este sentido, encontramos tres clados de MIPs que probablemente tengan orígenes evolutivos muy antiguos. Sin embargo, aún seguimos investigando si el origen de estos clados ocurrió durante la transición del primer ancestro común eucariota (FECA) al LECA, o si ya estaban presentes en FECA. Con respecto a la evolución de las AQPs/MIPs en peces

anfibios, obtuvimos un catálogo final de 356 secuencias de acuaporinas de 22 genomas de peces anfibios (además de cuatro secuencias de peces pulmonados). A diferencia de los sarcopterigios, no encontramos evidencia de la aparición de nuevas clases de AQPs que puedan estar relacionadas con la transición de agua a tierra en las especies estudiadas de peces anfibios actinopterigios. En cambio, sí detectamos señales de selección adaptativa en 19 ramas de AQPs (incluidas las clases AQP1, AQP3, AQP8, AQP10, AQP11 y AQP12) en 13 linajes de peces anfibios diferentes, y detectamos cambios de secuencia específicos en 12 de dichas ramas. Algunos de estos cambios están ubicados en, o cerca de, motivos importantes de la secuencia de AQP involucrados en la formación del poro o en la selectividad del sustrato (tal como el motivo NPA o el filtro de selectividad ar/R), lo que sugiere un cambio en la estructura, función o regulación de la proteína. De todos nuestros resultados, los relacionados con los ortólogos de la AQP11 sugieren que estos podrían constituir los candidatos más prometedores para futuras investigaciones. Hasta 15 posiciones de secuencia seleccionadas positivamente (una de hecho compartida entre tres linajes) corresponden a seis ramas de AQP11 diferentes. Además, es notable el caso de la selección adaptativa detectada en la rama ancestral de la AQP11b del clado Gobiidae, que, en nuestro conjunto de datos, está representada por los saltarines del fango (de modo de vida anfibio) y dos especies emparentadas totalmente acuáticas. En dicho clado, identificamos un cambio en el motivo NPA donde la N (asparagina) canónica es sustituida por una S (serina). La modificación de la secuencia de esta AQP11b ocurrió antes de la evolución del linaje de los saltarines del fango y podría representar un posible caso de exaptación en Gobiidae. Concluimos que, debido a la importancia de las acuaporinas en la osmorregulación de los peces, las posiciones de secuencia encontradas bajo selección adaptativa podrían haber dado lugar a modificaciones en la estructura o función de estas proteínas que podrían haber jugado un papel en las transiciones agua-tierra de los peces anfibios estudiados.

INTRODUCTION

INTRODUCCIÓN

Aquaporins or major intrinsic proteins: function, evolution, and diversity

Water is the major component of cells and tissues in living organisms. In liquid water, many molecules can dissolve and the essential biochemical reactions for life can occur. Besides, as a fluid, liquid water can easily move across different cells, tissues, and organisms maintaining the homeostasis, which is key for the proper function of many physiological processes, such as respiration, digestion, circulation, etc. It is known that the lipid bilayer of cells allows the transport of water by simple diffusion (Spring, 2011). However, the observed evidence of rapid water flux in toad bladders, mammalian kidneys and red blood cells (reviewed in Finkelstein, 1987), along with the pharmacological inhibition of water transport by low concentrations of mercurial sulfhydryl (Macey, 1984), suggested the presence of a channel-mediated water transport. Several groups tried to isolate these putative water channels, nevertheless, it was not until 1992 when the first evidence of water transport through a protein channel was described (Preston et al., 1992). Such protein, originally named CHIP28 due to its molecular weight of 28kDa, was isolated from human erythrocytes and renal tubules membranes (Preston & Agre, 1991). After that, the term aquaporin (AQP) was established to name this family of water transporters and the CHIP28 protein was renamed AQP1 (Agre, Preston, et al., 1993; Agre, Sasaki, et al., 1993). Studies based in sequence similarities revealed homologies among these aquaporins and members of an ancient family of transmembrane proteins named major intrinsic proteins (MIP) (Chrispeels & Agre, 1994; Pao et al., 1991). This superfamily was named after the characterisation of a protein from the bovine lens fiber membrane (Broekhuyse et al., 1976; Gorin et al., 1984). Early phylogenetic studies proved that AQPs are part of the MIP superfamily and recovered an ancient split between water (AQPs) and glycerol transporters (GLPs) that occurred early in the evolution of the cell (Heymann & Engel, 1999; Park & Saier, 1996; Zardoya, 2005). Since then, the term AQP has been generally used as a synonym of MIP, to name the entire superfamily, and simultaneously just to name the clade of water channels (e.g. Finn et al., 2014; Zardoya, 2005; Zardoya & Villalba, 2001). In general, studies focused on animals—including humans—preferred the term AQP to name the different paralogues within the entire superfamily, naming the GLPs as aquaglyceroporins (e.g. Cerdà & Finn, 2010; Finn et al., 2014; Ishibashi, 2006; Yilmaz et al., 2020). Conversely, MIPs is more broadly used when referring to plants, unicellular eukaryotes or in the context of the entire Tree of Life (e.g. Abascal et al., 2014; Anderberg et al., 2011; Khabudaev et al., 2014; Tesan et al.,

2021). Following the cited literature, in this Ph.D. thesis we use the term MIPs—distinguishing between AQPs and GLPs—when discussing the diversity of this superfamily within eukaryotes (Chapter II) and the term aquaporin or AQP to name all the amphibious fishes orthologues (Chapters III and IV).

MIPs constitute a highly diverse superfamily. Orthologues of these proteins have been well characterised in organisms ranging from bacteria, yeast, animals, and plants (Abascal et al., 2014; Heymann & Engel, 1999; Zardoya, 2005). In prokaryotes, five clusters of MIPs have been described: AqpM, AqpN, AQPX, AqpZ, and GlpF (Finn et al., 2014; Tesan et al., 2021). This division suggests the polyphyly of the term AQP and provides a far more complex interpretation of the origin of eukaryotic MIPs. The first eukaryotic common ancestor (FECA) could have evolved from Asgard archaeal cells (Zaremba-Niedzwiedzka et al., 2017). Consequently, if as suggested the five MIP clades appeared very early in the evolution of the cell, FECA should have encoded at least one copy of each. However, in the eukaryogenesis process from FECA to the last eukaryotic common ancestor (LECA), many events of duplications generated a fully-fledged expansion within the eukaryotic genome (Makarova et al., 2005; Vosseberg et al., 2021). The term LECA refers to the last common ancestor of extant eukaryotes plus extinct post-LECA lineages (Eme et al., 2017; Margulis et al., 2006). In contrast, FECA represents the oldest common ancestor of all eukaryotes that ever existed that is not also an ancestor of an extant archaeal lineage (Eme et al., 2017; Roger et al., 2017). Several studies have shown that most of the hallmarks related with eukaryotic cells could have arisen during the eukaryogenesis process being present in LECA. These include, for example, mitochondria, complex cell cycle with meiosis, intricate intracellular organisation with an endomembrane system and organelles, actin- and tubulin-based cytoskeleton enabling intracellular trafficking and cell motility, nucleus with linear chromosomes, different chromatin states, and regulation of gene expression (reviewed in Eme et al., 2017). Early phylogenetic studies divided eukaryotes into six major supergroups: Excavata, Rhizaria, Chromalveolata, Plantae, Amoebozoa, and Opisthokonta—containing Fungi and Animalia (Cavalier-Smith, 2002; Stechmann & Cavalier-Smith, 2002). Nonetheless, more recent phylogenomic studies have shown a far more complex evolution by describing new unicellular eukaryotes clades such as TSAR (Telonemia, Stramenopila, Alveolata, and Rhizaria), Haptista and Cryptista, as well as some orphan taxa (Burki et al., 2020).

In Eukarya, there are up to seven recognised families of land plants MIPs: plasma membrane intrinsic protein (PIP), tonoplast intrinsic protein (TIP), Nodulin 26-like intrinsic protein (NIP), small basic intrinsic protein (SIP), X or uncharacterised intrinsic protein (XIP), hybrid intrinsic protein (HIP), and GlpF-like intrinsic protein (GIP) (Abascal et al., 2014; Danielson & Johanson, 2008; Gustavsson et al., 2005; Soto et al., 2012; Strassert et al., 2021). GIPs, PIPs, SIPs, and TIPs are also present in green algae and can be traced back to the common ancestor of all the Viridiplantae clade (Anderberg et al., 2011; Li et al., 2022). Besides, green algae encode five exclusive subfamilies (named MIP A–E) that are not present in land plants (Anderberg et al., 2011). Likewise, NIPs emergence seems to have occurred earlier to the diversification of all plants (Archaeplastida) (Li et al., 2022). In animals, four well defined clusters are described (AQP1-like, AQP8-like, AQP3-like, and AQP11-like)(Finn et al., 2014; Soto et al., 2012). Recent phylogenetic studies suggested that PIPs and AQP1-like orthologues—which are basically water channels—are clustered together forming the classical AQPs (Abascal et al., 2014; Finn et al., 2014; Soto et al., 2012). TIPs form a clade with the animal AQP8-like, and SIPs with AQP11-like, although this latter may be due to a long-branch attraction artefact (Abascal et al., 2014; Soto et al., 2012). AQP3-like channels and Bacteria GLPs seem to share a common ancestor, suggesting that the term GLP refers to a monophyletic group that also includes plant GIPs (Abascal et al., 2014; Soto et al., 2012). Finally, plant NIPs seem to have evolved from horizontal gene transfer (HGT) from a cyanobacteria (Abascal et al., 2014; Zardoya et al., 2002). In particular, a recent study on NIPs has indicated that the most probable hypothesis is that these proteins evolved from a bacterial AQPN sequence (Pommerrenig et al., 2020). In contrast to animals and plants, very little is known about other supergroups, such as Discoba, TSAR, and Haptista. Nevertheless, with the advent of the genomic era and the rapid generation of new genome sequences, MIP diversity has been increasingly explored in other eukaryotic groups such as fungi (Pettersson et al., 2005) and the oomycete *Phytophthora* (Azad et al., 2021). Within the TSAR supergroup, some MIPs cluster with the families PIP, GIP, and MIPE, whereas other MIPs cluster in a new family specific to TSAR organisms, named Large Intrinsic Proteins (LIPs) (Khabudaev et al., 2014). The description of a new clade of AQPs (AQPX) only in unicellular eukaryotes and prokaryotes (Tesan et al., 2021), along with the polyphyletic AQP clade suggest a more complex pattern of evolution within eukaryotes than previously thought (Abascal et al., 2014; Soto et al., 2012; Zardoya, 2005).

MIPs function and their permeants are also highly diverse across the different organisms. Even though most of these proteins are able to transport water—suggesting a role in water homeostasis that likely arose very early during the evolution of the cell—some of them, such as the plant NIP6 orthologue, are not able to transport water when expressed in amphibian oocytes (Pommerrenig et al., 2020; Wallace et al., 2006). MIPs have also shown ability to transport glycerol, urea, lactic acid, reactive oxygen species (ROS), hydrogen peroxide, gases (like ammonia, carbon dioxide, and nitric oxide), and metalloids (like boron, silicon, arsenic, and antimony) (reviewed in Finn & Cerdà, 2015). MIPs have been associated with several physiological processes such as: osmoregulation in fishes (Madsen et al., 2015), defence against pathogens in plants (Li et al., 2020), mitochondrial polarisation and function in humans (Marchissio et al., 2012), water maintenance in tetrapods (Finn et al., 2014). They are also important targets for drugs against human diseases and parasites (Munday et al., 2015; Soveral & Casini, 2017) and central to the development of drought tolerant crops (Maurel et al., 2015). The ancient functional division among water and glycerol transporters—based on the difference between AQPs and GLPs (Heymann & Engel, 1999)—was initially questioned by studies that described overlapping functions among different paralogues (reviewed in Finn & Cerdà, 2015; Zardoya, 2005). For instance, AQP6 orthologues—which are AQP1-like and therefore classified as water channels (Finn et al., 2014)—can also transport large molecules such as glycerol and urea (Holm et al., 2004). Likewise, the ability for ammonia transport has been described in AQP1 and 8-like, TIPs-like, and GLPs orthologues (reviewed in Finn & Cerdà, 2015). Finally, MIPs have been associated with the transport of reactive oxygen species, suggesting that this capacity likely evolved in most eukaryotic MIPs (Bienert & Chaumont, 2014).

In functionally characterised MIPs, the selectivity to transport depends on two sets of conserved amino acids that define the size and affinity of the pore: (I) two opposite asparagine-proline-alanine (NPA) motifs that form hydrogen bonds with the water molecule and electrostatically repulse protons (Murata et al., 2000) and (II) four residues—known as the aromatic/arginine (ar/R) selectivity filter—that form the narrowest pore section and determine substrate specificity (Fu, 2000). In addition, five amino acid residues known as P1–P5 or Froger positions are important in defining MIP subfamilies and in substrate transport selectivity (Froger et al., 1998). Of these, P2 (located downstream the second NPA motif) is crucial in increasing pore size to allow for transport of larger molecules, such as glycerol (Hub & de Groot, 2008). The

reconstruction of the molecular structure of AQP1 showed that this protein is constituted by monomers of six α-helices (1–6) embedded in the cell membranes that are connected by five regions (A–B) that loop at either side of the membrane (loops A, C and E are extracellular whereas loops B and D are intracellular) (Cheng et al., 1997; Walz et al., 1997). Each monomer constitutes a water channel and clusters with other three to form tetrameric structures in the membranes (Fig. 1; Bienert et al., 2012). The central area of the tetramer constitutes a fifth pore (delimited by the junctions of the monomers) for which effective transport through has been reported as well (reviewed in Ozu et al., 2018). This basic molecular pore structure remains highly conserved in most of the different MIP paralogues that have been discovered more recently (Kruse et al., 2006).

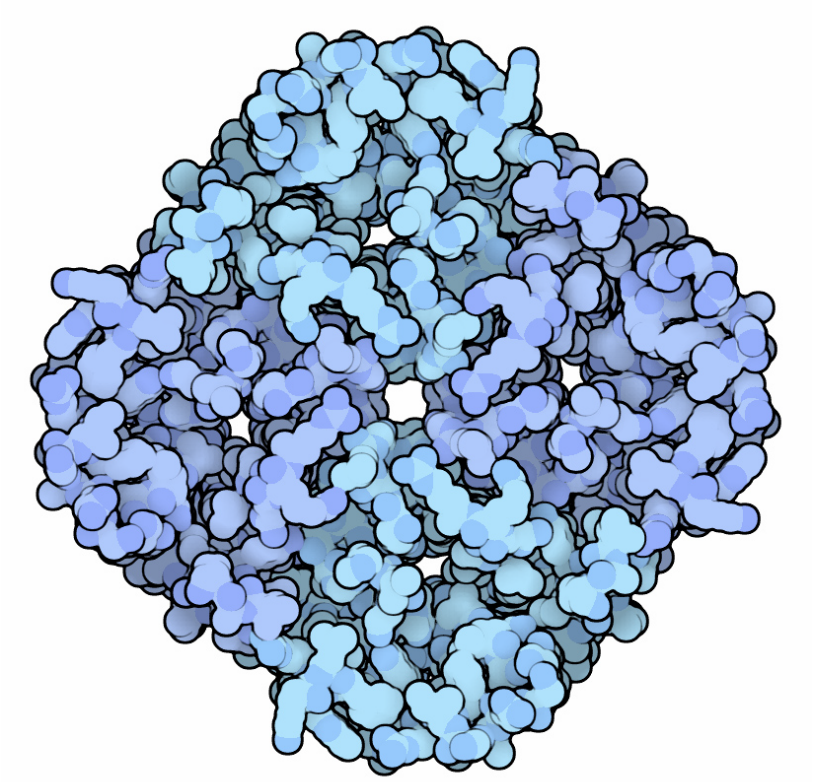


Figure 1. Space-fill 3D drawing (top view from outside the cell membrane) of the tetrameric structure of an aquaporin (from RCSB Protein Data Bank file 1FQY). Each of the four identical monomer (distinctly highlighted in two blue tones) contain a pore/channel at its centre. The squared configuration of the four monomers delimits a channel at the centre of the tetramer that constitutes a fifth pore. Drawing by David Goodsell retrieved from https://wordpress.org/openverse/ with Creative Commons open license CC BY 3.0.

MIPs regulation, inhibition, and stabilisation seems to rely on point changes at the sequence level. For instance, in some MIPs, mercury inhibition was suggested to be connected to a cysteine involved in ar/R constriction (Maclver et al., 2009; Tingaud-Sequeira et al., 2008, 2009, 2010; Watanabe et al., 2005). However, other MIPs do not present this cysteine and they can be inhibited by mercury as well (Maclver et al., 2009; Tingaud-Sequeira et al., 2010). MIPs function can be regulated by gating and trafficking (reviewed in Kreida & Törnroth-Horsefield, 2015). Gating refers to the conformational changes that can prevent permeant pass through the pore (Hedfalk et al., 2006; Törnroth-Horsefield et al., 2006). In MIPs, gating can be mediated by phosphorylation, pH, and Ca2+ concentration (reviewed in Wspalz et al., 2009). On the other hand, trafficking refers to the molecular transport of proteins to the membrane by means of vesicles (Törnroth-Horsefield et al., 2010). In lungfishes, the transport to the membrane (Öberg et al., 2011) of an AQP2-like paralogue is activated by the phosphorylation of a C-terminal serine (Finn et al., 2014; Konno et al., 2010). In humans, the glycosylation of an asparagine in loop C of an AQP10 orthologue could be involved in tetramer stabilisation (Öberg et al., 2011), and, in plants, a conserved cysteine has been proposed as key for the proper bond of MIP monomers (Bienert et al., 2012).

Adaptation from water to land in amphibious fishes

Adaptation to new, unexplored environments is generally challenging, but can also provide advantageous conditions that could drive adaptive evolution (Graham, 1997; Nosil, 2012; Schluter, 2000). In this sense, water-to-land transitions appears to be among the most remarkable and challenging adaptations in the history of life (Laurin, 2010; Long & Gordon, 2004). Water provides a buoyant medium where osmoregulation is a relatively straightforward process, whereas terrestrial environments pose higher gravitational pressure and desiccation conditions to organisms. Therefore, emersion from water involves several morphological (biomechanical) and physiological (metabolic and biochemical) changes, mostly associated with locomotion, vision, audition, respiration, and desiccation (Graham, 1997; Laurin, 2010; Sayer, 2005). These restrictions have made that just a small fraction of all of the major clades of plants, animals, fungi, and microbes have diversified on land (Grosberg et al., 2012). In vertebrates, the emergence of tetrapods from their ancestral aquatic habitats during the Devonian entailed a key event that ultimately led to the successful radiation of vertebrates in terrestrial environments

(Carrol, 2001). To cope with aerial respiration, these animals evolved lungs (Morris, 1892), and the acquisition of limbs allowed them to move on land (Schneider & Shubin, 2012; Shubin et al., 1997). Air-breathing can be traced back early in the evolution of vertebrates (Daniels et al., 2004), as are the genes involved in the formation of lungs and limbs that appear to be ubiquitous among all jawed vertebrates (Gnathostomata) (Freitas et al., 2012; Ruvinsky & Gibson-Brown, 2000). In a similar way, Finn et al. (2014) proposed that three aquaporin classes (AQP2, AQP5 and AQP6) could have appeared very early in the evolution of osteichthyan fishes (Euteleostomi), but they were only retained in the Sarcopterygian branch, presumably contributing to water conservation against desiccation. In mammals, an AQP2 channel located in the apical membrane of the cells of renal collecting ducts mediates in water conservation (Boone & Deen, 2008; Deen et al., 1994). Likewise, avian AQP5 is transcriptionally downregulated in the nasal gland in order to promote hyperosmotic salt secretion and water conservation (Müller et al., 2006). In amphibians and mammals, an AQP2-like protein localised in the kidney tubules is involved in water conservation in response to arginine vasotocin (Ogushi et al., 2007). Likewise, an AQP5-like is suggested to aid in the maintenance of moist skin, cutaneous gas exchange and thermoregulation (Suzuki & Tanaka, 2009), and it is thus partly reminiscent of mammalian exocrine sweat glands, which also express AQP5 (Takata et al., 2004). These three aquaporin paralogues are included in the so-called water-selective classical aquaporins. In vertebrates, up to 17 aquaporins classes have been discovered thus far (Finn et al., 2014; Yilmaz et al., 2020). In this animal group, the term AQP is generally used as a synonym of MIP, and therefore, it refers to the entire superfamily (e.g. Cerdà & Finn, 2010; Finn et al., 2014; Finn & Cerdà, 2015; Ishibashi, 2006). Thus, we can find glycerol transporters named as AQPs (e.g. AQP3 and AQP7). The vertebrate AQP classes or subfamilies can be classified in four main groups: the aquaglyceroporins or AQP3-like—traditionally containing aquaporin classes/paralogues AQP3, AQP7, AQP9, AQP10 and AQP13—, AQP8-like or ammonia channels—containing AQP8 and AQP16 classes—, unorthodox or super aquaporins or AQP11-like—containing classes AQP11 and AQP12—, and water-selective classical aquaporins or AQP1-like—AQP0, AQP1, AQP2, AQP4, AQP5, AQP6, AQP14 and AQP15 (Abascal et al., 2014; Cerdà & Finn, 2010; Tingaud-Sequeira et al., 2010; Yilmaz et al., 2020).


A
B
C
D
E
F
G
H
I
J

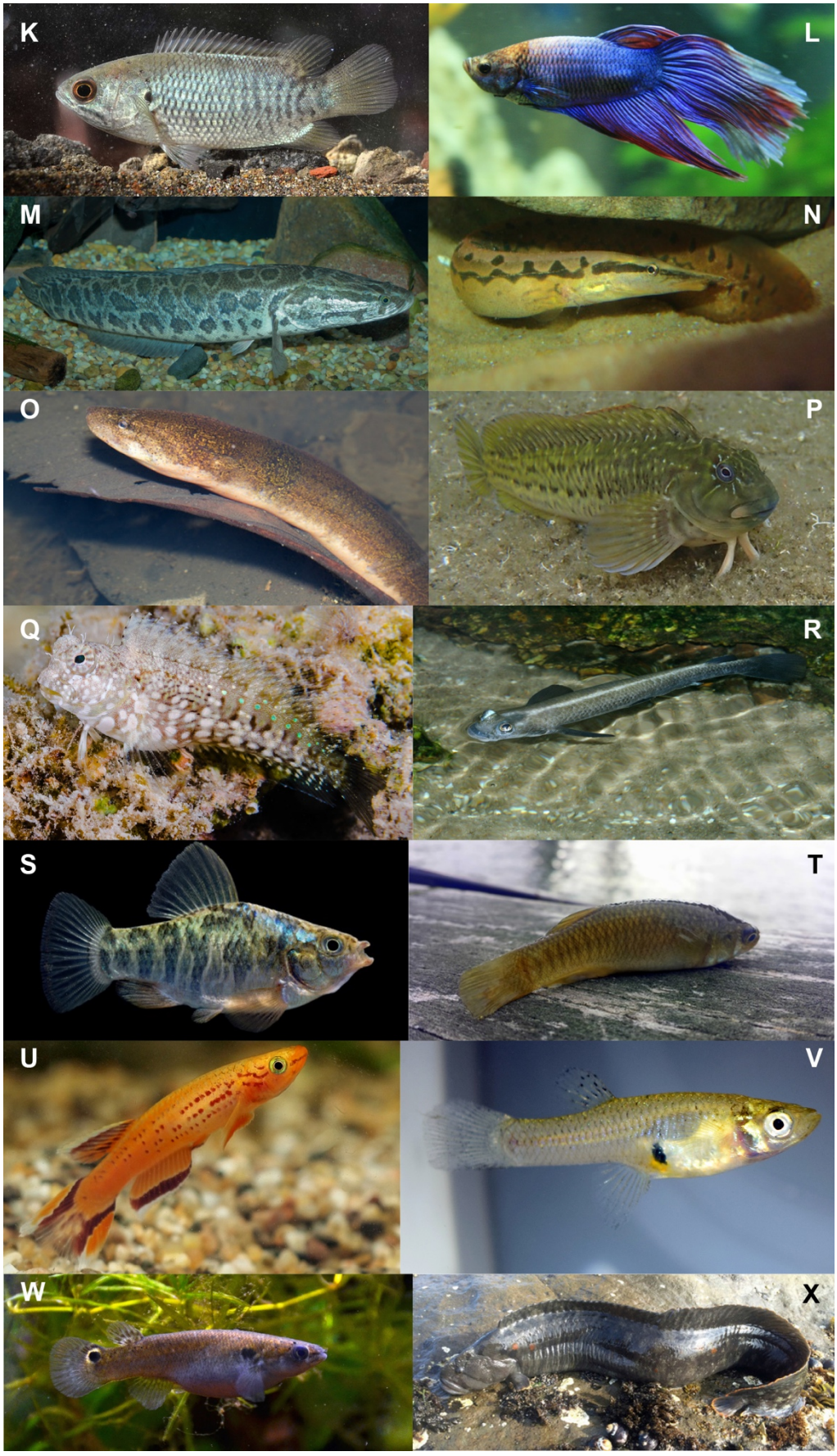

K
L
M
N
O
P
Q
R
S
T
U
V
W
X

Figure 2. Diversity of amphibious fishes. Species name and author/attribution of each photo (indicated in parentheses) are: A) *Protopterus annectens* (Takada@mex), B) *Neoceratodus forsteri* (Mitch Ames), C) *Erpetoichthys calabaricus* (Daiju Azuma), D) *Polypterus senegalus* (5snake5), E) *Anguilla anguilla* (Belarusian Backwoods), F) *Clarias batrachus* (rc6750), G) *Boleophthalmus pectinirostris* (yaoho93), H) *Periophthalmodon schlosseri* (berniedup), I) *Periophthalmus magnuspinnatus* (tommyhui), J) *Scartelaos histophorus* (wildsingapore), K) *Anabas testudineus* (Abu hamas), L) *Betta splendens* (h080), M) *Channa argus* (brian.gratwicke), N) *Mastacembelus armatus* (Adolsomee), O) *Monopterus albus* (budak), P) *Parablennius parvicornis* (stanvrem), Q) *Salarias fasciatus* (zsispeo), R) *Anableps anableps* (Quartl), S) *Cyprinodon variegatus* (reptiles_and_wildlifeliny), T) *Fundulus heteroclitus* (pmk00001), U) *Aphyosemion australe* (Dornenwolf), V) *Gambusia affinis* (David A. Hofmann), W) *Kryptolebias marmoratus* (Vassil), X) *Cebidichthys violaceus* (anudibranchmom). All photos retrieved from https://wordpress.org/openverse/ with Creative Commons open licenses CC BY-SA 2.0, 3.0, and 4.0.

Apart from early tetrapod lineages, the achievement of an amphibious lifestyle has evolved independently many times in vertebrates, and there are reports of over 200 extant species of actinopterygian fishes (spanning 17 orders) that present various degrees of terrestrial adaptation (see e.g. Fig.2; Graham, 1997; Ord & Cooke, 2016; Thompson et al., 2017; Wright & Turko, 2016). These, dubbed amphibious fishes, typically inhabit intertidal areas, taking refuge in small pools during low tides (Bridges, 2015; Graham, 1997; Ord & Cooke, 2016). It has been hypothesised that the low concentrations of O_2 available in such pools could be the main driver for fish emersion (e.g. Graham, 1997; Martin, 1995; Randall et al., 1981; Turko et al., 2012). During terrestrial adaptation, fish gills tend to collapse (e.g. Long & Gordon, 2004; Sayer, 2005; Sayer & Davenport, 1991; Turko et al., 2012) and gas exchange switches to other organs, such as the skin and the gastrointestinal tract (reviewed in Graham, 1997). Consequently, in the absence of gills normal function, amphibious fishes have also developed alternative air-breathing strategies, such as gulping air (reviewed in Graham, 1997). On the other hand, gills are responsible of osmo- and ionoregulation in normal conditions. Hence, when emerged, the skin of amphibious fishes has to take the main role as water and ion transporter (reviewed in Pelster & Wood, 2018; Wright & Turko, 2016). Some amphibious fishes can produce mucus to prevent dehydration while maintaining O_2 uptake (e.g. LeBlanc et al., 2010; Zhang et al., 2000, 2003). However, apart from low O_2 conditions, predator avoidance, competition, foraging, migration, and simply drought conditions are all important drivers

of amphibious behaviour (reviewed in Sayer & Davenport, 1991). Each of these activities requires some degree of terrestrial locomotion. Out of water, amphibious fishes can move by performing axial or appendage movements, or, more commonly, a combination of both (reviewed in Pace & Gibb, 2014). Likewise, the pectoral fins are the main appendages used for walking, although some can also use their operculum (Davenport & Matin, 1990; Flammang et al., 2016). For instance, the climbing perch (*Anabas testudineus)* can walk up obstacles that are at least half its body length (Davenport & Matin, 1990). Both the European eel (*Anguilla anguilla)* and the reedfish *(Erpetoichthys calabaricus*) have an elongated body form (Gillis, 1998; Pace & Gibb, 2011) that facilitates displacement by undulations of their body axis, thus creating a thrust through a body-tail wave (Ellerby et al., 2001; Gillis, 1998; Pace & Gibb, 2011, 2014). This movement may resemble those observed in other terrestrial vertebrates with elongated body forms, such as snakes, although their efficiency in amphibious fishes is substantially lower (Pace & Gibb, 2014). On the other hand, locomotion is an energy-intensive mechanism due to the lack of buoyancy in terrestrial environments. Some lungfishes (Dipnoi) are able to aestivate, reducing their metabolism, thus avoiding the production of toxic excretion products that are more difficult to eliminate during drought conditions (Janssens, 1964; Smith, 1930). Likewise, other amphibious fish species such as the Asian swamp eel (*Monopterus albus*) have the ability to reduce the proteolysis and catabolism of amino acids under low O_2 conditions, in some cases practically stopping the endogenous production of ammonia (Chew, Gan, et al., 2005; Ip, Tay, et al., 2004; Tay et al., 2003). Higher ammonia tolerance and the ability to actively excrete it appear to be a widespread evolutionary solution in amphibious fishes (e.g. Chew, Gan, et al., 2005; Ip, Chew, et al., 2004; Ip, Randall, et al., 2004; Randall et al., 2015). This molecule is extremely toxic and it needs a large amount of water to be diluted (reviewed in Chew, Wilson, et al., 2005; Ip et al., 2001; Wright, 1995). Moreover, ammonia concentration can be remarkably high in the small pools where many amphibious fishes usually inhabit (Ishimatsu et al., 1998; Ishimatsu & Gonzales, 2011; Taylor, 2000).

Amphibious behaviour of the different fish groups is not equally well documented, and much of our knowledge is limited to a few clades, namely oxudercides (mudskippers), blennies (rockskippers), killifishes, and lungfishes (reviewed in Ord & Cooke, 2016; Thompson et al., 2017; Turko & Wright, 2015). Among these, mudskippers (suborder Gobioidei; family Oxudercidae, Periophthalmus-lineage) represent a remarkable case with a relatively high degree of terrestrialisation (Wright & Turko, 2016;

You et al., 2014). These amphibious fishes can emerge during low tide and keep very active on land (i.e., to forage, seek mates and defend territories) (Jaafar & Murdy, 2017; Wright & Turko, 2016). They can gulp air and obtain oxygen through the skin and a highly-vascularised buccopharyngeal epithelium as well (Graham, 1997; Randall et al., 1981; Zander, 2011). Using their pectoral fins, they can perform long jumps—in stress response or predation avoidance—and even 'crutch' (resembling human move on crutches) (Harris, 1960; Pace & Gibb, 2009; Swanson & Gibb, 2004). To cope with predators, they have also developed a great areial vision that, in terms of colour recognition, can be comparable to humans (You et al., 2018). Mudskippers can also reduce evaporative water loss through their skin in dry environments (Dabruzzi et al., 2011) and produce mucus to prevent dehydration while maintaining O_2 uptake (Zhang et al., 2000, 2003). Besides, one special type of cells—likely ubiquitous in all mudskippers species—named the swollen middle cells act as a water warehouse (Zhang et al., 2000, 2003). However, their dependence of water is still important for reproduction (they possess aquatic larvae) and to minimise desiccation. For this latter, mudskippers roll on mud, immerse in small pools that remain at low tides, or frequently move from land to burrows built in the mud that are submerged at high tides or retain water inside at low tides. These burrows are also used as refuges to hide from predators and for laying eggs (Graham & Lee, 2004; Ishimatsu & Gonzales, 2011; Tsuhako et al., 2003; Wright & Turko, 2016). Mudskippers live on mudflats and mangrove swamps and have colonised peritidal habitats of tropical and subtropical Western Africa, the Indian Ocean, and the whole Indo-West Pacific region (Murdy, 2011). Only four oxudercide genera are strictly considered to be mudskippers, namely *Boleophthalmus*, *Periophthalmodon*, *Periophthalmus* and *Scartelaos,* and they present different degrees of adaptation to terrestrial conditions. *Periophthalmus* and *Periophthalmodon* are the most terrestrially-adapted, spending the majority of time out of water, although their degree of terrestriality gradually varies among species of these genera. In contrast, *Scartelaos* is predominantly aquatic (spending less time out of water) followed by less water-dependent *Boleophthalmus* (Graham, 1997; Ishimatsu & Gonzales, 2011; Polgar et al., 2017; You et al., 2014; Zhang et al., 2003). Apart from terrestrial behaviour, the differences among genera can also be found in genetic and physiological aspects (You et al., 2014). For instance, within the *Periophthalmodon* clade there are species that can tolerate and excrete ammonia more efficiently than other mudskippers (Chew et al., 2003, 2007; Ip, Randall, et al., 2004; Randall et al., 2015).

In contrast to tetrapods, the relationship between aquaporins and the terrestrialisation of amphibious fish remains largely unknown thus far. Only in lungfishes there is some evidence of the involvement of aquaporins in their adaptation to an amphibious lifestyle. Like other extant sarcopterygians (Meyer et al., 2021), lungfishes encode an AQP2-like that is involved in water recovery in the kidney (Konno et al., 2009, 2010). This mechanism involves the hormone arginine vasotocin (AVT), which also seems to modulate the amphibious response in mudskippers (Sakamoto et al., 2015). However, like other actinopterygian fishes, mudskippers lack orthologues for the AQP2, AQP5 and AQP6 classes (Cerdà & Finn, 2010; Finn et al., 2014; Finn & Cerdà, 2011; Tingaud-Sequeira et al., 2010), and there is no evidence of aquaporin regulation mediated by the AVT hormone in this group of fishes. On the other hand, AQP3 appears to be downregulated in embryos of the mummichog (*Fundulus heteroclitus)* during aerial exposure, probably to reduce water loss (Tingaud-Sequeira et al., 2009). In contrast, AQP1 is upregulated in the gills of the anabantoid *A. testudineus* during emersion, maybe to increase ammonia excretion (Ip et al., 2013).

More broadly, the role of AQPs in fish osmoregulation has been extensively researched. Seawater fish must actively drink water and excrete salt to prevent desiccation, while freshwater fish gain water through their gills (Evans, 2008; Grosell, 2010; Larsen et al., 2014). When fishes acclimate to seawater conditions, aquaporins tend to be downregulated in the gills and upregulated in the intestine (reviewed in Madsen et al., 2015). This mechanism could be related to water maintenance in the gills and with water recovery in the intestine. However, the pattern of expression is not constant among fish species, conditions, tissues, or even aquaporin classes. For example, both AQP11 and AQP12 paralogues are downregulated in seawater conditions in the roughskin sculpin (*Trachidermus fasciatus*) intestine (Ma et al., 2020). In several fishes, AQP3 is downregulated in gills during seawater acclimation (e.g. Giffard-Mena et al., 2007; Lignot et al., 2002; Madsen et al., 2014; Tipsmark et al., 2010). However, in *F. heteroclitus*, this occurs when it acclimates to freshwater under hypoxia conditions (Ruhr et al., 2020). In the kidney of the European eel (*Anguilla anguilla*) and the Atlantic salmon (*Salmo salar*), AQP3 is upregulated when facing high salinity concentrations, whereas in the Mozambique tilapia (*Oreochromis mossambicus*) AQP3 is downregulated in similar situations (Martinez et al., 2005; Tipsmark et al., 2010; Watanabe et al., 2005). In this sense, terrestrial and marine conditions can be compared in terms of water maintenance. For example, cortisol is usually classified as a ‘seawater-acclimation' hormone (reviewed

in Breves, 2020), but it can also act as an amphibious-driver hormone in mudskippers (Sakamoto et al., 2011). However, as in the case of the AVT hormone, there is no evidence for the relationship between cortisol-induced amphibious behaviour and an aquaporin gene in mudskippers.

OBJECTIVES

OBJETIVOS

OBJECTIVES

This Ph.D. thesis aims to study the molecular evolution and phylogenetic context of aquaporins (major intrinsic proteins) in amphibious fishes and their role in water-to-land transitions.

The specific objectives of the thesis are the following:

1. To present a solid cross-platform workflow and pipeline for gene isolation from large-scale genomic data retrieved from publicly available databases and phylogenetic reconstruction that can be used for molecular evolution analyses, such as detection of adaptive/purifying selection or ancestral protein reconstruction.

2. To study the phylogeny and molecular evolution of the major intrinsic proteins in the broad context of major eukaryotic supergroups, with special emphasis on unicellular forms, by combining sequence data from a comprehensive eukaryotic database and previous analyses focused on MIP diversity.

3. To infer the MIP catalogue of the last eukaryote common ancestor by investigating deep eukaryotic MIP clades.

4. To identify and catalogue the repertory of aquaporin genes in the genomes of several amphibious fishes, with particular emphasis on mudskippers.

5. To reconstruct the phylogenetic position of the isolated amphibious fish aquaporins in the context of vertebrates with special emphasis on actinopterygian lineages.

6. To detect signatures of branch and site adaptive (positive) selection on the aquaporin sequences of the studied amphibious fishes, and relate these to possible roles in their water-to-land adaptation.

7. To map adaptive sites on the aquaporin 3D structure and infer possible function or regulation effects.

CHAPTER I

Genomic fishing and data processing for molecular evolution research

---

Pesca genómica y procesamiento de datos para investigaciones sobre evolución molecular

This chapter is a reproduction of the following article:

Lorente-Martínez, H., Agorreta, A., & San Mauro, D. (2022). Genomic fishing and data processing for molecular evolution research. *Methods and Protocols*, 5(2), 26.

Abstract

Molecular evolution analyses, such as detection of adaptive/purifying selection or ancestral protein reconstruction, typically require three inputs for a target gene (or gene family) in a particular group of organisms: sequence alignment, model of evolution, and phylogenetic tree. While modern advances in high-throughput sequencing techniques have led to rapid accumulation of genomic-scale data in public repositories and databases, mining such vast amount of information often remains a challenging enterprise. Here, we describe a comprehensive, versatile workflow aimed at the preparation of genome-extracted datasets readily available for molecular evolution research. The workflow involves: (1) fishing (searching and capturing) specific gene sequences of interest from taxonomically diverse genomic data available in databases at variable levels of annotation, (2) processing and depuration of retrieved sequences, (3) production of a multiple sequence alignment, (4) selection of best-fit model of evolution, and (5) solid reconstruction of a phylogenetic tree.

Keywords: genomics; high-throughput sequencing; data mining; gene family; blast search; sequence alignment; phylogeny; molecular evolution.

Resumen

Los análisis de evolución molecular, tales como la detección de selección adaptativa/purificadora o la reconstrucción de proteínas ancestrales, por lo general, implican tres pasos a tener en cuenta para un determinado gen (o familia de genes) de interés en un grupo particular de organismos: el alineamiento de las secuencias, el modelo de evolución y el árbol filogenético. Si bien los recientes avances en las técnicas de secuenciación masiva han dado lugar a una rápida acumulación de datos genómicos en repositorios y bases de datos públicas, el manejo y tratamiento de una cantidad tan grande de información sigue siendo a menudo un desafío. En este estudio, describimos un protocolo de trabajo detallado y versátil destinado a la preparación de conjuntos de datos a partir de genomas para su uso en investigaciones de evolución molecular. El protocolo de trabajo implica: (1) pesca (búsqueda y captura) de las secuencias génicas específicas de interés a partir de datos genómicos taxonómicamente diversos que ya estén disponibles en bases de datos y a niveles variables de anotación, (2) procesado y depuración de las secuencias extraídas, (3) producción de un alineamiento múltiple de secuencias, (4) selección del mejor modelo de evolución y (5) reconstrucción de un árbol filogenético robusto.

Palabras clave: genómica, secuenciación de alto rendimiento, extracción de datos, familia génica, búsquedas de blast, alineamiento de secuencias, filogenia, evolución molecular.

1. Introduction

The relatively low cost and increasing power of modern (high-throughput) sequencing technologies have resulted in a massive increase of the number of genome projects on non-model organisms (Ekblom & Galindo, 2011; Lee et al., 2013). This has prompted the development of numerous databases to store all the generated data. As a logical consequence, figuring out how to extract relevant information from such vast amounts of data becomes fundamental. In this context, bioinformatics has emerged as a key tool for handling and processing sequence data derived from genome-scale sequencing experiments. Proteomics, genomics, transcriptomics, and other new disciplines have emerged as a result of the fusion of programming languages and cutting-edge sequencing technologies. However, gene sequence retrieval and assessment are still arduous and complex processes. High-throughput sequencing (HTS) platforms based on DNA amplification, such as Illumina, typically yield short reads of around 100 base pairs (Lee et al., 2013; Shendure & Ji, 2008); hence, systematic assembly of the data (e.g., reads to contigs, contigs to scaffolds, etc.) is mandatory to obtain final sequences of genes or genomic regions. Molecule-sequencing platforms, such as PacBio or NanoPore, can yield much longer reads, thus reducing the gap between sequencing output data and real gene or genomic sequences (Lee et al., 2013; Schadt et al., 2010). In all cases, the identification and annotation of relevant and meaningful genomic regions always remains a mandatory step, and the comparative analysis of genome sequences is central to such an endeavour (Hardison, 2003). Basically, conserved functions between two organisms are assumed to be encoded in DNA in a similar way. Therefore, similar DNA, RNA, or protein sequences are likely involved in relatively similar functions and assumed to be homologous (orthologues or paralogues). In this context, comparative genomics can make use of sequence alignment and phylogenetic analysis as a framework to try to understand the evolutionary processes that trigger sequence diversification. Knowing the pattern of historical relationships among groups (lineages) of elements (organisms, sequences) allows for possible biases and dependencies derived from shared ancestry to be amended when interpreting a function, structure, or any other pattern involving genes or genomic regions (San Mauro & Agorreta, 2010).

Nevertheless, the pathway from raw reads to gene alignments and phylogenetic trees is not necessarily straightforward but rather challenging and often very intense in

terms of both time and resources (e.g., computing power). In the last few years, a number of programs and pipelines for relatively automated extraction of relevant information from modern sequencing technology outputs have been developed. Of the several tools available, software such as geneid (Alioto et al., 2007), Prokka (Seemann, 2014), or GenMark (Brůna et al., 2020) allow for complete gene mapping all across the genome. In a similar way, several platforms have appeared for RNA-seq analysis, such as the TRUFA web server (Kornobis et al., 2015), eventually intended for sequence assembly (either de novo or referenced) and gene annotation and alignment. In terms of gene isolation, BLAST (Wheeler & Bhagwat, 2007) is certainly the most broadly used tool, but other programs, such as ORTHOSCOPE (Inoue & Satoh, 2019), make use of it for identification and isolation of groups of related orthologues. In general, all these pipelines and web servers are intended for analysis of genomic and/or HTS data oriented toward gene mapping and identification; however, to our knowledge, direct processing and preparation of resulting outputs for molecular evolution analysis are lacking.

In this study, we provide a protocol and pipeline for gene search and capture/isolation (fishing) for particular sets of organisms (at any degree of taxonomic diversity) from large-scale genomic data retrieved from publicly available databases. The isolated gene sequences are subsequently aligned and submitted to robust phylogenetic reconstruction using adequate modelling of the substitution process. Altogether, this constitutes baseline data for conducting molecular evolution analyses, such as detection of adaptive/purifying selection or ancestral protein reconstruction. The protocol described here is relatively unique in its span: from genomic mining to phylogeny reconstruction in a comprehensive step-by-step workflow.

Our protocol is intended to be a cross-platform workflow that can be executed on Linux, macOS, and Windows machines. As mentioned above, programming languages are very useful (often mandatory) when working with genomic-scale data. Therefore, working with command-line interfaces on system terminals (such as those of Linux- and Unix-based macOS) becomes a sensible choice in practice. In order to facilitate particular steps of the process, we have developed a suite of small Python programs (GNFish package), taking advantage of the free-access Biopython project environment (Cock et al., 2009). We strongly recommend using this package, especially for programming language beginners. As an alternative itinerary for the protocol (e.g., for those preferring not to deal with command-line procedures), we additionally

describe most steps using windows-style interfaces with local programs or remote/online web tools. The protocol presented here is perfectly adequate for research studies on just one or multiple gene families alike. Theoretically, it can also be used for complete gene mapping all across the genome, although there are more specific tools and pipelines publicly available for this purpose (such as Prokka (Seemann, 2014) or geneid (Alioto et al., 2007), as detailed above). A schematic flowchart of the main steps of the protocol is shown in Fig. 1.

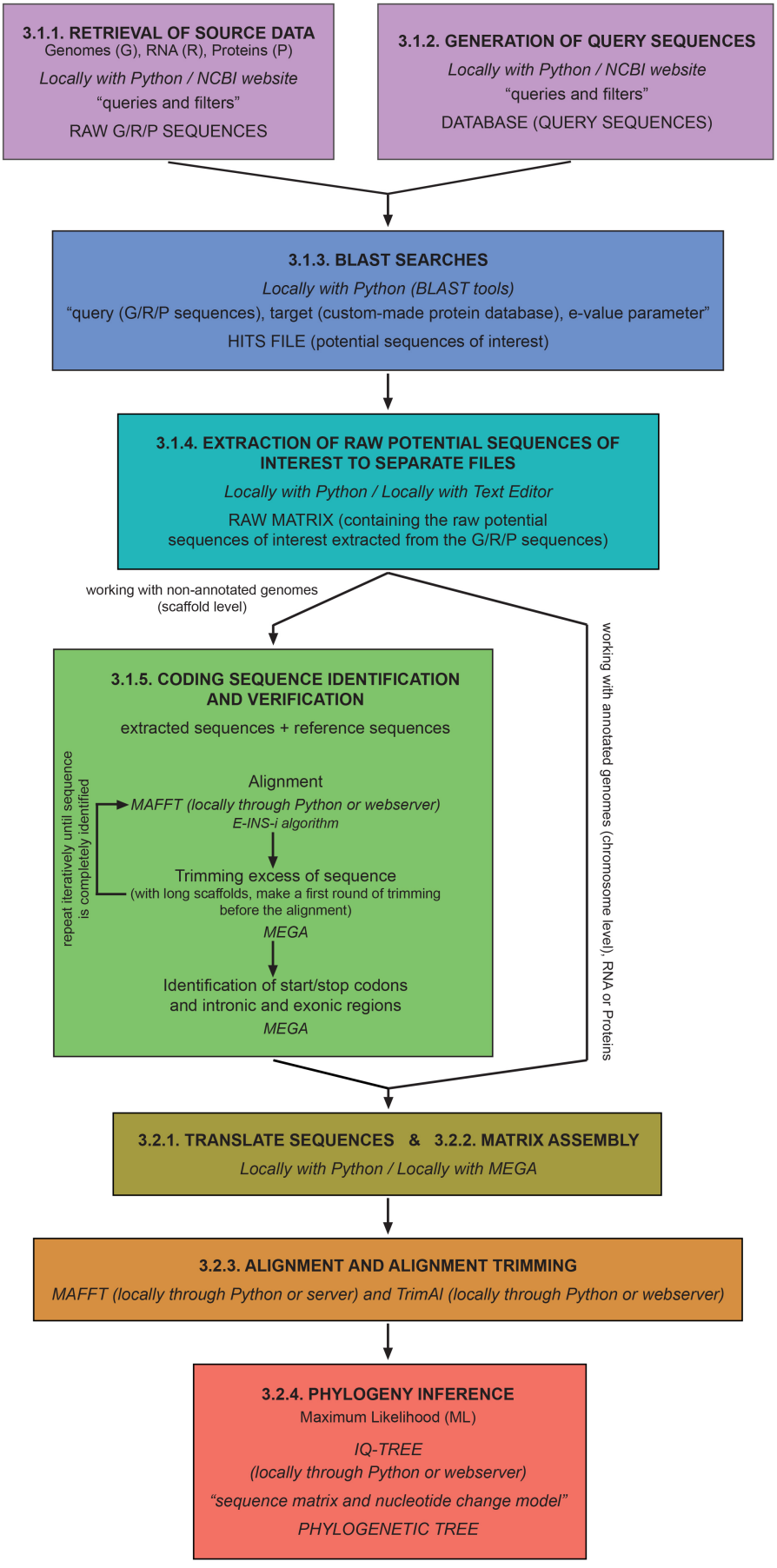


Figure 1. Schematic flowchart of the main steps of the protocol.

2. Experimental Design

2.1. Data Mining and Sequence Identification

Working with whole genome sequences is still an arduous task that is generally not affordable for users not familiar with programming language. A single genome can represent more than one gigabyte of data. Automated annotation programs have facilitated the identification and extraction of protein-coding sequences by mapping potential open reading frames in the genomes to place gene locations. However, the annotation process is not evenly spread among all the available genome data. In some cases, only preliminary rounds of assembly have been developed, resulting in records that contain multiple scaffold entries. These scaffolds correspond to huge portions of the genome sequence in which gene identification has not been fully accomplished yet. Even coding RNA or protein datasets derived from genomic data can contain thousands of entries. Therefore, establishing a solid protocol for specific gene identification is essential when working with genomic data. This becomes of particular interest for research oriented toward molecular evolution insights at any taxonomic level. The first step of such a protocol consists of retrieving the relevant genomic data. This genomic data can be determined anew, although there is certainly a huge amount of genomic data already available and stored in publicly available databases and repositories (such as EukProt (Richter et al., 2020) or Ensembl (Howe et al., 2021)) that remains underused or at least not fully exploited for its entire potential. The National Center for Biotechnology Information (NCBI) comprises a comprehensive collection of connected databases, including gene, protein, genomic, and transcriptomic data, among others (Sayers et al., 2021). Within all the diversity covered by the different databases of the NCBI project, RefSeq emerges as a reliable choice because it maintains and curates a publicly available database of annotated genomic records (Pruitt et al., 2007). Besides, the NCBI Entrez system provides an easy and powerful means to retrieve data (Schuler et al., 1996).

When genomes are stored without annotations, a basic, straightforward method for gene identification can be implemented based on similarity searches of DNA and protein sequences, such as those performed by the BLAST algorithm for local alignment (Altschul et al., 1990). Basically, this tool is able to find related sequences in a database. Each genome can be treated as an independent database, and a list of query

sequences of interest can be used as template. Typically, this query dataset is made of protein sequences, although the target corresponds to nucleotide sequences. Due to the degeneration of the genetic code, some of the changes that occur at the nucleotide level have no impact on the protein primary structure. Therefore, it is more plausible that the protein primary structure and sequence remain more conserved, thus facilitating gene identification. The RefSeq database provides reliable information for protein sequences, and the BLAST tool allows for a certain level of custom parametrization (Wheeler & Bhagwat, 2007). Of these, we simply explore/manipulate two: (1) *e-value*, a parameter that describes the number of hits that one can "expect" to see by chance when searching a database of a particular size, and (2) *outfmt*, which sets the output format.

BLAST output files must be parsed to extract the relevant information, but the specific procedure can vary depending on the type of data. For example, for protein and coding RNA data, every BLAST hit generally corresponds to a unique sequence. Therefore, it is enough to just extract the target-sequence match to obtain the gene sequence of interest. Based on our personal experience, these sequences are usually well annotated. Nonetheless, some hits can contain mismatched positions at the beginning or at the end of the sequences, especially when working with coding RNA sequences because they can present untranslated regions (UTR). On the other hand, working with whole genome sequences or big genomic portions (such as scaffolds) often implies a considerable challenge because of the huge amount of data to be handled and mined, as well as their intrinsically higher complexity. Indeed, single scaffolds typically comprise more than one gene; hence, BLAST output files must be thoroughly inspected to identify possible multiple-gene matches present in each file. Moreover, genomic-level sequences typically include the intronic fractions that must be removed to isolate coding and, ultimately, protein sequences.

In this sense, a relatively simple an effective method to identify coding regions in genomic portions is to align these against some of the query sequences of interest as a template. There is a plethora of programs and algorithms to compute multiple-sequence alignment (Notredame, 2007), but among them, MAFFT (Katoh et al., 2002, 2019) stands among the most reliable and widely used, outperforming more classical programs, such as CLUSTALW or T-COFFEE (Di Tommaso et al., 2011; Thompson et al., 2003). Among its strengths, MAFFT implements several distinct algorithms, each adequate for different types of data and situations (Katoh et al., 2019). One of these, the E-INS-i algorithm, performs well for aligning coding-gene sequences that present

intronic regions against closely related template sequences, such as those resulting from BLAST searches. Once the alignment is completed, the next step consists of cleaning up the (presumed) intronic portions (those mismatched by very long gaps) from the alignment to just isolate the coding-sequence regions. This trimming step can be conducted with general packages for sequence editing and manipulation, such as MEGA, which is a desktop application designed for comparative analysis of homologous gene sequences that allows for alignment visualization and modification with a user-friendly interface (Tamura et al., 2021). It may be the case that the process of alignment and trimming needs to be iteratively repeated and refined (progressively cleaning up intronic portions) until coding regions can be conclusively identified and isolated.

2.2. Phylogenetic Analysis

Studying the molecular evolution of genes (commonly across different organisms) is important to understand the biological processes in which they are involved. For this aim, first establishing a robust framework for the phylogenetic relationships of the different gene sequences involved provides the necessary groundwork that precedes further evolutionary analysis. In our particular context, the phylogenetic experimental design begins with the gathering of all gene sequences of interest (typically from a particular organism or group of organisms) into individual gene matrices (one for each gene of interest). At this point, it may be relevant to also include some external sequences (outgroups) if possible/available. These will provide information about ancestral states, serving as relative indicators of the direction of evolutionary change (e.g., which nodes are the oldest in the phylogenetic tree) (San Mauro & Agorreta, 2010). The next step consists of aligning the sequences included in each of the matrices, which is crucial because all subsequent phylogenetic inferences rely on these alignments (Goldman, 1998; Ogden & Rosenberg, 2006; Phillips et al., 2000). As mentioned above, MAFFT is our recommended choice for this task, either running an automatic strategy (which selects the most adequate alignment algorithm among those available depending on data size) or selecting a particular alignment algorithm (in this case, the L-INS-i may be a good option for sequences from the same gene family). Once the alignment is done, trimming ambiguously aligned positions can increase quality and, consequently, the reliability and accuracy of subsequent analyses

(Talavera & Castresana, 2007). TrimAl is a tool for automated alignment trimming that is especially suited for large-scale phylogenetic analyses (Capella-Gutiérrez et al., 2009). It is free and portable to all platforms, and it can be used online through the Phylemon web server (Sánchez et al., 2011). trimAl implements modes for automated selection of trimming parameters, although the use of some can be computationally demanding, especially when working with very large datasets (a simpler option for genomic-scale data is the *conservation threshold* parameter based on the percentage of gaps).

Apart from robust alignments, proper characterization of the process of sequence evolution is essential in molecular phylogenetic inference (Cunningham et al., 1998) because phylogenetic methods tend to be less accurate or inconsistent when an incorrect model of sequence evolution is assumed (Bruno & Halpern, 1999; Huelsenbeck & Hillis, 1993). Phylogenetic inference in a probabilistic framework, such as maximum likelihood (ML), allows for the estimation of complex substitution model parameters, branch lengths, and tree topology using heuristic methods. In this sense, the IQ-TREE software presents a set of fast and effective stochastic algorithms for ML phylogenetic analysis, including automated assessment of the best-fit substitution model (Nguyen et al., 2015). IQ-TREE implements modern measures of branch support, such as the ultrafast bootstrap approximation approach (UFBoot) (Minh et al., 2013), which can reduce computing times compared to traditional bootstrap. Phylogenetic inference can be conducted at the nucleotide (DNA) or amino acid (protein) sequence level. In general, protein alignments (often obtained by conceptual translation of primary DNA records) are more adequate for inference of old relationships because the higher character-state space (20) of amino acids (compared to 4 of nucleotides) makes it less likely to observe homoplasy events due to sequence saturation (San Mauro & Agorreta, 2010). The process of nucleotide or amino acid substitution is further complicated by the fact that the evolution of sites is often highly heterogeneous, with some sites changing rapidly, whereas others are highly conserved. In general, this heterogeneity of evolutionary rates among sites is modelled using specific parameters, such as a proportion of invariant sites or a discrete approximation (usually with four categories) of the continuous gamma function (Yang, 1993). Furthermore, the substitution process can be affected by other factors, such as solvent exposure and secondary structure (Goldman et al., 1998; Koshi & Goldstein, 1995; Thorne et al., 1996); therefore, more

complex models can be devised to better explain protein evolution (Le et al., 2008). Nevertheless, further discussion of these issues is outside the scope of the present study.

Although we developed our protocol with a preferred list of software for each of the different tasks of the process (retrieval of genomic data, gene sequence identification and isolation, multiple alignment, model selection, and phylogenetic inference), there are often alternatives that can do and perform equally well. We next discuss some of these alternatives. In the case of genome retrieval, EukProt is a database of published and publicly available predicted protein sets and unannotated genomes selected to represent eukaryotic diversity, including species from all major supergroups, as well as orphan taxa (Richter et al., 2020). On the other hand, UniProt is a reference database for protein data that can be used to obtain the query sequences necessary for similarity searches (Bateman, 2019). For this purpose, HMMER can be used as an alternative to BLAST or even in combination with it. HMMER is based on probabilistic models called profile hidden Markov models (profile HMMs) (Finn et al., 2011), and it often works with protein profiles downloaded from Pfam (El-Gebali et al., 2019) and Interpro (Hunter et al., 2009) databases. Like BLAST, HMMER can also work with query sequences. In the case of multiple sequence alignments, CLUSTALW (Thompson et al., 2003), T-COFFEE (Di Tommaso et al., 2011), and PRANK (Löytynoja, 2014) are good alternatives to MAFFT. For phylogenetic inference, IQ-TREE is comparable to other ML programs, such as PhyML (Guindon et al., 2005) and RAxML (Stamatakis, 2014). However, in these best-fit substitution models, selection is not automated (only in the online version of PhyML), and an external program is required for this task, such as ProtTest for protein alignments (Darriba et al., 2011) or jModelTest for nucleotide alignments (Darriba et al., 2012). It is often the case that applying different best-fit models for distinct alignment sections that differ in rates of evolution (e.g., different genes, codon positions, stems vs. loops) might be preferred over averaging a single model for the entire set. Programs such as PartitionFinder (Lanfear et al., 2012) allow for simultaneous selection of best-fit partitioning strategy and substitution models, and this information can be easily implemented in RAxML and IQ-TREE. The latter implements an option for automated selection of partitions and models (see http://www.iqtree.org/doc/Advanced-Tutorial#partitioned-analysis-for-multi-gene-alignments, 27 January 2022). The MEGA package can also be used for phylogenetic inference, as well as for alignment and selection of best-fit model of substitution (Tamura et al., 2021) but without taking into account partitioning schemes.

Finally, phylogenetic trees can be graphically inspected and enriched using several publicly available programs, with FigTree (Rambaut) being our preferred choice because of its versatility and ability to produce publication-ready figures. A good alternative for this task may be Dendroscope (Huson & Scornavacca, 2012).

3. Procedure

3.1. Data Mining and Sequence Identification

3.1.1. Retrieval of Source Data (Genomic, RNA, and Protein)

Locally Using Python

- Go to https://github.com/hectorloma/GNFish (27 January 2022) to obtain the GNFish package. The "README.md" file contains a detailed explanation for the suite of Python programs used throughout this protocol, as well as some running examples.
- The following protocol details how to run Python programs using the Anaconda platform (See below). However, on the "README.md" file, you will find details on how to run it directly on a Linux terminal. Functionality and output files are the same.
- Download the main directory containing the scripts and examples by clicking on Code → Download Zip and decompress the file (Fig. 2).

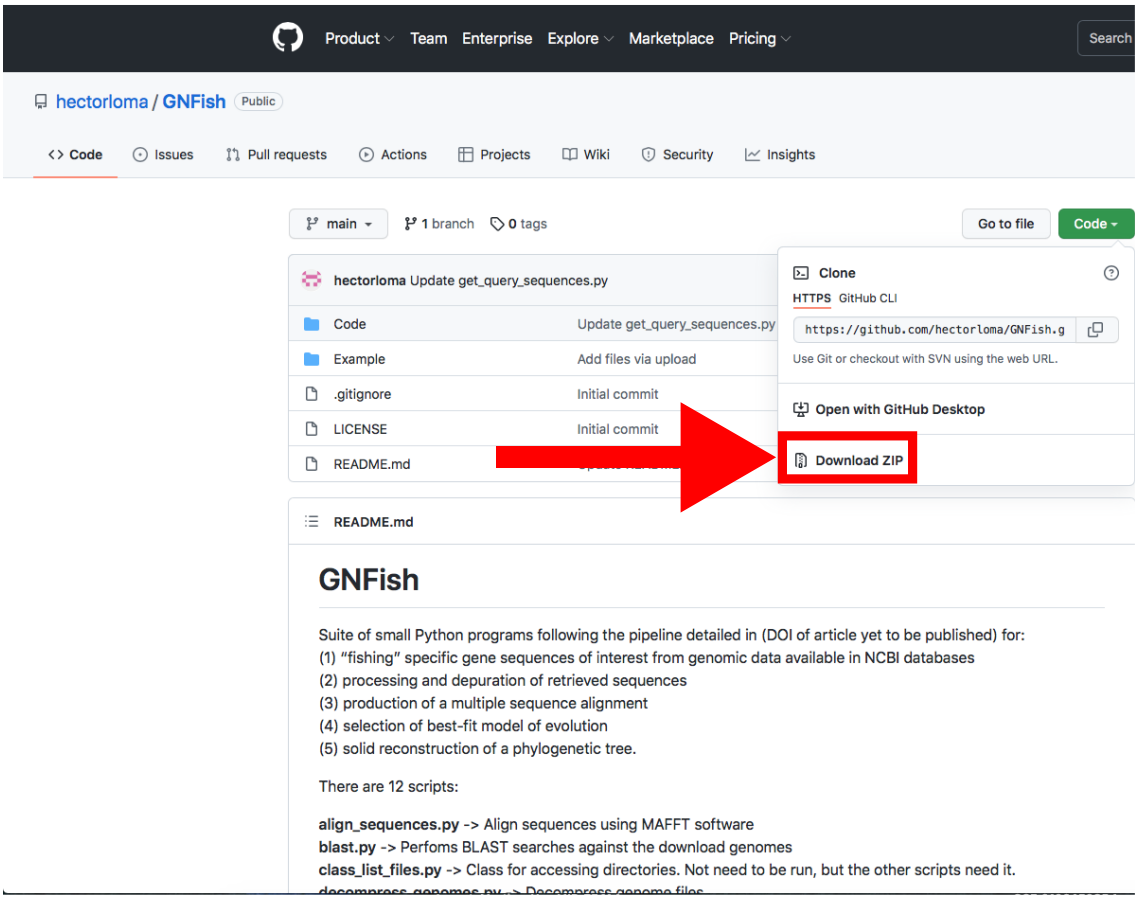


Figure 2. GitHub web server where the GNFish package is stored. The red arrow indicates the button for downloading. The Code button appears in green (top right). Accessed on 27 January 2022.

- Install Anaconda following the instructions at (https://conda.io/projects/conda/en/latest/user-guide/install/index.html, 27 January 2022) and open the Spyder program.
- Install Biophyton (if not installed yet). In the Spyder console (Fig. 3), type pip install biopython.
- Again, in the Spyder console, type cd and drag GNFish/Code directory (remove quotes if you are a Windows user). Note that this step is mandatory every time you close the Spyder program.
- Type *run get_genomes.py -h* in the Spyder console to display help information and read the "README.md" file for further information and some running examples. This applies to all the ".py" programs used throughout this pipeline.

- Create a file with all your queries (usually species or higher taxon names). You can also add specific field tags (e.g., organism, assembly level, etc.) and some filters (e.g., latest [filter] or unambiguous [filter]).
- More information about how to concatenate specific field tags is found at https://www.ncbi.nlm.nih.gov/books/NBK3837/-EntrezHelp.Entrez_Searching_Options, 27 January 2022.
- For information about filters, check out https://www.ncbi.nlm.nih.gov/assembly/help/, 27 January 2022.
- For both filters and field tags, you can obtain more information after conducting a manual search. We recommend first trying a simple query with one taxa as an example following Section 3.1.1 B.
- Run the program typing run get_genomes.py [e-mail address] [path to query file] [data-type] on the Spyder console. Add any optional arguments after these mandatory ones.
- We recommend that you to use the --*refine* argument in order to refine your search. By default, this will apply *Latest*, *Representative, Not Anomalous* but you can add your own settings by typing them, enclosed by quotes, after the argument.
- Use the proper arguments according to your search. By default, the program will download whole-genome data. If downloading protein or RNA data, and they are not available, the program will try with the whole genome version. Stop that feature using --*exclusive* argument.
- Downloaded sequences will be stored into *Genomic*, *RNA* and *Protein* directories located at *GNFish/Code/Data*. Information about the downloaded genomes will be stored at *Data/downloaded_genomes_log.tsv* as well.

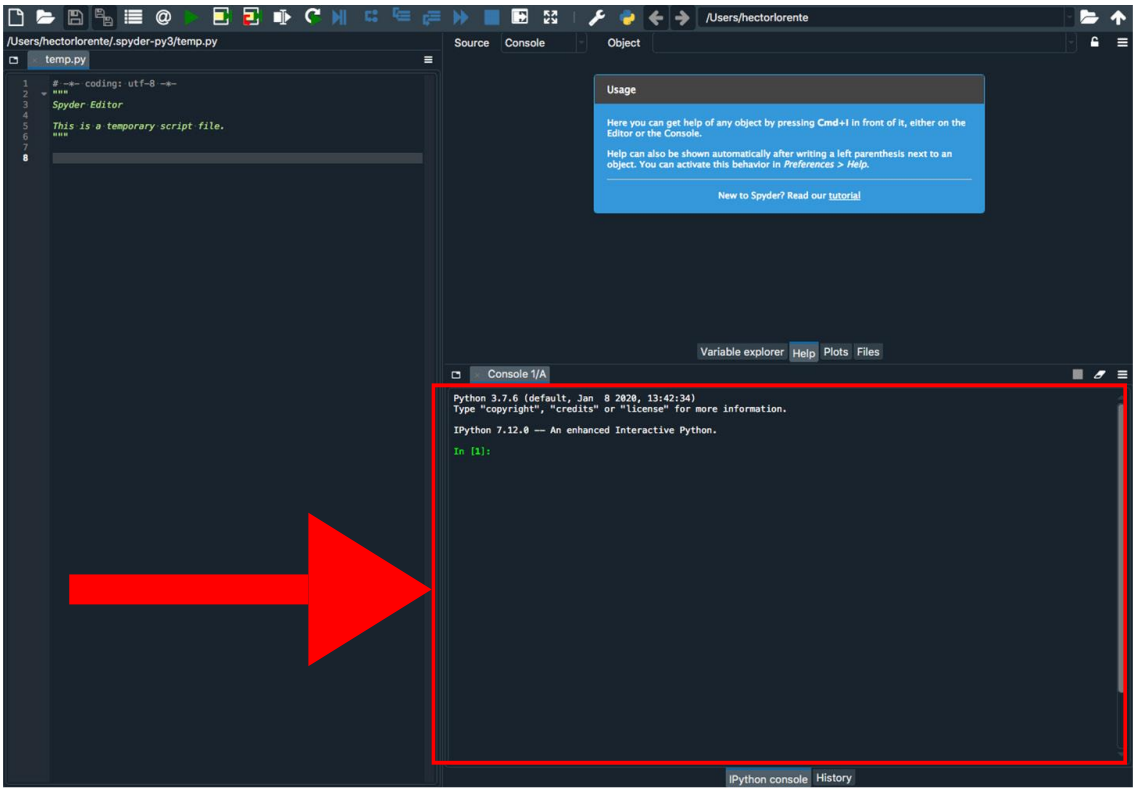


Figure 3. Spyder program interface. The red arrow indicates the Python console.

B. Through NCBI Website

- Go to https://www.ncbi.nlm.nih.gov/assembly/?term=, 27 January 2022.
- Type your query (usually species or higher taxon names) on the search box located at the top of the web page (Fig. 4).
- By clicking the *Advance* button right below the search box (Fig. 4), you can manually add field tags (e.g., organism, assembly level, etc.).
- Filters *Latest* and *Exclude anomalous* are applied by default (Fig. 4).
- After your search, you would find a side bar on the left with all the available filters (Fig. 4).
- In addition, you can find a text box named "Search details" on the right with the specific command of your query (Fig. 4).
- This text box can be useful for creating custom queries that can be used in the automatic path (See Section 3.1.1. A.).
- Note that when there are more than one field tag or filter, they appear enclosed in parentheses (Fig. 4).
- We recommend keeping the default filters and adding *Representative or Reference* from *RefSeq category* by clicking on the left side panel (Fig. 4) or by

typing *AND representative genome [filter] OR reference genome [filter]* within the filter parentheses in the search box.

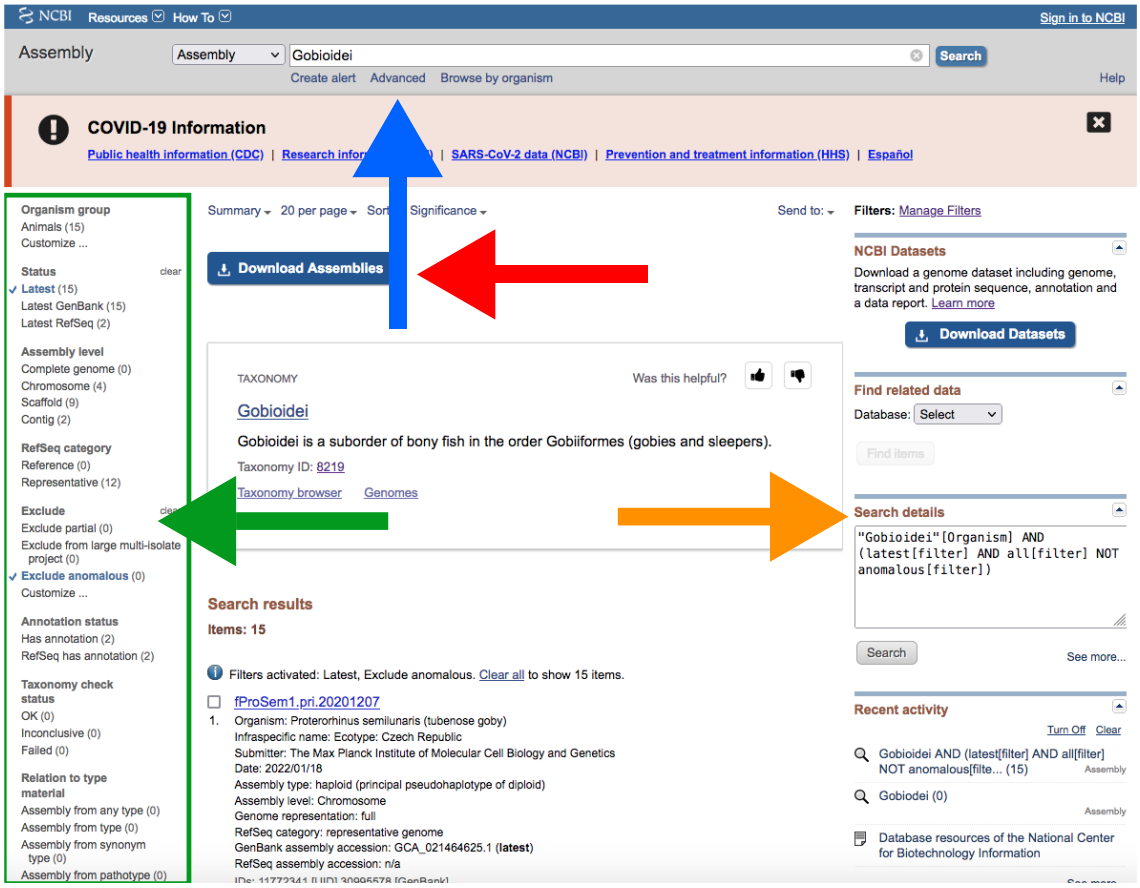


Figure 4. NCBI web server interface after conducting a search on assembly database. The blue arrow indicates the search box; the green arrow and green square indicate *Representative* and the filters side panel, respectively; the orange arrow indicates "Search details"; the red arrow indicates the *Download Assemblies* button (see Fig. 5 for more information about downloading). Accessed on 27 January 2022.

- Click on the *Download Assemblies* button to download the assembly (Fig. 5).
- As mentioned above, genome assembly may not be evenly distributed. This can be a problem when downloading.
- We recommend that you to choose *RefSeq* under "Source database" (Fig. 5) and either *Protein FASTA (.faa)* or *RNA FASTA (.fna)* under "File type".
- After decompressing the downloaded file, you will find a directory tree similar to *genome_assemblies_genome_fasta/ncbi-genomes-date*. Within this, you will find all the files for every species.

- Then, repeat the search, but this time, when downloading, choose *GenBank* under "Source database" (Fig. 5) and *Genomic FASTA (.faa)* under "File type". Right after that, check what species were previously downloaded, and then check just the boxes of the remaining species.
- Of course, these steps are not intended for downloading a large number of genomes; for such purposes, we recommend following Section 3.1.1. A.

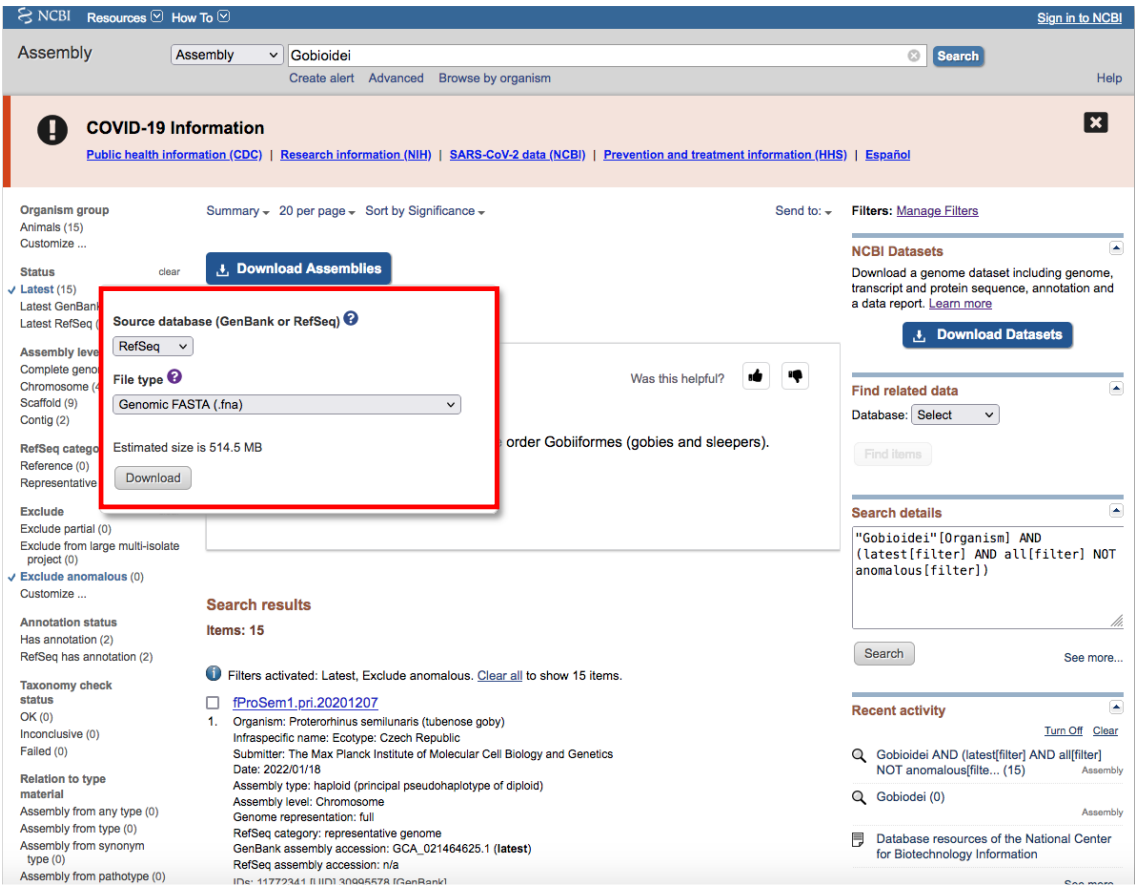


Figure 5. *Download Assemblies* button in detail. Accessed on 27 January 2022.

- If there is just one assembly for the selected taxa, you do not need to apply any filter. The web will automatically redirect you to the assembly entry page (Fig. 6).
- We recommend storing every ".gz" file in a separate folder according to its data type (*Genomic*, *RNA*, *Protein*) first and then according to the species/organism.
- Download the file by clicking on *Download Assemblies* (Fig. 6) as detailed above.

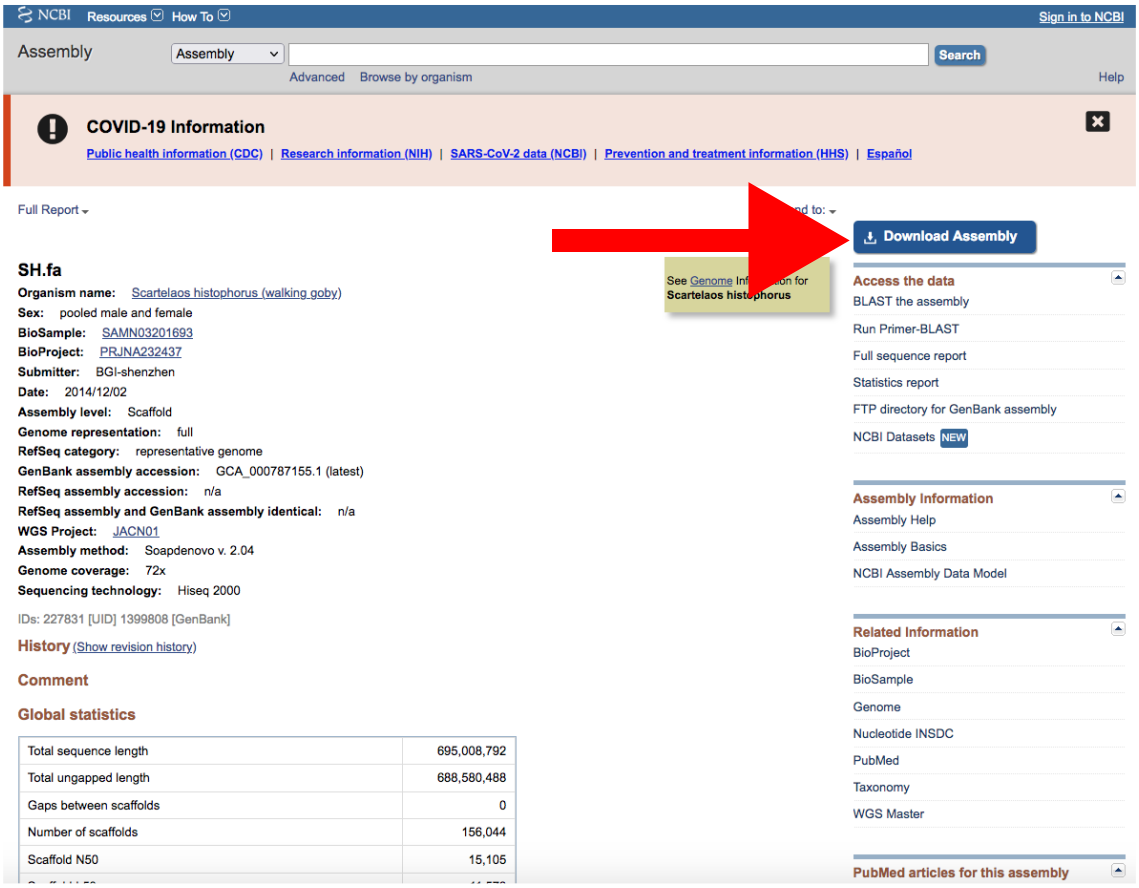


Figure 6. Example of an assembly entry. The NCBI web server will automatically redirect to a page like this if there is only one assembly that matches your query parameters. The red arrow indicates the *Download Assembly* button. Accessed on 27 January 2022.

3.1.2. Generation of Query Sequences

A. Locally Using Python

- Type *run get_query_sequences.py [email address] [path to query file] --curated --refine*. This will download a protein dataset containing 200 sequences that includes the name in "Protein Feature" (curated parameter) from "RefSeq" (refine parameter).
- The "query.txt" file can include several queries. The programs expect a gene name, with fields and filters enclosed in parentheses right after it.
- You can type your own field tags and filters, typing them after *--refine* argument in a similar way as when downloading genomes (See Section 3.1.1.)
- In addition, you can restrict the number of downloaded sequences to a maximum number using *--retmax* arguments. When using *--curated* arguments, the program should curate sequences based upon this number; therefore, you may obtain a smaller number of sequences than with *--retmax*.

- There is not a perfect number of query sequences. Ideally, the best number should maximize the diversity of the studied gene family and minimize computing time. Our approach (200) aims for a great coverage of this diversity.
- BLAST can perform well with just a few sequences (around 10), reducing computing time. Therefore, another strategy could be to manually select some key sequences and download them one at a time. Of course, this requires a solid knowledge of the studied gene family.
- All this is for protein downloading, recommend as query when using BLAST searches. However, download of nucleotide sequences is also allowed (if needed for alignment; see below). However, this is not refined, so we recommend using *biomol_mrna[PROP]* to download just the transcripts.
- The database will be stored at *Data/Query_seqs*, and its name will be the name of the gene you entered, followed by "query_sequences_data_type.fas".

B. Through NCBI Website

- Go to https://www.ncbi.nlm.nih.gov/protein/?term=, 27 January 2022, (Fig. 7).
- Type the name of your protein in the search box at the top (Fig. 7).
- In a similar way as explained for downloading genomes (see above), you can manually add field tags by clicking the *Advance* button right below the search box (Fig. 7). You can also add filters to refine your search.
- We recommend that you use the default *refseq[filter]* (Fig. 7). In the "Search box" on the right, you can look at the command that you are applying to your search (Fig. 7).
- To download the database, click on *Send to* → *File* → *Forma*t → *FASTA* → *Create File* (Fig. 7).
- Note that this will download the whole list of results. Therefore, it is important to be as precise as possible with your query.
- You can download nucleotide sequences in a similar way at https://www.ncbi.nlm.nih.gov/nuccocre/?term=, 27 January 2022. Type *biomol_mrna[PROP]* after your query to download transcripts.

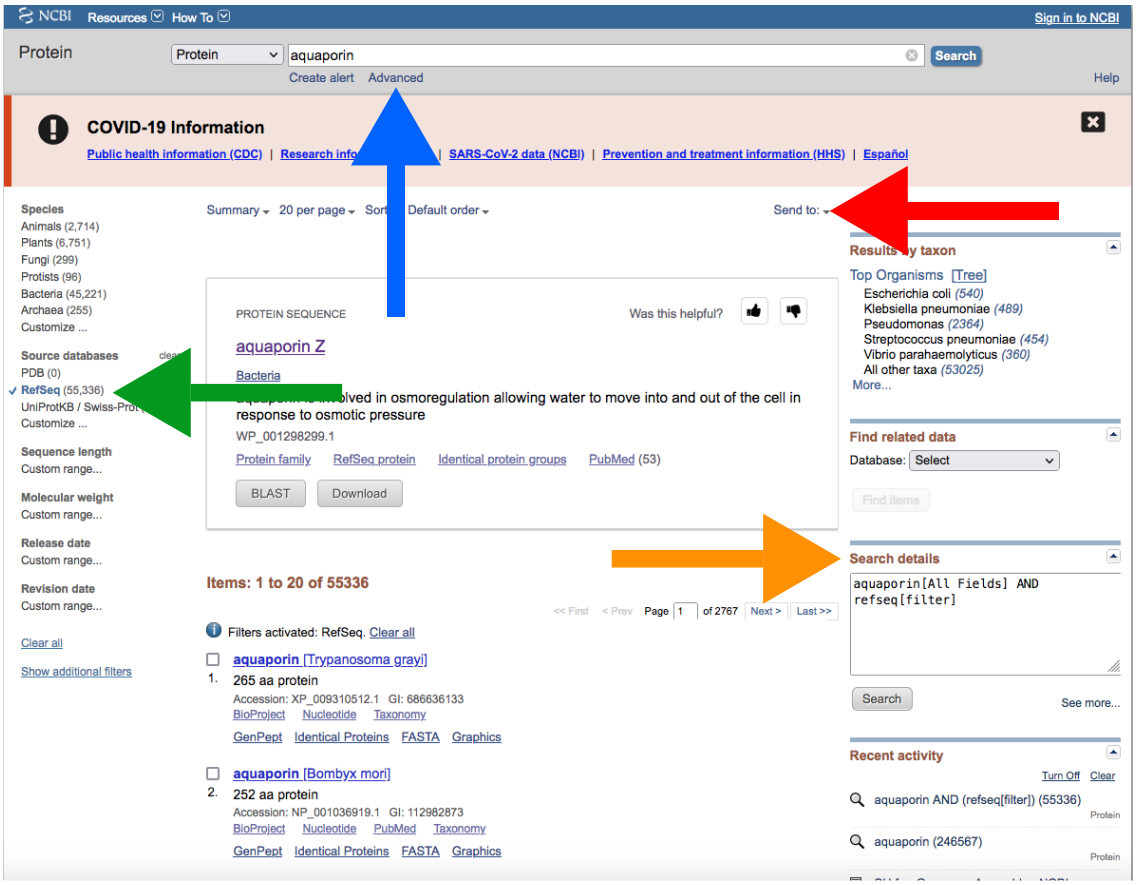


Figure 7. The NCBI web server interface after search in the protein database. The blue arrow indicates the search box; the green arrow and green square indicate "RefSeq" and the filters side bar, respectively; the orange arrow indicates "Search details", and the red arrow button and red square indicate the "Send to" button and the "Create file" window for downloading, respectively. Accessed on 27 January 2022.

3.1.3. BLAST Searches

A. Locally Using Python

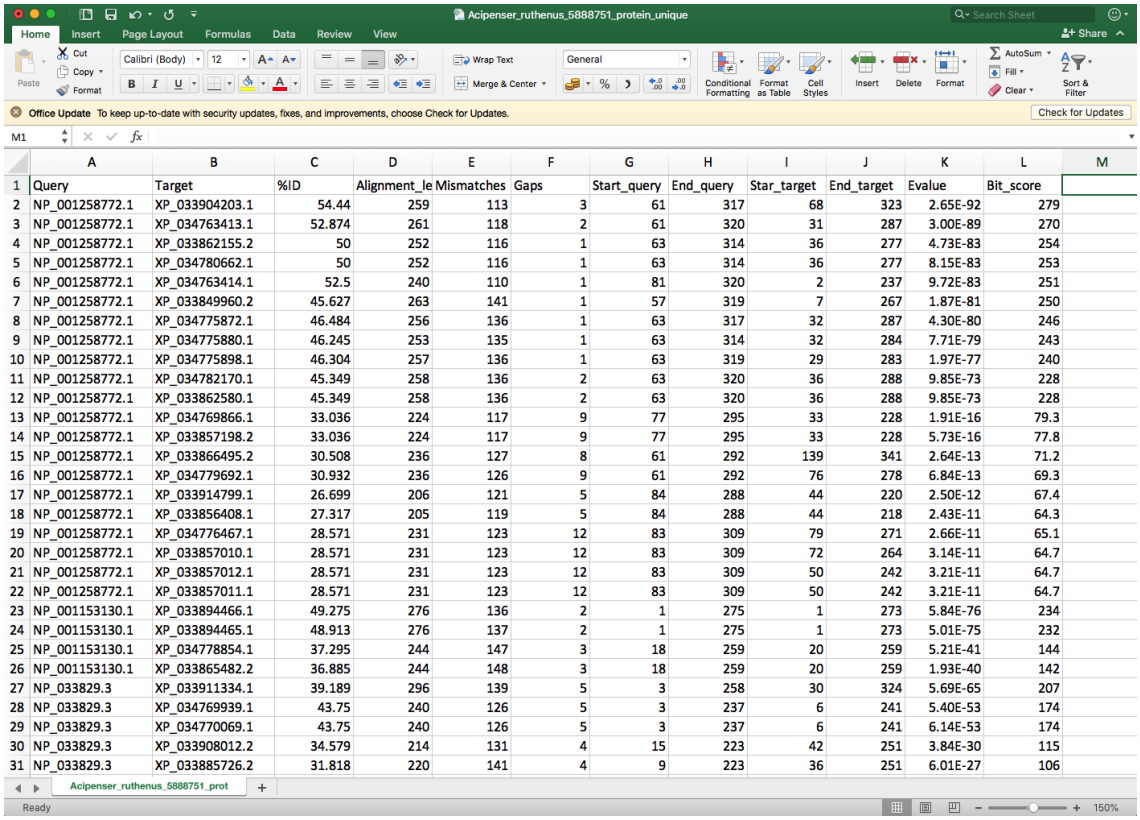
- Type *run decompress_genomes.py* in the Spyder Python console to decompress the genomic data files. Use *genomic*, *RNA*, or *protein* arguments or *directory* for your own custom path.
- Go to https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/ (27 January 2022) and download ncbi-blast-version-x64-win64.tar.gz (Windows users) or ncbi-blast-version-x64-linux.tar.gz (Linux and Mac users).
- Decompress the BLAST program.
- Run *blast.py [blast_directory] [data_type]*.
- BLAST programs are located in the *bin* folder inside the BLAST directory. Drag the *bin* folder to the Spyder console after *--blast_path* parameter when running.

- The program will check in *Genomic*, *RNA*, and *Protein* directories automatically. You can change the directory by using --*directory* arguments, but you must also specify the genomic data type.
- You can modify the e-value parameter (see https://www.ncbi.nlm.nih.gov/books/NBK279690/, 27 January 2022, and https://www.ncbi.nlm.nih.gov/books/NBK279684/table/appendices.T.options_common_to_all_blast/ for more information, 27 January 2022). You will obtain a ".tsv" file with all the hits found in your target genomic data.

3.1.4. Extraction of RAW Sequences

A. Locally Using Python

- Type *run get_unique_hits.py* to obtain the best hit for every of the entries of your target data.
- For whole-genome data, go to Section 3.1.4. Multiple Gene Inspection below and then continue with the next step.
- Type *run get_RAW_sequences.py [data_type]* to extract every sequence corresponding to each unique hit. The extracted sequences will be stored in the *Extraction* directory located in the same folder as the whole genome file.
- Change directory using --*directory* argument but keep using genomic data type.
- If using whole genome sequences, you may have to modify the --*in_len* parameter to control the intron length.
- Using --*query_seqs* arguments and typing your database file path allows you to attach some of the database sequences that match your query entry.
- By default, the program will attach the five (arbitrary number) first non-redundant sequences according to the entries of the BLAST output file. Change this using --*query_seqs_num* parameter.
- Note that for whole genome entries that include more than one gene, this number depends on the number of modified query entry IDs (see above).
- You can manually add any sequence. Preferably, add sequences from closely related species. You can download single sequences in the same way as the query dataset (see Section 3.1.2).



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Query	Target	%ID	Alignment_le	Mismatches	Gaps	Start_query	End_query	Star_target	End_target	Evalue	Bit_score	
2	NP_001258772.1	XP_033904203.1	54.44	259	113	3	61	317	68	323	2.65E-92	279	
3	NP_001258772.1	XP_034763413.1	52.874	261	118	2	61	320	31	287	3.00E-89	270	
4	NP_001258772.1	XP_033862155.2	50	252	116	1	63	314	36	277	4.73E-83	254	
5	NP_001258772.1	XP_034780662.1	50	252	116	1	63	314	36	277	8.15E-83	253	
6	NP_001258772.1	XP_034763414.1	52.5	240	110	1	81	320	2	237	9.72E-83	251	
7	NP_001258772.1	XP_033849960.2	45.627	263	141	1	57	319	7	267	1.87E-81	250	
8	NP_001258772.1	XP_034775872.1	46.484	256	136	1	63	317	32	287	4.30E-80	246	
9	NP_001258772.1	XP_034775880.1	46.245	253	135	1	63	314	32	284	7.71E-79	243	
10	NP_001258772.1	XP_034775898.1	46.304	257	136	1	63	319	29	283	1.97E-77	240	
11	NP_001258772.1	XP_034782170.1	45.349	258	136	2	63	320	36	288	9.85E-73	228	
12	NP_001258772.1	XP_033862580.1	45.349	258	136	2	63	320	36	288	9.85E-73	228	
13	NP_001258772.1	XP_034769866.1	33.036	224	117	9	77	295	33	228	1.91E-16	79.3	
14	NP_001258772.1	XP_033857198.2	33.036	224	117	9	77	295	33	228	5.73E-16	77.8	
15	NP_001258772.1	XP_033866495.2	30.508	236	127	8	61	292	139	341	2.64E-13	71.2	
16	NP_001258772.1	XP_034779692.1	30.932	236	126	9	61	292	76	278	6.84E-13	69.3	
17	NP_001258772.1	XP_033914799.1	26.699	206	121	5	84	288	44	220	2.50E-12	67.4	
18	NP_001258772.1	XP_033856408.1	27.317	205	119	5	84	288	44	218	2.43E-11	64.3	
19	NP_001258772.1	XP_034776467.1	28.571	231	123	12	83	309	79	271	2.66E-11	65.1	
20	NP_001258772.1	XP_033857010.1	28.571	231	123	12	83	309	72	264	3.14E-11	64.7	
21	NP_001258772.1	XP_033857012.1	28.571	231	123	12	83	309	50	242	3.21E-11	64.7	
22	NP_001258772.1	XP_033857011.1	28.571	231	123	12	83	309	50	242	3.21E-11	64.7	
23	NP_001153130.1	XP_033894466.1	49.275	276	136	2	1	275	1	273	5.84E-76	234	
24	NP_001153130.1	XP_033894465.1	48.913	276	137	2	1	275	1	273	5.01E-75	232	
25	NP_001153130.1	XP_034778854.1	37.295	244	147	3	18	259	20	259	5.21E-41	144	
26	NP_001153130.1	XP_033865482.2	36.885	244	148	3	18	259	20	259	1.93E-40	142	
27	NP_033829.3	XP_033911334.1	39.189	296	139	5	3	258	30	324	5.69E-65	207	
28	NP_033829.3	XP_034769939.1	43.75	240	126	5	3	237	6	241	5.40E-53	174	
29	NP_033829.3	XP_034770069.1	43.75	240	126	5	3	237	6	241	6.14E-53	174	
30	NP_033829.3	XP_033908012.2	34.579	214	131	4	15	223	42	251	3.84E-30	115	
31	NP_033829.3	XP_033885726.2	31.818	220	141	4	9	223	36	251	6.01E-27	106	

Figure 8. Example of a BLAST output file viewed in Excel.

B. Locally Using a Text Editor

- For whole-genome data go to Section 3.1.4. Multiple Gene Inspection below and then continue with the next step.
- Open every "whole_genome_name_out.tsv" file. Look at the second column (target ID) and keep just unique IDs. For whole-genome data, follow Section 3.1.4. A.
- Open your genomic data (i.e., *Genus_species_id.faa*) with a text editor. This step cannot be performed in some cases, especially those that imply the use of non-annotated genome sequences.
- Pick up every unique target ID and search for it in the corresponding genomic data file.

  Multiple Gene Inspection (Mandatory for Whole-Genome Data; Skip if Using RNA or Protein Data).
- The following steps assume that you have used *--outfmt* 6 (i.e., BLAST tabular output format 6).
- Open the "whole_genome_name_out.tsv*"* file with a spreadsheet program (such as Microsoft Excel) or a text editor.

- Every row corresponds to a different hit, and the second column indicates the target identifier (scaffold ID) (Fig. 8).
- Controlling by the 2nd column, you must check the 9th and 10th columns that contain the start and end positions where the query sequence aligned within the target entry and compare entries below to identify different start or end positions that could be associated with two different genes.
- In order to facilitate visibility, you can highlight every target by clicking on *Conditional Formatting → Highlight Cells Rules → Equal to → Choose the corresponding cell → OK*.
- You can also highlight the 9th and 10th columns in the same way using *Between to* instead of *Equal to* for controlling.
- If you spot another gene in the same query entry, you must modify the query entry ID (first column) by adding "_1" to the rows that belong to the first one, "_2" to the second one, etc. We recommend changing at least five query entry IDs if possible in order to facilitate proper gene fishing (see below). Additionally, you must update the "_unique.tsv" file with the new query names, and you must add at least one row containing the information of the new gene(s).
- Before continuing with the next target ID, click on *Conditional Formatting → Clear Rules → Clear Rules for Entire set*.

3.1.5. Coding Sequence Identification (Although This Step Is Mandatory When Working with Whole Genome Sequences, You Can Skip It When Working with Protein and RNA Sequences)

- You should have used --query_seqs earlier (Section 3.1.4 A.) to attach template sequences or have manually added some.

  Alignment

A. Locally Using Python

- Go to MAFFT software web (https://mafft.cbrc.jp/alignment/software/, 27 January 2022) and navigate to the specific page according to your operating system. Follow the steps to install MAFFT software on your computer.
- On the Spyder Python console, type *run align_sequences.py*. You can choose a specific alignment algorithm using *--algorithm*. The MAFFT manual

recommends using the *Auto* function when you know little about your data. For genomic data and working with one gene family, we recommend using the *E-INS-i* algorithm.

- As in other cases, you can use *genomic*, *rna*, and *protein* or *directory* arguments.
- Windows users may encounter some problems in either installing or running MAFFT, especially those using older system versions. If this is the case, look at the next section.

B. Through MAFFT Server

- Go to the MAFFT online version page (https://mafft.cbrc.jp/alignment/server/, 27 January 2022).
- Paste the content of the file you want to align in the available text box or browse for your file by clicking on the *Choose File* button.
- Select *Same as input* for the options: *UPPERCASE/lowercase*, *Direction of nucleotide sequences*, and *Output order*.
- Scroll down to the *Advance settings* section. In the *Strategy* section, we recommend using *Auto* if you know little about your data. For genomic data, we recommend using the *E-INS-i* algorithm.
- Download the alignment from the results page (Fig. 9). This page will pop up after the alignment run is completed.
- At the top of the page, click the right button of the mouse over *Fasta format* → *Save link as* → *Save it adding "_final.fas" suffix* (Fig. 9).

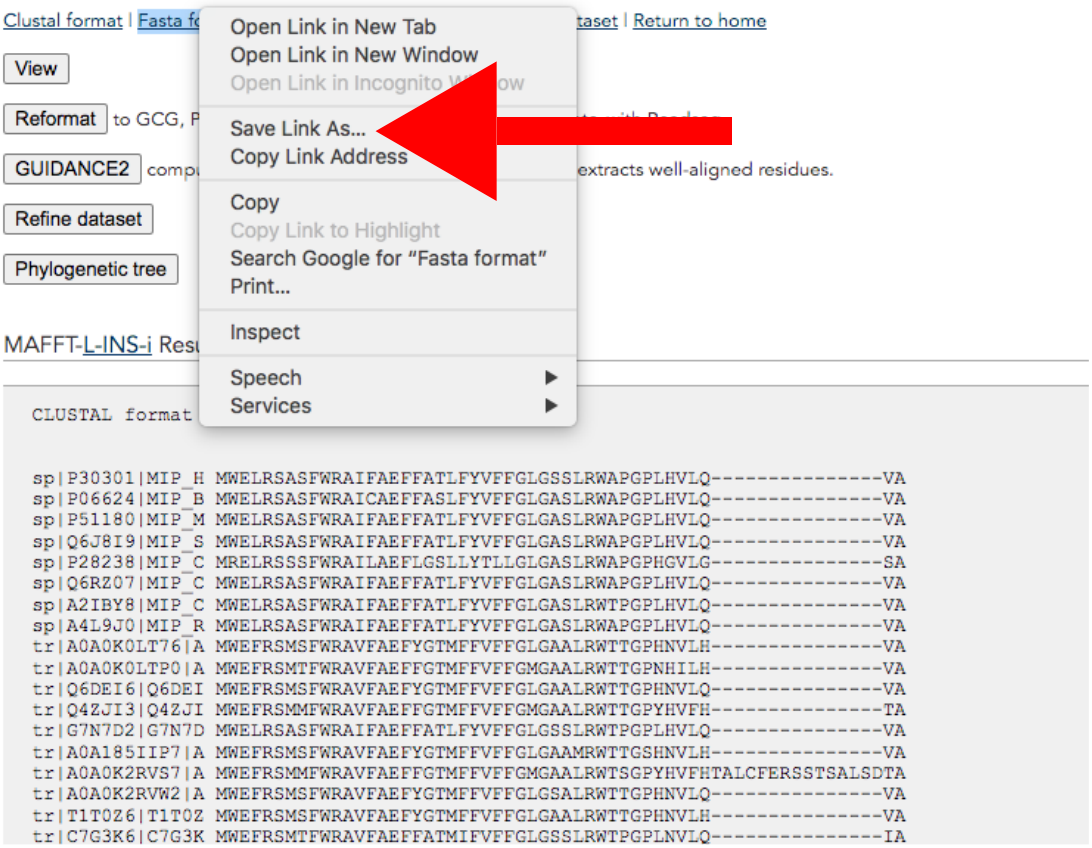


Figure 9. MAFFT result web page. The red arrow indicates *Save link As…* button for downloading. Accessed on 27 January 2022.

Trimming and Retrieving Coding Sequences (Using MEGA Version 11)

- Go to the MEGA home page (https://www.megasoftware.net/, 27 January 2022), select your operating system, *Graphical GUI*, and *MEGA 11* (or newer versions if available), and click on the *Download* button.
- Follow the steps for MEGA downloading and installation.
- Open the MEGA program and load every alignment file (*ALIGN → Edit/Build Alignment → Open a saved alignment version → OK → Open the downloaded file*) (Fig. 10).
- Trim the sequences using the *Scissors* tool (Fig. 10) or using *Ctrl* or *Cmmd + X*. If you are using genomic or RNA data, you can click on *Translated Protein Sequences* to obtain the deduced amino acid sequences (see also Section 3.2.1.), which can be useful for delimiting open reading frames (identification of start/stop codons and intronic/exonic regions by visual inspection).
- Finally, click on *Data → Export Alignment → Fasta Format → Save it.*

- When working with genomic sequences, it may be necessary to conduct steps in Section 3.1.5. several times (iterative refinement) in order to eventually obtain the coding sequence of interest. The alignment becomes progressively refined by iteratively trimming intronic regions and leftover positions at the beginning and the end of the sequence.
- In some cases, you will have to rerun *get_RAW_sequences.py*, changing the *--in_len* parameter value in order to cover all the sequence. Sometimes the chosen value may be too small, and part of the sequence can be left out unintentionally.
- Once the coding sequence has been fully identified, save the alignment with MEGA as detailed above and add the suffix, "_final.fas".

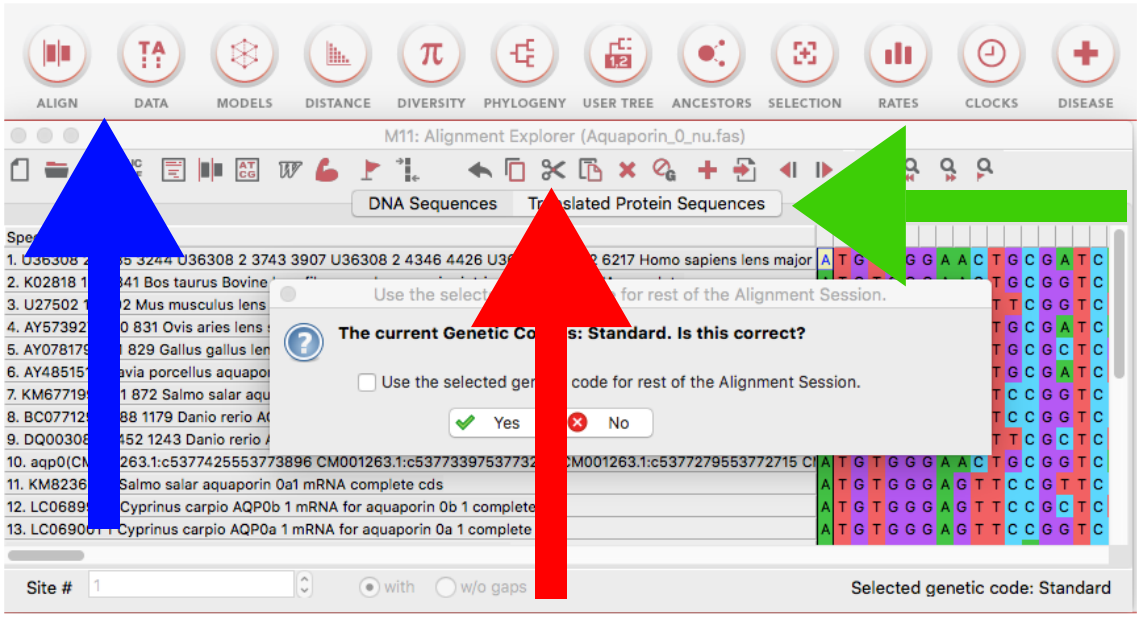


Figure 10. MEGA version 11 program showing a nucleotide alignment window. The blue arrow indicates the tool bar where the *ALIGN* and *DATA* buttons are located. The red arrow indicates the scissors tool. The green arrow indicates the *Translated sequences* tab.

3.2. Phylogenetic Analyses

3.2.1. Translate Sequences

A. Locally Using Python

- This step is mandatory if you want to obtain a protein data matrix and you are working with genomic or RNA sequences.
- In the Spyder console, type *run translate_sequences.py [data_type]*. Typically, you are going to use it with genomic sequences or, in some cases, with RNA sequences (particularly for checking and removing ambiguously aligned positions).
- By default, the program will look for files with "final.fas" or "RAW.fas". Change this using the --*pattern* parameter.
- If you are working with several species that each have a different genetic code, you will have to run this program several times. We recommend cutting those folders that share the same genetic code and pasting them into a new folder. Use --*directory* arguments to indicate the new path and run the program. Then, put the folders back in their initial location and repeat the step with the different genetic codes.

  B. Locally Using MEGA
- Open the MEGA program (see Section 3.1.5. Trimming and Retrieving Coding Sequences for MEGA installation).
- Import alignment as in Section 3.1.5. Trimming and Retrieving Coding Sequences (Fig. 10).
- On the emergent window, click on *Translated Protein Sequences → Choose the adequate genetic code* (Fig. 10).
- Finally, click on *Data → Export Alignment → Fasta Format → Save with "_translated.fas" suffix.*
- Repeat with all the files that need to be translated.

3.2.2. Matrix Assembly

A. Locally Using Python

- In the Spyder Python console, type *run get_combined_seqs.py [output name][data_type]*. The three data types can be combined. However, note that if

you have information for more than one data type for the same species, then you may obtain redundant sequences.

- By default, the pattern for file searching is "final.fas". However, it is programmed to look for "final_translated.fas", "RAW.fas", and "RAW_translated.fas" when it cannot find the first pattern.
- You can change this using the *--pattern* parameter. Then, the program will search for "new_pattern.fas", "new_pattern_translated.fas", "RAW.fas", and "RAW_translated.fas".
- The program will produce a data matrix named "[output name]_all_combined.fas*"*, which will be downloaded to the *Data* folder.

  B. Locally Using MEGA
- Import a sequence file as in Section 3.1.5. Alignment B. (Fig. 10).
- In the emergent window, click on *Edit → Insert Sequence From File → Open every sequence file* (Fig. 10).
- This step can also be conducted manually with a text editor; simply open every file containing the downloaded sequences and paste their content into a new file, one after the other.

3.2.3. Alignment

- See Section 3.1.5. Alignment. If using the Python version, use *–directory*, and drag the combined matrix file.

  Alignment Filtering

  A. Using trimAl Locally through Python
- Go to http://trimal.cgenomics.org/downloads (27 January 2022) and download the specific program according to your operating system.
- Decompress the trimAl file. For Windows and Mac/Unix users, open the terminal and follow the steps on the trimAl README.md file. You can run the program on the terminal following the instructions at http://trimal.cgenomics.org/use_of_the_command_line_trimal_v1.2 (27 January 2022) or the following steps.
- Type *run alignment_trimming.py [path_to trimAl] [path_to_matrix] [combined_matrix]* in the Spyder console. For Windows users, trimAl will be

stored in the *bin* directory. For Mac and Unix users, trimAl will be stored in the *source* directory.

- This will remove all positions in the alignment with gaps in 90% or more of the sequences (-gt 0.9), which is the default option of the program.
- The trimming algorithm can be changed using --trimal_command. See http://trimal.cgenomics.org/use_of_the_command_line_trimal_v1.2 (27 January 2022) for additional information.

B. Using trimAl through the Phylemon Web Server

- Go to http://phylemon2.bioinfo.cipf.es/, 27 January 2022.
- Create and account and login or star as anonymous user.
- Go to *Utilities → Alignment Utilities → TrimAl* (Fig. 11).
- Paste the content of the matrix or upload the file clicking on *browse server → Upload new file → Browse → Open the matrix → Select format → Aligned sequences → Upload → Accept.*
- Click on *Method → User define →* In *Gap threshold, fraction of positions without gaps in a column* set 0.1. Similar output as using Python version.
- Run the job.
- Refresh the page. On the right panel, open the job when finished.

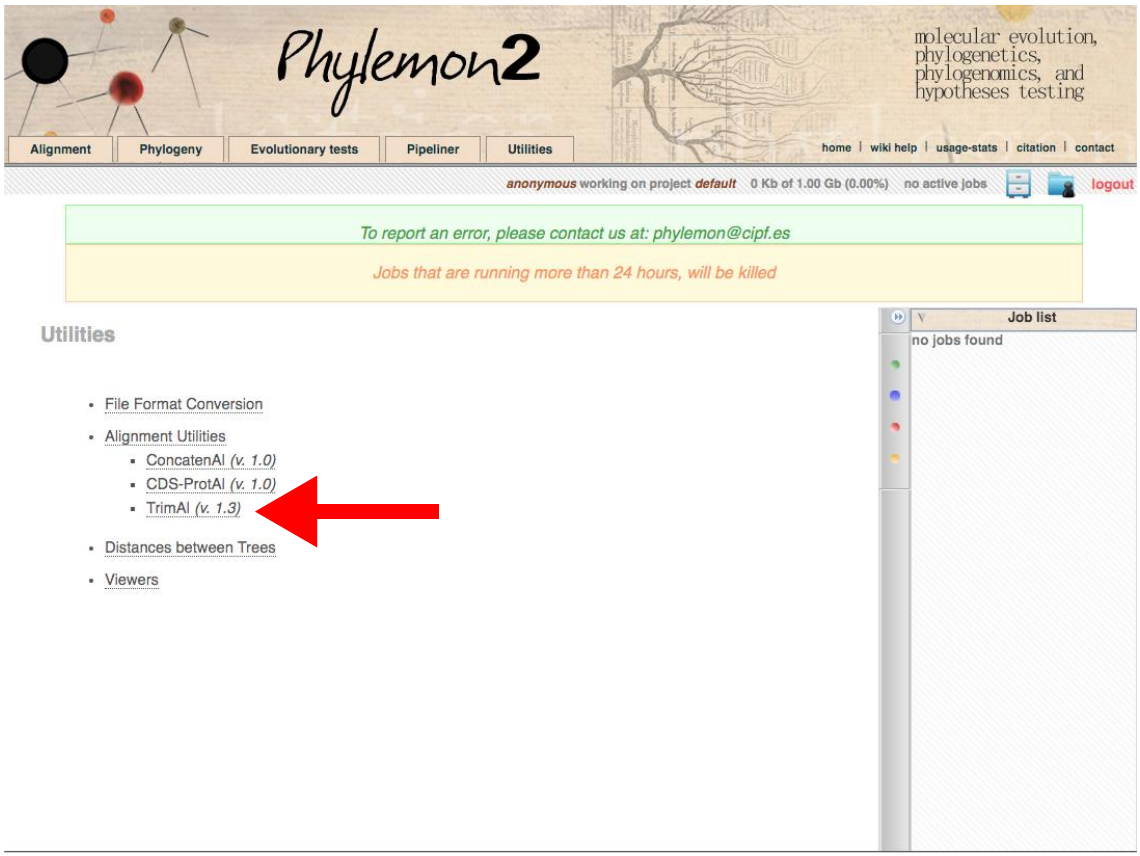


Figure 11. Phylemon2 *Utilities* web server. The red arrow indicates the trimAl program. Accessed on 27 January 2022.

3.2.4. Phylogeny Inference

A. Using IQ-TREE Locally through Python

- Go to http://www.iqtree.org/#download (27 January 2022) and download the adequate version for your operating system.
- Decompress the folder.
- Type *run phylogenetic_inference.py [iqtree_folder] [trimmed_matrix] [data_type]*.
- IQ-TREE program is located in the *bin* folder in the IQ-TREE program folder. Drag this folder to the Spyder console when running.
- The ".treefile*"* will be stored at *Data/Phylogenetic_inference*.

B. Using IQ-TREE Web Server

- Go to http://iqtree.cibiv.univie.ac.at/, 27 January 2022.
- Browse the trimmed matrix in the *Alignment file* field.
- Select sequence type.

- Do not change any more parameters for a similar run and output as the Python version.
- Enter your e-mail and click on *SUBMIT JOB.*
- When finished, you will receive and e-mail. Click on the provided link, and with the button on the left, click *DOWNLOAD SELECTED JOBS.*

Tree Visualization

- Go to the FigTree website (https://github.com/rambaut/figtree/releases (27 January 2022) and download the specific version for your operating system.
- Decompress the file and open FigTree (Fig. 12).
- Click on *File → Open → Open ".treefile" file.*
- Provide a label for the values of support (or leave unchanged).
- The left panel allows for modification of multiple tree features displayed as collapsible menus. For example, tree appearance options can be changed from the *Appearance* menu.
- Display the values of support. Click on *Node Labels → Display → Select the name of the label provided before* (Fig. 12).
- The root of the tree can be changed by selecting a specific branch and then clicking the *Reroot* button at the top of the window.

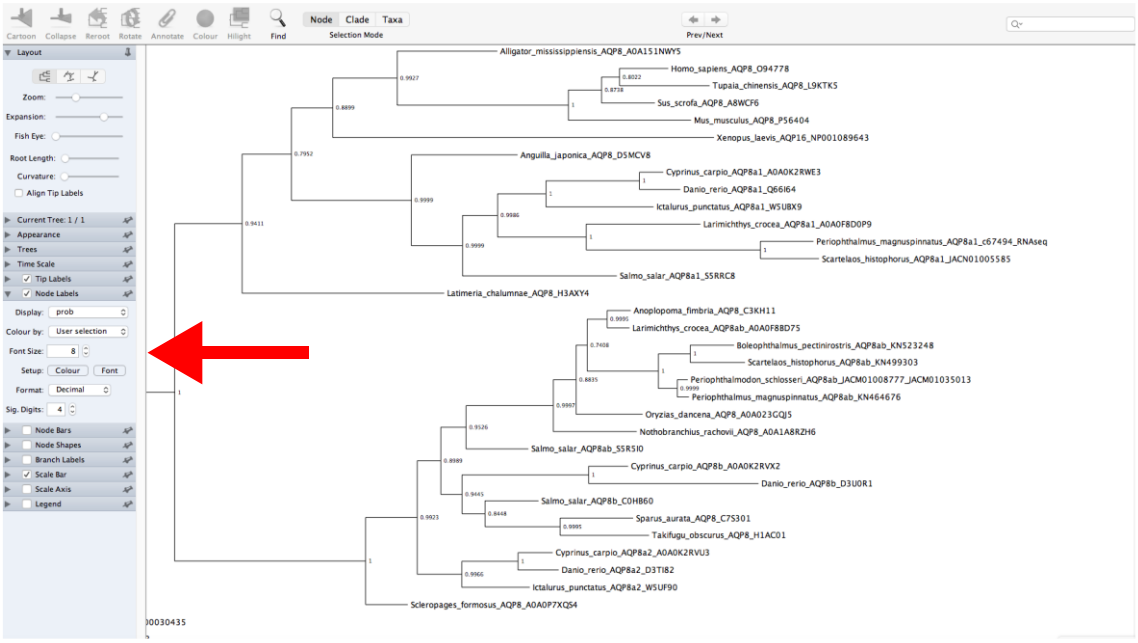


Figure 12. FigTree program. The red arrow indicates the *Node Label* panel.

4. Expected Results

The detailed protocol will primarily yield a sequence alignment and a phylogenetic tree (along with a best-fit model of substitution) for a particular group of organisms of interest. Altogether, these resulting data are the typical starting point of molecular evolution analyses, such as those aimed at detection of adaptive and/or purifying selection, ancestral protein reconstruction, or protein structure. Additionally, the protocol can provide refined annotation of raw genome sequences, which can be used for other goals, such as gene mapping or functional analyses. Finally, the provided pipeline and protocol constitute a basis for bioinformatic work with DNA and protein sequences that can be easily modified and adapted for many other specific tasks in the present context of massive generation of genomic data.

5. Troubleshooting

The first issue that readers may encounter relates to software versions and maintenance of websites and online resources. Even if some of these are discontinued, the methods outlined in the workflow would stand as a solid guideline. Every day, new applications and versions are released, and it is not difficult to find programs that perform similar tasks to the ones mentioned in the protocol. It is also possible that the functions of some of our Python applications may become deprecated in the future. This may represent a more challenging issue because the user might have to modify the source code (therefore requiring at least some basic notions of programming languages). The user should also be aware that unexpected inputs and commands can result in illogical outcomes. It is therefore of critical importance to follow the instructions provided here or in the README.md file and to carefully inspect the output files generated in each step.

Additionally, there may be internet connection issues when downloading a large number of genomes (especially genomic sequences) using GNFish that may produce partly empty or incomplete files (apparently correct sometimes). This will prevent further progression in the workflow, as subsequent steps become logically impossible even if the inputted commands are valid. A similar issue can occur when downloading the query sequences, although this is less common.

Generally speaking, it is always a challenge to work with genomic sequences because their analysis usually requires substantial investment of time and use of computational resources. If possible, we recommend running the GNFish package on an HPC (high-performance computing) cluster, particularly when working with scaffold sequences. As previously stated, these data must be iteratively curated, and identifying open reading frames can be arduous or even impractical in some difficult cases. In all cases, outcome sequences must be carefully examined by the user. The more information about the gene is available, the easier and more accurate the retrieval becomes. In some particular cases, the user would benefit from repeating the retrieval process if additional knowledge of the data is gained.

Another issue may be related to phylogenetic inference if the maximum likelihood analysis becomes trapped in local optima (particularly an issue for very large datasets). If this is suspected, the IQ-TREE developers recommend repeating the analysis with at least 10 independent runs.

Last but not least, high-throughput sequencing technologies are not flawless, and sequencing errors may occur in the source data. We recommend using the RefSeq database throughout the protocol because it only contains curated sequence data. We also suggest the use of some filters when searching for sequences, which may improve the quality of the outcome. Nonetheless, the user should be aware that full exclusion of low-quality sequences may not always be possible, and dealing with a certain proportion of them in the datasets is likely unavoidable. In principle, this proportion should be low, with little or a negligible impact on subsequent analyses.

Author Contributions

Conceptualization, H.L.-M., A.A., and D.S.M.; methodology, H.L.-M. and D.S.M.; software, H.L.-M.; validation, H.L.-M., A.A., and D.S.M.; formal analysis, H.L.-M.; investigation, H.L.-M. and A.A.; resources, H.L.-M.; data curation, H.L.-M.; writing—original draft preparation, H.L.-M.; writing—review and editing, A.A. and D.S.M.; visualization, H.L.-M. and A.A.; supervision, A.A. and D.S.M.; project administration, A.A. and D.S.M.; funding acquisition, D.S.M. All authors have read and agreed to the published version of the manuscript.

Funding

This work received financial support from the Ministry of Science and Innovation of Spain (grant PID2020-115481GB-I00 to D.S.M.). H.L.-M. was sponsored by a predoctoral contract of the Complutense University of Madrid in partnership with the Real Colegio Complutense at Harvard University (RCC-UCM).

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The GNFish package containing code scripts, readme file, and examples is available from the GitHub platform at https://github.com/hectorloma/GNFish, 27 January 2022.

Acknowledgments

Three anonymous reviewers provided insightful comments on an earlier version of the manuscript. Support was provided by the Spanish network of research laboratories working on adaptation genomics (AdaptNET).

Conflicts of Interest

The authors declare no conflict of interest.

CHAPTER II

Early diversification of major intrinsic proteins (MIP) in eukaryotes

---

Diversificación inicial de las proteínas intrínsecas principales (MIP) en eucariotas

This chapter is a reproduction of the following manuscript, currently under review in *Proceedings of The Royal Society B, Biological Sciences*:

Irisarri, I.*, Lorente-Martínez, H.*, Strassert, J.F.H., Agorreta, A., Zardoya, R., San Mauro, D., & de Vries, J. Early diversification of major intrinsic proteins (MIP) in eukaryotes. (*contributed equally)

Abstract

Aquaporins (AQPs) or major intrinsic proteins (MIPs) form an ancient family of transporters for water and small solute across biological membranes. The evolutionary history and functions of MIPs have been extensively studied in vertebrates and plants, but their widespread presence across the eukaryotic tree of life suggests both a more complex evolutionary history and broader set of functions than previously thought. Yet, the early evolution of MIPs remains obscure. The co-occurrence of up to two MIPs in most prokaryotes suggests that the first eukaryotes likely had one aquaporin (AQP) and one aquaglyceroporin (GLP). Here, we report on a previously unknown richness in eukaryotic MIP diversity across all major lineages, including protists, which make up the bulk of eukaryotic diversity. Three MIP clades have likely deep evolutionary origins, dating back to the last eukaryotic common ancestor (LECA), and supporting the presence of a complex MIP repertoire in early eukaryotes. Overall, our findings highlight the growing complexity of the reconstructed LECA genome: the dynamic evolutionary history of MIPs was set in motion when eukaryotes were in their infancy followed by radiative bursts across all main eukaryotic lineages.

Keywords: aquaporin; deep eukaryote evolution; aquaglyceroporin; last
43 eukaryote common ancestor; water transport.

Resumen

Las acuaporinas (AQPs) o proteínas intrínsecas principales (MIPs) forman una antigua familia de transportadores de agua y pequeños solutos a través de las membranas biológicas. La historia evolutiva y las funciones de los MIPs han sido ampliamente estudiadas en vertebrados y plantas, pero su extensa presencia a lo largo del árbol de la vida de los eucariotas sugiere una historia evolutiva más compleja y un conjunto de funciones más amplio de lo que se pensaba anteriormente. Sin embargo, la evolución temprana de los MIPs sigue siendo poco conocida. La concurrencia de hasta dos MIPs en la mayoría de los procariotas sugiere que los primeros eucariotas probablemente tenían una acuaporina (AQP) y una acuagliceroporina (GLP). En este estudio, informamos sobre una riqueza previamente desconocida de la diversidad de MIPs en eucariotas en todos los linajes principales, incluidos los protistas, que constituyen la mayor parte de la diversidad eucariota. Tres clados de MIPs probablemente tienen orígenes evolutivos profundos, que se remontan al último ancestro común eucariota (LECA) y que respaldan la presencia de un complejo repertorio de MIPs en los primeros eucariotas. En general, nuestros hallazgos resaltan la creciente complejidad del genoma LECA reconstruido: la historia evolutiva dinámica de los MIPs se puso en marcha cuando los eucariotas estaban en su infancia, seguida de explosiones radiativas en todos los linajes principales de eucariotas.

Palabras clave: acuaporina; evolución profunda en eucariotas; acuagliceroporina; último antepasado común eucariota; transporte de agua.

Introduction

Eukaryogenesis was an evolutionary singularity. One of the conundrums in early eukaryote evolution is that their last common ancestor is inferred to have possessed a wide range of features. Almost any feature that one would deem a hallmark of eukaryotes was likely present in the last eukaryotic common ancestor (LECA). These include, for example, mitochondria, complex cell cycle with meiosis, intricate intracellular organisation with an endomembrane system and organelles, actin- and tubulin-based cytoskeleton enabling intracellular trafficking and cell motility, nucleus with linear chromosomes and different chromatin states, and regulation of gene expression (Eme et al., 2017) Building on genome data across the eukaryotic tree of life, it is now possible to piece together the conserved genetic framework for these important traits. A recent comparative genomics study inferred a complex genome for LECA with ~13,000 genes (Vosseberg et al., 2021). Major intrinsic proteins (MIPs) are likely a point in case despite their early evolution remains obscure.

MIPs are channel proteins with key physiological roles as transporters of water and small solutes across biological membranes, where they form pores as tetramers, each with six transmembrane helices (Bienert et al., 2012; Walz et al., 1997). Because they are essential for life, MIPs are important drug targets for human disease (Soveral & Casini, 2017) and central in the development of drought tolerant crops (Maurel et al., 2015). MIPs have an ancient origin and are highly diversified (Abascal et al., 2014; Heymann & Engel, 1999). Most bacteria and archaea generally have one aquaporin (AQP) and one aquaglyceroporin (GLP) that function as water and glycerol transporters, respectively (Abascal et al., 2014). By contrast, eukaryotes display a much richer set of MIPs and are very diverse in terms of structure and function. The expansion of eukaryotic MIPs is often linked to tissue-specific expression in multicellular organisms (subfunctionalisation) and less frequently to functional divergence (Ikeda et al., 2011; Ishibashi, 2006; Zardoya et al., 2002). Most MIP diversity and function has been studied invertebrates (Finn et al., 2014; Finn & Cerdà, 2015) and flowering plants (Li et al., 2022; Maurel, 2007), with ~20 described MIP subfamilies in total. Yet, eukaryotic MIP diversity extends well beyond. With the advent of new sequenced genomes, MIP diversity has been increasingly explored in other eukaryotic groups such as fungi (Pettersson et al., 2005), green algae (Anderberg et al., 2011), diatoms (Khabudaev et al., 2014), the oomycete Phytophthora (Azad et al.,

2021), or kinetoplastid parasites (Tesan et al., 2021). These often report on new MIP subfamilies outside of the known MIP diversity and are suggestive of a vast yet unknown MIP diversity in understudied eukaryotic lineages.

Here, we report on the hidden diversity of MIPs across the eukaryotic tree of life. Our study takes advantage of recently available genomic data for all major eukaryotic supergroups-most of which are unicellular protists to better understand the diversity and evolution of this important protein family. By constructing a phylogenetic framework of thousands of MIPs sampled from across seven eukaryotic supergroups, we provide a comprehensive view of MIP diversity. Our data pinpoint deep orthologous relationships among MIP clades that remained hidden in taxonomically-restricted studies; we highlight that the deep diversification of MIPs can be dated back to the infancy of eukaryote evolution.

Results and discussion

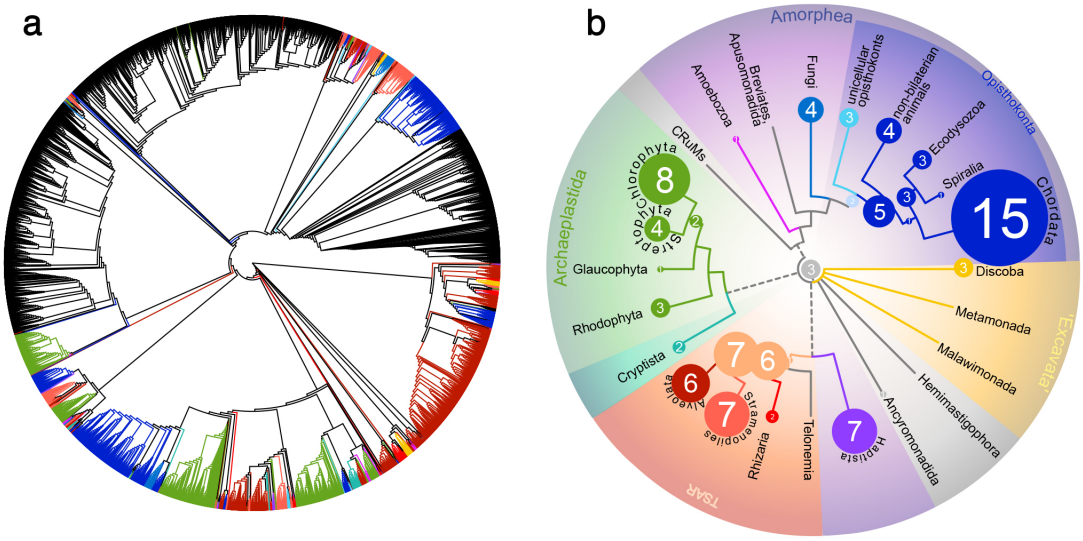


Figure 1. Overview of major intrinsic protein (MIP) diversity. (a) Maximum likelihood phylogeny of 7 ,541 proteins (full dataset, midpoint rooting). Branch colours reflect eukaryotic supergroups according to those labelled in (b); bacteria and archaea are in black (a detailed version of this tree is shown in Fig. S1). (b) MIP paralogues (circles) and their inferred evolutionary origins mapped onto the eukaryotic tree of life (Burki et al., 2020) (dotted lines represent uncertain relationships). MIP paralogues include previously described subfamilies and new clades defined here; MIP deep clades (MDC1–3; grey) are shown at the tree root. The origin of a MIP paralogue is defined by the most recent common ancestor (MRCA) of all included proteins (i.e., assuming gene loss); for clarity, some taxonomically restricted MIP

clades were summarised at higher-level lineages for (reconstructed MRCAs for each MIP clade are available in Table S1).

A previously unrecognised diversity of eukaryotic MIPs

We assembled a large dataset of 7,541 MIP homologues (Figs. 1a and S1) aiming to represent the protein family diversity and emphasising previously overlooked eukaryotic lineages (a total of 4,235 eukaryotic proteins from 484 species). Our dataset builds from a large set of eukaryotic proteins (Richter et al., 2020) as well as previous broad datasets for eukaryotes, bacteria and archaea (Abascal et al., 2014; Anderberg et al., 2011) and the MIPdb database (Delamarche & Le Bechec; http://genoweb1.irisa.fr/), and it was subjected to careful tree-based decontamination steps in order to remove likely contaminants and symbionts that can bias the inferred MIP diversity patterns. The dataset was split into ten phylogenetically-defined subsets (Fig. S1) for increased internal phylogenetic resolution in otherwise short protein alignments. The ten subsets were analysed by maximum likelihood (ML) (Fig. 2) and used to define MIP paralogue clades on the basis of phylogeny and the conservation of amino acid residues of functional or evolutionary relevance. In functionally characterised MIPs, the selectivity to transport depends on two sets of conserved amino acids that define the size and affinity of the pore: two opposite asparagine-proline-alanine (NPA) motifs that form hydrogen bonds with the water molecule and electrostatically repulse protons (Murata et al., 2000) and four residues forming the narrowest pore section that determine substrate specificity (known as the ar/R selectivity filter; Fu, 2000). In addition, five amino acids (Froger's residues or P1–P5) define MIP subfamilies and substrate transport selectivity (Froger et al., 1998).

Our datasets (Fig. 2) recovered all described MIP clades of canonical vertebrate aquaporins (AQP0, 1, 2, 4, 5, 6, 14, 15), aguaglyceroporins (AQP3, 7, 9, 10), intracellular aquaporins or superquaporins (AQP11, AQP12), as well as land plant aquaporins (PIPS, TIPS, XIPs, HIPS, SIPS, NIPs) and glyceroporins (GIPs) (Abascal et al., 2014; Finn et al., 2014). As in a very recent study (Li et al., 2022), we identified older origins for several plant MIP subfamilies than it was previously accepted (Abascal et al., 2014; Soto et al., 2012). In our dataset, TIPs can be traced back to a more ancient streptophyte algal ancestor (MRCA of Phragmoplastophyta, i.e. of Embryophyta, Zygnematophyceae, Coleochaetophyceae, and Charophyceae) and multiple duplications

in the MRCA of ferns, gymnosperms, and angiosperms (Euphyllophyta) gave rise to clades TIP2–5 (Figs. 2c and S4). SIPs likely originated in the MRCA of streptophytes and chlorophytes (Chloroplastida) and early duplications occurred in the land plant ancestor (Figs. 2f and S7). Both NIPs and their early duplications into clades NIP1–4 likely date back to the MRCA of land plants and their closest algal relatives, the Zygnematophyceae (Figs. 2g and S8). This implies an earlier HGT event from bacteria (Zardoya et al., 2002) and their immediate diversification. With regards to animal MIPs, we identified several new clades of invertebrate and holozoan MIPs that are orthologous to known vertebrate MIPs: one clade of invertebrate co-orthologues to vertebrate AQP11 and AQP12 (Figs. 2f and S7); three clades of co-orthologues to vertebrate aquaporins (Figs. 2a and S2); and four orthology groups to vertebrate aquaglyceroporins (Figs. 2h and S9). One of the invertebrate clades closely related to vertebrate aquaporins (named Metazoa I; Fig. S2) contains the recently described 'entomoglyceroporins' that have secondarily evolved glycerol selectivity in insects (Finn et al., 2015). Our tree supports deep orthology relationships of 'entomoglyceroporins' with other pancrustacean and lophotrochozoan MIPs. Beyond animal and plant MIPs, we identify all other described MIP subfamilies including the five green algal MIPs lineages MIPA–E that remain restricted to Chlorophyta (Anderberg et al., 2011), fungal aquaporins and aquaglyceroporins (Verma et al., 2014), large intrinsic proteins (LIPs) recently described in diatoms (Khabudaev et al., 2014), Phytophthora MIPs (Azad et al., 2021), and kinetoplastid AQPXs (Tesan et al., 2021). LIPs were identified as part of a clade including other ochrophytes beyond diatoms, dinoflagellates, and ciliates and thus dating back at least to the MRCA of Stramenopiles + Alveolata within SAR (Figs. 2f and S7), much older than initially thought (Khabudaev et al., 2014). The majority of Phytophthora MIPs (clades PMIP-A–H) clustered together, and likely originated from ancient duplications within oomycetes (probably the MRCA of Peronosporales and Phytiales), and the entire clade probably dates back to the Stramenopiles + Alveolata ancestor within SAR as suggested by closely related oomycetes and dinoflagellates with conserved amino acid residues (Figs. 2h and S9). The Phytophthora PMIP-1 clade originally defined by a single sequence (XP_008909057; Azad et al., 2021) is corroborated by two additional species but remains restricted to Phytophthora (Figs. 2i and S10). Kinetoplastid AQPXs are recovered as two distantly-related clades that suggest their origin by an ancient duplication predating Discoba (Figs. 2d–e and S5,6), but the overall low statistical support of basal branches and the fast evolutionary rates of

AQPXs (that make them prone to long-branch attraction artefacts) suggests that this hypothesis should be taken with caution. The possibility of independent origins of two AQPX clades could not be identified by analysing exclusively discoban proteins (Tesan et al., 2021).

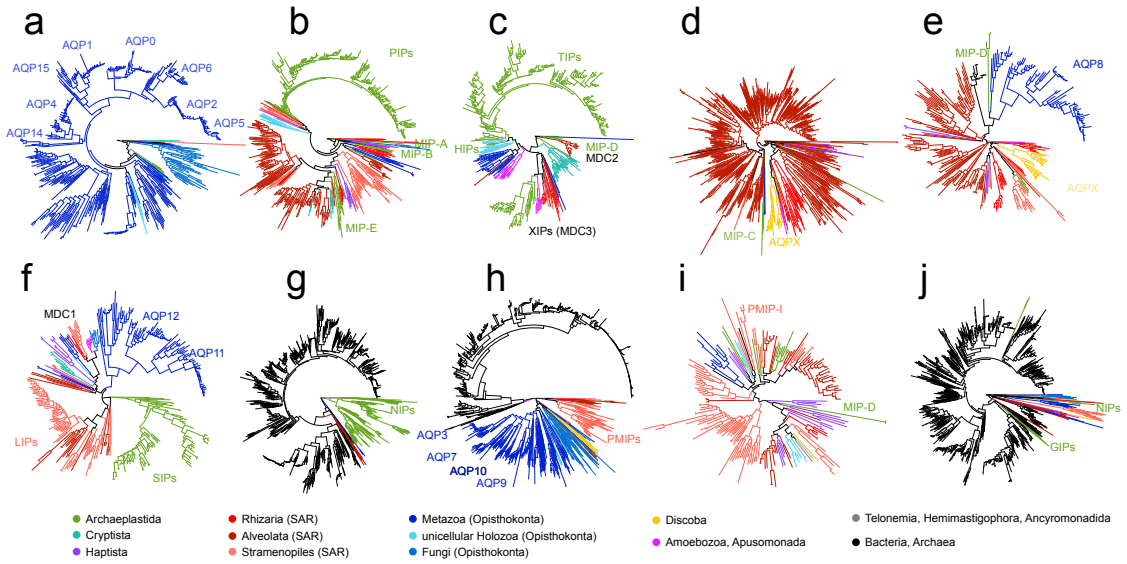


Figure 2. Schematic representation of the maximum likelihood phylogenies (midpoint rooting) of the ten data subsets (a–j) showing the diversity of major intrinsic proteins. Described MIP groups and here-defined deep MIP clades (MDC1, MDC2, MDC3) are highlighted; branch colours reflect taxonomic affiliation. Detailed versions of these trees with additional information and functional residues are shown in Figs. S2–11.

MIPs are much more diverse than previously thought: they are found across all major eukaryotic lineages and several new MIP clades lay outside of the previously described subfamilies (Fig. 1b and Table S1). Most of the new clades, defined on the basis of phylogenetic affinity and key residue conservation, correspond to unicellular eukaryotes, which have remained largely understudied with respect to MIPs. Quantitatively, the majority of the new MIP clades correspond to the TSAR supergroup (Telonemia, Stramenopiles, Alveolata, Rhizaria; Strassert et al., 2019), with one MIP clade dating back to its MRCA, five clades dating back to the SAR ancestor, five further clades to the Stramenopiles + Alveolate ancestor, one each to the MRCAs of Alveolata and Rhizaria, in addition to many taxonomically-restricted clades of dinofagellates (Alveolata; five clades), ciliates (Alveolata; 2 clades), ochrophytes (Stramenopiles, 3 clades), diatoms (Stramenopiles, 2 clades), Bigyra (Stramenopiles, 1 clade), and

foraminiferans (Rhizaria, 1 clade). Seven clades of Haptista were recovered containing representatives of either Prymnesiophyceae or Pavlovales. For Cryptista, we found two MIP clades dating back to at least the MRCA of its largest clade Cryptophyceae (no significant MIP homologues were found in Centrohelida transcriptomes). In Archaeplastida, one clade of Glaucophyta and three clades of Rhodophyta were recovered; two out of the three clades in Rhodophyta likely date back to its MRCA. A new trypanosomatid (Discoba) MIP clade was found in addition to the previously described AQPXs (Tesan et al., 2021). In Amoebozoa, a new MIP clade was identified for slime molds (Eumycetozoa). We also found a clade of Acyromonadida MIPs (Fig. 2b and S3). Three clades of choanoflagellates and one clade of ichthyosporeans were found among the closest unicellular relatives of animals (Opisthokonta, Holozoa). We also found additional MIP clades for invertebrates, including two likely dating back to the metazoan ancestor (Figs. 2b–c and S3, 4) and several taxonomically restricted clades including mollusks, copepods, brachiopods, nematodes, and sponges. Interestingly, one sponge MIP clade clusters with a bacterial homologue and might represent a horizontal gene transfer (HGT) of an aquaglyceroporin in sponges (Kenny et al., 2020), here shown to date back at least to the MRCA of Heteroscleromorpha (Fig.2i and S10).

The recovered high diversity of eukaryotic MIPs with large paralogy groups outside of the known subfamilies and the non-conservation of functional residues might be suggestive of new roles for some of these MIPs. Besides water, MIPs are known to transport small molecules including non-polar compounds such as glycerol, urea, or lactic acid; reactive oxygen species (ROS) and hydrogen peroxide, gasses (ammonia, carbon dioxide, nitric oxide), and metalloids (boron, silicon, arsenic, antimony)(Gupta & Sankararamakrishnan, 2009; Mukhopadhyay et al., 2014). This broad range of compounds highlights the functional versatility of MIPs and their roles in many cellular processes. In the case of unicellular eukaryotes, the presence of multiple MIPs likely provides a better control of solute transport compared to the passive exchange through the membrane (larger eukaryotic cells have a lower surface-to-volume ratio and thus slower equilibration times) (Tanghe et al., 2006) and the presence of specific MIPs for vacuoles (TIPs) and intracellular membranes (SIPs, AQP11, AQP12) further supports the possibility of functional diversification. MIPs are also central for stress responses to abiotic and biotic factors including low temperature and freezing (membrane permeability is reduced) (Tanghe et al., 2006), plant-mycorrhizal symbioses, the

formation of resistance forms such as spores (Tanghe et al., 2006), or the maturation of siliceous cell walls in diatoms (Grachev et al., 2008). Further functional and structural studies of the vast eukaryotic MIPs diversity is likely to broaden the set of functions performed by them. For example, tripanosomatid AQPXs are likely poor transporters of water and glycerol and possess wider selectivity filters to permeate larger (so far unknown) solutes (Tesan et al., 2021). From a biotechnological viewpoint, MIPs have been proposed as drug targets against fungal (Verma et al., 2014) and tripanosomatid (Tesan et al., 2021) parasites.

MIPs are not only present across all major eukaryotic supergroups—they diversified within each. The number of available genomes and transcriptomes per eukaryotic supergroup predicts well the number of MIP homologues in the final dataset (Pearson's R=0.93, p=0.0021; Fig. S12). While a definitive relationship between new MIPs and genomic data availability cannot be inferred here because transcriptomic data do not represent the full gene repertoire of a species, it is clear that additional genomic data on lesser-studied deep eukaryotic groups is needed to complete the understanding of MIP diversity. These data push MIPs of unicellular eukaryotes into the limelight.

A complex repertoire of MIPs in LECA

Despite overall low statistical support for many basal relationships among defined MIP paralogy groups (expected for short proteins that diverged a long time ago), we identified further MIP paralogue clades that might date back to ancient gene duplications during the early evolution of eukaryotes. We identified a strongly supported clade termed MDC1 (MIP deep clade 1) that groups several distantly related unicellular protists including slime molds (Amoebozoa, Eumycetozoa), blastocladiomycete and chytridiomycete fungi (Fungi), golden algae (Stramenopiles, Ochrophyta), and rhizarians such as *Paulinella* (Rhizaria; Figs. 2f and S7). Froger's residues suggest that these might be aquaporins (e.g., small uncharged P2–P3 and aromatic P4–5) but display aromatic amino acids in P1 typical of GLPs (Fig. 3a). A second clade of deep evolutionary origin (termed MDC2) is that formed by chlorophyte algal MIPD (Archaeplastida) and a dinoflagellate (Alveolata) clade (Figs. 2c, 3b and S4), which received strong statistical support and displays conserved key residues. This clade could represent the first deep orthology proposed for any of the enigmatic clades named MIPA–E that are exclusively found in green algae (Archaeplastida,

Chlorophyta). Residue conservation suggests that MDC2 might be aquaporins (e.g., non-aromatic P1, small uncharged P2–P3, and aromatic P4), but also have non-aromatic P5 as in GLPs. The third deep clade (MDC3) is that of XIPs, which in our analyses encompass plant XIPs with slime molds and other amoebas (Amoebozoa), chlorarachnophyte algae (Rhizaria), and one diatom (Stramenopiles) (Figs. 2c, 3c and S4). A deep evolutionary origin of XIPs has already been proposed based on phylogenetic clustering of plant XIPs with *Dictyostelium* (Amoebozoa) and fungi (Abascal et al., 2014; Danielson & Johanson, 2008; Gupta & Sankararamakrishnan, 2009). Previously identified fungal XIPs are recovered elsewhere in our analyses (Figs. 2j and S11) likely due to a long branch attraction artefact, but the conservation of key residues (Fig. 3c) and a set of synapomorphic amino acids (Abascal et al., 2014) are strong indicators of their deep orthology. All three deep MIP clades are likely salient to very ancient gene duplication events. They contain representatives of both Obazoa (e.g., Opisthokonta, Amoebozoa) and Diaphoretickes (e.g., SAR, Archaeplastida). According to the current understanding of the deep eukaryotic phylogeny (Burki et al., 2020) (Fig. 1b), such duplication date back straight to LECA under the assumption of a unikont/bikont root (Derelle et al., 2015), or would date back to the second deepest node in the tree under the assumption of a neozoan/excavate root (He et al., 2014). According to the latest molecular clock estimations, these events occurred >2 billion years ago (Strassert et al., 2021).

Two other MIP clades could represent additional instances of deep orthology within Diaphoretickes: one clade of TSAR and Haptista (Fig. S6) and another clade with representatives of SAR+ Cryptista (Fig. S3). However, such genes might be affected by endosymbiotic gene transfer (EGT) events, as the above lineages were involved in the ancient serial endosymbioses that gave rise to secondary red-algal plastids (Strassert et al., 2021). Abascal et al. (2014) suggested the possibility of deep orthology for plant XIPs + HIPs + TIPs + animal AQP8, and for plant PIPs + animal aquaporins (AQPs 0–2 and 4–6) but also warned about the possibility for phylogenetic artefacts and low statistical support for basal tree branches. Ishibashi et al. (2017) considered intracellular SIPs, AQP11, and AQP12 a subfamily on their own (superaquaporins) but the non-conservation of key amino acids and the possibility of a long-branch attraction artefact question their true deep orthology (Abascal et al., 2014).

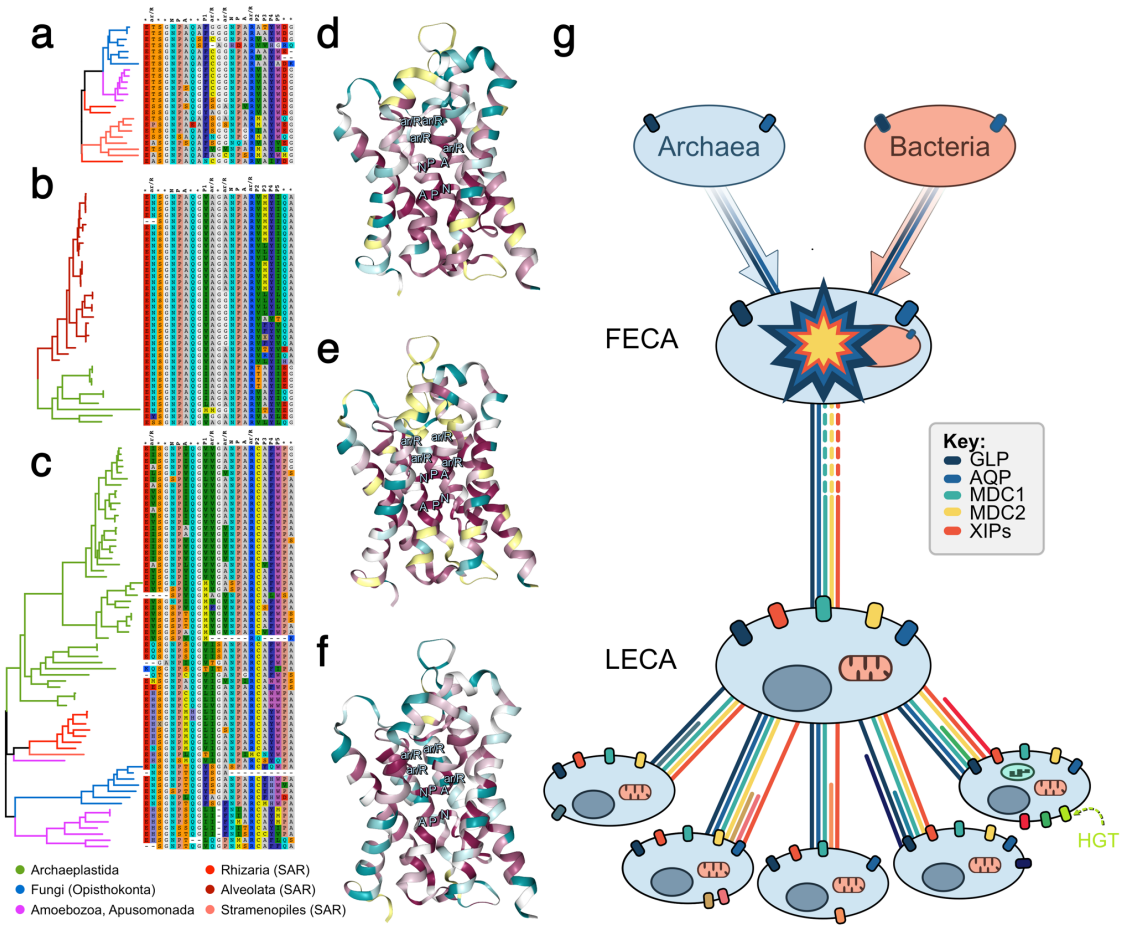


Figure 3. Early evolution of eukaryotic MIPs. Residue conservation for the three MIP clades of likely ancient origin highlighted in the main text. Alignments highlight key amino acids of functional and evolutionary relevance (tree branches coloured by eukaryotic lineage) while structures reflect overall amino acid conservation as inferred with ConSurf for (a, d) MDC1, (b, e) MDC2, (c, f) XIPs. (g) Hypothetical scenario for the early evolution of MIPs, showing the transition from the first (FECA) to the last eukaryotic common ancestor (LECA) and early duplications within main eukaryotic supergroups; one horizontal gene transfer (HGT) event is also shown.

The presence of MIP clades whose origin can be traced back to deepest nodes in the eukaryotic tree of life is indicative of a dynamic evolutionary pattern of gene duplications, losses, and divergences in the MIP family already very early in eukaryotic evolution. In particular, some of the deep MIP clades might be traceable to LECA, suggesting a complex ancestral repertoire of MIP homologues for the first eukaryotes. This is apparently counterintuitive considering that most bacteria have one AQP and one GLP, and archaea have either one AQP, one GLP, or no MIPs (Abascal et al., 2014; Tanghe et al., 2006; Zardoya et al., 2016). Despite the difficulties of reconstructing the

evolution of MIPs in bacteria and archaea, mainly due to high HGT rates and the lack of phylogenetic component in MIP contents that seem to depend on lifestyle (Abascal et al., 2014), it is conceivable that FECA had at least one aquaporin and one aquaglyceroporin (Abascal et al., 2014; Ishibashi et al., 2017). This reconstruction agrees with the identification of MIP homologues in the genomic repertoire of the first eukaryotes (Ku et al., 2015; Pittis & Gabaldón, 2016). Our results further indicate a rather dynamic evolution of MIPs in the earliest eukaryotes and point out to the presence of at least five MIP paralogues (MDC1, MDC2, XIP, plus the ancestral AQP and GLP) very early in eukaryote evolution and probably directly in LECA (Fig. 3d). We further stress that this is likely an underestimation, given the difficulty of identifying ancient orthology relationships with certainty in short protein alignments.

Conclusions

MIPs have undergone pronounced expansions within most of the eukaryotic supergroups, mostly by gene duplication but also through non-vertical inheritance such as EGT and HGT from bacteria (Fig. 3d). Our hypothesised scenario roughly agrees with the numbers of MIP homologues recovered by a large-scale protein clustering (Ku et al., 2015) that found three ancestral (eukaryote-prokaryote clusters) and 23 eukaryote-specific MIP clusters. The early burst of MIP homologues in eukaryotes are also in line with the high rates of gene duplication inferred for early eukaryotes, which apparently doubled the number of genes in the transition from FECA to LECA (Vosseberg et al., 2021). An early diversity in MIPs suggests an early diversification in function. This diversification likely facilitated complexity in physiological properties in an ancient single-celled eukaryote, allowing a versatile transport of small solutes. Versatility in solute transport through MIPs thus is a cornerstone of eukaryotic functions that can be traced back to LECA—if not to an earlier infancy of eukaryote evolution.

Acknowledgements

This work used the Scientific Compute Cluster at GWDG, the joint data centre of Max Planck Society for the Advancement of Science (MPG) and University of Gottingen. H.L-M. was sponsored by a predoctoral contract of the Complutense

University of Madrid and the Real Colegio Complutense at Harvard University (RCC-UCM), and an Erasmus+ research mobility traineeship from the European Commission. J.d.V. thanks the European Research Council for funding under the European Union's Horizon 2020 programme (Grant Agreement No. 852725; ERC-StG "TerreStriAL") and the German Research Foundation (DFG) within the framework of the Priority Programme "MAdland-Molecular Adaptation to Land: Plant Evolution to Change" (SPP 2237; VR132/4-1), in which I.I. partakes as associate members.

Author Contributions

Conceptualisation, I.I., J.d.V., H.L.-M.; Methodology, I.I., J.d.V., H.L.-M.; Formal analysis and Investigation, I.I., H.L.-M.; Data Curation, I.I., H.L.-M. J.F.H.S., Writing - Original draft, I.I; Writing - Review & Editing, all authors; Visualisation, I.I., J.d.V.; Administration and Funding, J.d.V.

Declaration of Interests

The authors declare no competing interests.

Material and methods

MIP dataset assembly

A large dataset of MIP homologues was assembled from the following sources: (I) the most taxonomically broad analysis of MIPs (Abascal et al., 2014), (II) bacterial and archaeal homologues in MIPdb (http://mipdb.genouest.org/), (III) chlorophyte MIPs (Anderberg et al., 2011), and (IV) newly identified eukaryotic MIP homologues from EukProt (Richter et al., 2020).Thus far, EukProt is the largest and taxonomically broadest collection of eukaryotic genomes and transcriptomes (742 species, with emphasis on unicellular eukaryotes. MIP homologues in EukProt were identified by (I) BLASTP v2.11 .0 searches (Altschul et al., 1990) using 105 MIP proteins from *Homo sapiens*, *Arabidopsis* spp., *Trypanosoma* spp. and *Capsaspora owczarzaki* as queries that represent the known MIP diversity (e-value threshold: 1e-20), and (II) HMMER v3.3 (Finn et al., 2011) searches using Pfam's MIP Hidden Markov Model (HMM) profile (PF00230). In both cases, hits were required to align at least 150 amino acids.

MIP homologues from all four sources were merged into a single dataset of 8,308 proteins, aligned with MAFFT v7.304 (Katoh & Standley, 2013) ('--auto') and alignment columns with >90% missing data were removed with trimAL v1.3 (Capella-Gutiérrez et al., 2009).A ML phylogeny was inferred with IQ-Tree v1 .6.12 (Nguyen et al., 2015) under BIG-selected best-fit substitution model LG+F+Γ4 and branch support assessed with 1,000 replicates each of ultrafast bootstrapping (Hoang et al., 2018) and SH-like approximate likelihood ratio test (Guindon et al., 2010).

Dataset decontamination

The initial tree was used to identify and remove duplicates and likely contaminants based on unexpected phylogenetic clustering. For this, the tree was visualised with FigTree v1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/) using colours for the different eukaryotic supergroups as obtained from the NCBI taxonomy (Schoch et al., 2020). Proteins that clustered outside of their respective eukaryotic lineage (except if three or more different species were present), were excluded. Small clades containing a mix of very different eukaryotic groups where BLASTP-searched against non-redundant (NR) database and removed if their best hits were from a different eukaryotic supergroup. In the event of multiple versions (or isoforms) of the same protein, the one with the shortest branch was retained. After a first round of decontamination, the resulting clean dataset was re-aligned and a new ML tree was constructed as detailed above. The tree was used to select 18 subsets of more closely related sequences, for which independent ML trees were built upon re-alignment and trimming (as detailed above) to facilitate identification of contaminants among less divergent proteins. The 18 cleaned subsets were combined, re-aligned, trimmed, and a new inclusive ML tree was inferred as detailed above. A third round of decontamination specifically targeted known endosymbionts or food sources, as detected in a previous study that used EukProt data (Strassert et al., 2021). For instance, dinoflagellates (Alveolata) inside ochrophyte (Stramenopiles) or *Sorites* (Rhizaria) could be endosymbionts; *Paramecium* clustering in algal clades are likely symbionts; *Tiarina* transcriptomes are highly contaminated, often with ochrophytes and diatoms; *Pseudokeeronopsis* sp. Brazil is also often contaminated by ochrophytes; *Durinskia baltica* is often contaminated with Chromulinaceae; *Colponema* transcriptomes are often contaminated by the excavate *Procyptobia* used to feed cultures. Finally, all

proteins were screened again for the presence of the MIP domain using hmmsearch (e-value threshold: 1 e-3) and non-significant hits were further checked by BLASTP searches against NR, resulting in the exclusion of ten proteins. The final cleaned dataset contained a total of 7,541 proteins.

Phylogenetic inference and identification of MIP clades

The set of 7,541 MIPs was subjected to alignment (MAFFT --auto), trimming (trimAL -gt 0.9), and IQ-Tree ML phylogenetic inference with BIG-selected best-fit LG+F+Γ4 model of amino acid replacement. Branch support was assessed by transfer bootstrap expectation (TBE; Lemoine et al., 2018) as calculated by RAxML-NG v1.1.0 (Lutteropp et al., 2020) using 1,000 standard bootstrapped trees calculated in IQ-TREE. To gain phylogenetic resolution in the study of MIP clades, we partitioned the large dataset of 7,541 sequences into ten subsets of closely related sequences using the ML tree as a guide; each subset was then subjected to alignment, trimming and independent ML inference as detailed above (support values in this case were UFBoot and SH-like LRT with 1,000 replicates).

The ten smaller ML trees were used to define MIP clades. All eukaryotic MIP clades described in the literature were searched in the ten subsets. The remaining diversity of eukaryotic MIP was then systematically defined by taxonomic clades mostly restricted to a single eukaryotic supergroup total or partially (highly-supported clades with SH-aLRT >0.85 or UFBoot >0.95 were favored). For each defined clade, we inferred the most recent common ancestor based on the taxonomic representation in our dataset (Table S1), which was then used as base for Fig. 1b.

Conservation of amino acid residues

To further understand the evolutionary conservation of key residues, we extracted previously defined key residue homologues, as identified after re-aligning each of the ten subsets with the *Escherichia coli* aqpZ reference sequence (Uniprot accession P60845). Residues include (I) the two NPA boxes (N63, P64, A65, N 186, P187, A 188), (II) four ar/R selectivity filters (F43, H 17 4, T183, R189), (III) five Froger residues (Froger et al., 1998) (1103, 5190, A194, F208, W209), and (IV) conserved residues identified by Abascal et al. (2014; EB, S58, G59, 088, G91, N182,

G212, G215). Residues were plotted with ETE3 (Huerta-Cepas et al., 2016). For the three identified deep MIP clades, the conservation of amino acids was inferred with ConSurf (Ashkenazy et al., 2016) and plotted onto the Arabidopsis thaliana PIP2–4 structure (Uniprot accession 6QIM) using custom alignments (MAFFT E-INS-i) and ML trees (IQ-Tree).

Statistical analyses

The correlation between the initial number of EukProt protein sets per eukaryotic supergroup and number of MIP homologues in the final dataset was tested with Pearson's correlation test in R.

SUPPLEMENTARY MATERIAL

Table S1. MIP clades as defined here based on phylogenetic trees from the ten data subsets and their reconstructed evolutionary origin (most recent common ancestor).

Figure	MIP clade	Inferred MRCA
Fig. 2a	AQP5	Synapsida + Diapsida (Chordata)
	AQP2	Synapsida + Diapsida (Chordata)
	AQP6	Tetrapoda (Chordata)
	AQP0	Gnathostomata (Chordata)
	AQP1	Gnathostomata (Chordata)
	AQP15	Gnathostomata (Chordata)
	AQP4	Tetrapoda (Chordata)
	AQP14	Gnathostomata (Chordata)
	Metazoa I	Ecdysozoa + Lophotrochozoa
	Metazoa II	Metazoa
	Metazoa III	Metazoa
	Archaeplastida I	Gaucophyta
	Fungi I	Fungi
Fig. 2h	AQP3	Craniata (Chordata)
	AQP7	bony fishes (Osteichthyes)
	AQP10	bony fishes (Osteichthyes)
	AQP9	bony fishes (Osteichthyes)
	Metazoa IV	Metazoa
	Metazoa V	Ecdysozoa + Lophotrochozoa
	Opisthokonta I	Opisthokonta
	Opisthokonta II	Opisthokonta
	Fungi II	Fungi
	Discoba I	Kinetoplastea
	SAR I	Alveolata + Stramenopiles
Fig. 2e	AQP8	Bilateria
	MIPD	Chlamydomonaceae (Chlorophyta)
	SAR II	Alveolata + Stramenopiles
	Haptista I	Prymnesiophyceae
	SAR III	Dinophyceae + Colpodellida (Alveolata)
	TSAR I	TSAR + Haptista
	SAR IV	SAR + Rhodophyta
	AQPX	Kinetoplastea
	SAR V	Rhizaria

Fig. 2f	AQP11	Gnathostomata (Chordata)
	AQP12	Gnathostomata (Chordata)
	Metazoa VI	Ecdysozoa + Lophotrochozoa
	MDC1	LECA
	Metazoa VII	Gastropoda+Bivalvia
	Haptista II	Prymnesiophyceae
	Cryptista I	Cryptophyta
	SAR VI	Alveolata + Stramenopiles
	LIPs	Alveolata + Stramenopiles
	SAR VII	Alveolata + Stramenopiles
	SIPs	Chlorophyta
Fig. 2j	Archaeplastida II	Galdieria
	GIPs	Chlorophyta + Streptophyta
	SAR VIII	Ochrophyta
	Metazoa VIII	Calanoida (Copeppoda)
	NIPs II	Poales
	Fungi XIPs (MDC3)	Pezizomycotina
Fig. 2d	SAR IX	Alveolata + Stramenopiles
	MIPC	Mamiellaceae
	AQPX	Kinetoplastea
	SAR X	Foraminifera (Rhizaria)
	SAR XI	Ciliophora
	SAR XII	Dinophyceae (Alveolata)
Fig. 2i	SAR XIII	SAR
	Archaeplastida III	Rhodophyta
	Metazoa IX	Porifera
	SAR XIV	Phytophthora PMIP-I
	SAR XV	Bigyra (Stramenopiles)
	SAR XVI	Bacillariophyta (Stramenopiles)
	SAR XVII	Dinophyceae (Alveolata)
	Haptista III	Prymnesiophyceae
	Opisthokonta III	Choanoflagellata
	Haptista IV	Prymnesiophyceae
	Haptista V	Pavlovales
	MIPD	Volvox (Chlorophyta)
	Haptista VI	Prymnesiophyceae

Fig. 2g	SAR XVIII	Dinophyceae (Alveolata)
	Metazoa X	Lingula (Brachiopoda)
	Metazoa XI	Caenorhabditis (Nematoda)
	NIPs	Embryophyta
Fig. 2b	PIPs	Chloroplastida
	SAR XIX	Phaeophyceae + Xantophyceae (Ochrophyta, Stramenopiles)
	Opisthokonta IV	Ichthyosporea
	SAR XX	SAR
	SAR XXI + Cryptista	Alveoltata + Stramenopiles + Cryptista
	MIPE	Chlorophyta
	TSAR II	TSAR
	Haptista VII	Prymnesiophyceae
	SAR XXII	Ochrophyta (Stramenopiles)
	SAR XXIII	Bacillariophyta (Stramenopiles)
	SAR XXIV	Dinophyceae
	Metazoa XII	Metazoa
	SAR XXV	SAR
	Archaeplastida IV	Rhodophyta
	Ancyromonadida I	Ancyromonadida
	MIPA	Coccomyxa (Chlorophyta)
	MIPB	Ostreococcus (Chlorophyta)
	Fungi III	Microsporidia
	SAR XXVI	SAR
Fig. 2c	TIPs	Streptophytes
	Opisthokonta V	Salpingoecidae (Choanoflatellata)
	HIPs	Embryophyta (Viridiplantae)
	Opisthokonta VI	Salpingoecidae (Choanoflatellata)
	Metazoa XIII	Metazoa
	Amoebozoa I	Eumycetozoa
	XIPs (MDC3)	LECA
	Cryptista II	Cryptophyta
	MDC2 (incl. MIPD)	LECA

Trees represented in figures S1–S11 are too big to be presented in this Ph. D. thesis document. Instead they are available for visualisation and downloading at: **https://doi.org/10.5281/zenodo.7455743**. Figure captions of these trees are listed below:

Figure S1. Maximum likelihood tree of 7,541 MIP homologues (IQ-TREE, midpoint rooting). Branch support is indicated as transfer bootstrap expectation values. The ten data subsets that are independently analysed are indicated in the tree.
Figure S2. Maximum likelihood tree of alignment subset in Fig. 2a (IQ-TREE, midpoint rooting). Branch support is indicated as SH-aLRT and UFBoot, respectively. Taxa are coloured according to taxonomy and MIP clades mentioned in the text are highlighted. Key amino acid residues are shown.
Figure S3. Maximum likelihood tree of alignment subset in Fig. 2b (IQ-TREE, midpoint rooting). Branch support is indicated as SH-aLRT and UFBoot, respectively. Taxa are coloured according to taxonomy and MIP clades mentioned in the text are highlighted. Key amino acid residues are shown.
Figure S4. Maximum likelihood tree of alignment subset in Fig. 2c (IQ-TREE, midpoint rooting). Branch support is indicated as SH-aLRT and UFBoot, respectively. Taxa are coloured according to taxonomy and MIP clades mentioned in the text are highlighted. Key amino acid residues are shown.
Figure S5. Maximum likelihood tree of alignment subset in Fig. 2d (IQ-TREE, midpoint rooting). Branch support is indicated as SH-aLRT and UFBoot, respectively. Taxa are coloured according to taxonomy and MIP clades mentioned in the text are highlighted. Key amino acid residues are shown.
Figure S6. Maximum likelihood tree of alignment subset in Fig. 2e (IQ-TREE, midpoint rooting). Branch support is indicated as SH-aLRT and UFBoot, respectively. Taxa are coloured according to taxonomy and MIP clades mentioned in the text are highlighted. Key amino acid residues are shown.
Figure S7. Maximum likelihood tree of alignment subset in Fig. 2f (IQ-TREE, midpoint rooting). Branch support is indicated as SH-aLRT and UFBoot, respectively. Taxa are coloured according to taxonomy and MIP clades mentioned in the text are highlighted. Key amino acid residues are shown.
Figure S8. Maximum likelihood tree of alignment subset in Fig. 2g (IQ-TREE, midpoint rooting). Branch support is indicated as SH-aLRT and UFBoot, respectively. Taxa are coloured according to taxonomy and MIP clades mentioned in the text are highlighted. Key amino acid residues are shown.
Figure S9. Maximum likelihood tree of alignment subset in Fig. 2h (IQ-TREE, midpoint rooting). Branch support is indicated as SH-aLRT and UFBoot, respectively. Taxa are coloured according to taxonomy and MIP clades mentioned in the text are highlighted. Key amino acid residues are shown.
Figure S10. Maximum likelihood tree of alignment subset in Fig. 2i (IQ-TREE, midpoint rooting). Branch support is indicated as SH-aLRT and UFBoot, respectively. Taxa are coloured according to taxonomy and MIP clades mentioned in the text are highlighted. Key amino acid residues are shown.
Figure S11. Maximum likelihood tree of alignment subset in Fig. 2j (IQ-TREE, midpoint rooting). Branch support is indicated as SH-aLRT and UFBoot, respectively. Taxa are coloured according to taxonomy and MIP clades mentioned in the text are highlighted. Key amino acid residues are shown.

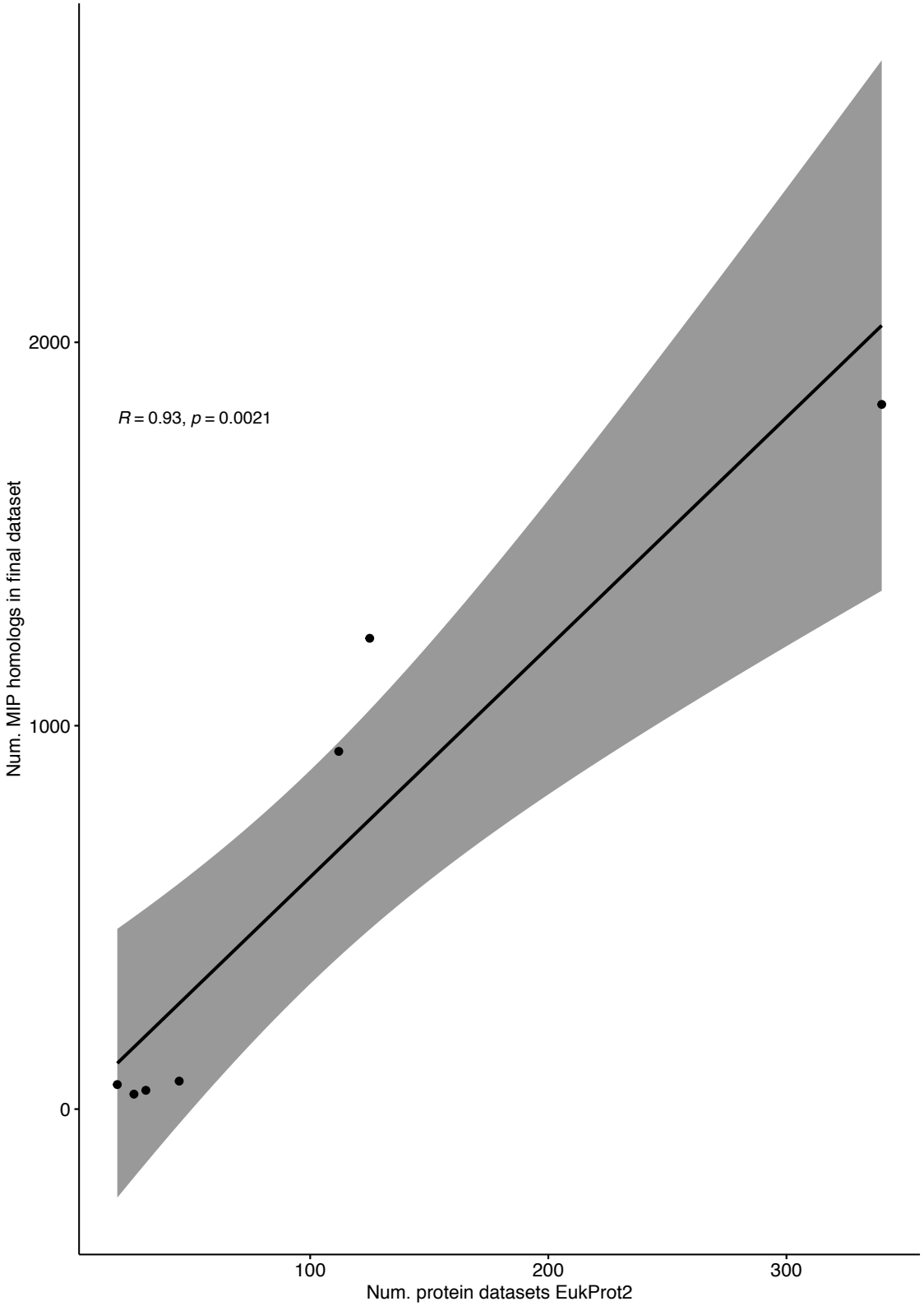


Figure S12. Correlation plot between the number of MIP homologues in the final 701 dataset per eukaryotic supergroup and the number of protein sets (i.e., genomes, transcriptomes) in EukProt for the different eukaryotic supergroups Pearson's correlation coefficient and the associated *P* value are shown.

CHAPTER III

Evidence of positive selection suggests possible role of aquaporins in the water-to-land transition of mudskippers

---

La evidencia de selección positiva sugiere un posible rol de las acuaporinas en la transición del agua a la tierra de los saltarines del fango

This chapter is a reproduction of the following article:

Lorente-Martínez, H., Agorreta, A., Torres-Sánchez, M., & San Mauro, D. (2018). Evidence of positive selection suggests possible role of aquaporins in the water-to-land transition of mudskippers. *Organisms Diversity & Evolution*, 18(4), 499–514.

Abstract

Aquaporins are integral membrane proteins that exchange water and small solutes. They played an important role in the colonization of terrestrial environments by tetrapod ancestors via the appearance of three exclusive paralogues. Like early tetrapods, mudskippers represent an independent case of amphibious lifestyle evolution that is unparalleled by other extant fish groups. Given this lifestyle parallelism and that aquaporins were relevant for tetrapod terrestrialisation, this study examines the aquaporins in mudskippers to investigate whether similar changes in aquaporins could have possibly occurred during their water-to-land transition. We have catalogued aquaporin genes in four mudskipper genomes, and studied their diversity and molecular evolution (including detection of positive selection) in a broad phylogenetic context of vertebrates. Our genomic screening returned 55 aquaporin genes for mudskippers (none of them constituting novel paralogues) that can be assigned to 10 different known classes. We detected signatures of positive selection in AQP10a and AQP11b in mudskippers (both entire clade and the clade containing the most terrestrial species, implying different evolutionary times). This suggests possible alteration of the molecular function of such paralogues caused by changes at specific protein sequence positions, some of them located in relatively close proximity to parts of the molecule involved in pore formation and substrate selectivity. Given the importance of aquaporins for osmotic regulation in fishes, it might be possible that these selective changes (perhaps allowing permeability to new solutes) could have played a role during the adaptation of mudskippers to an amphibious lifestyle.

Keywords: amphibious lifestyle, aquaporin, molecular evolution, mudskipper, positive selection.

Resumen

Las acuaporinas son proteínas integrales de membrana que intercambian agua y pequeños solutos y que desempeñaron un papel importante en la colonización de los ambientes terrestres por antepasados de los tetrápodos a través de la aparición de tres parálogos exclusivos. Al igual que los primeros tetrápodos, los saltarines del fango representan un caso independiente de evolución del estilo de vida anfibio sin precedentes en otros grupos de peces existentes. Dado este paralelismo de estilo de vida y que las acuaporinas fueron relevantes para la terrestrialización de los tetrápodos, este estudio examina las acuaporinas en los saltarines del fango para investigar si podrían haber ocurrido cambios similares en ellas durante su transición del agua a la tierra. Hemos catalogado los genes de acuaporinas en cuatro genomas de saltarines del fango, y estudiado su diversidad y evolución molecular (incluida la detección de selección positiva) en un amplio contexto filogenético de vertebrados. Nuestra búsqueda genómica dio lugar a 55 genes de acuaporinas para los saltarines del fango (ninguno de ellos constituye nuevos parálogos) que pueden asignarse a 10 clases conocidas diferentes. Detectamos señales de selección positiva en las AQP10a y AQP11b de los saltarines del fango (tanto en el clado completo como en el clado que contiene la mayoría de las especies terrestres, lo que implica diferentes tiempos evolutivos). Esto sugiere una posible alteración de la función molecular de dichos parálogos causada por cambios en posiciones específicas de las secuencias de proteínas, algunas de ellas localizadas relativamente cerca de partes de la molécula involucradas en la formación del poro y la selectividad del sustrato. Dada la importancia de las acuaporinas para la regulación osmótica en los peces, es posible que estos cambios selectivos (quizás permitiendo la permeabilidad a nuevos solutos) pudieran haber jugado un papel importante durante la adaptación de los saltarines del fango a un estilo de vida anfibio.

Palabras clave: estilo de vida anfibio, acuaporina, evolución molecular, saltarín del fango, selección positiva.

Introduction

Aquaporins (AQPs), also called water channels, are a family of transmembrane protein channels of the major intrinsic proteins (MIP) superfamily (Agre, Preston, et al., 1993; Connolly et al., 1998). These proteins have a molecular mass of 24–30 kDa, and their main function is the transport of water and small uncharged solutes (e.g. glycerol, ammonia) through the membranes of the cells (Agre & Kozono, 2003; Heymann & Engel, 1999; Jahn et al., 2004; Wu & Beitz, 2007). The first evidence of the presence of proteins able to transport water through the cell membrane was documented in bovine lens tissue (Broekhuyse et al., 1976), but the first genuine aquaporin was isolated from human erythrocytes and renal tubules membranes (Preston & Agre, 1991). Such protein, originally named CHIP28 due to its molecular mass of 28kDa, is now clustered within the aquaporin class AQP1 (Agre, Sasaki, et al., 1993). The reconstruction of the molecular structure of AQP1 showed that this protein is constituted by monomers of six α-helices (1–6) embedded in the cell membranes that are connected by five regions (A–E) that loop at either side of the membrane (loops A, C and E are extracellular whereas loops B and D are intracellular). Loops B and E are hydrophobic and present a transmembrane asparagine-proline-alanine (NPA) domain each, which are related with pore formation. Besides, there is a selectivity filter, the aromatic/arginine (ar/R), made of two groups of amino acids that delimit the diameter and the hydrophobia of the pore, determining the substrate specificity (Cheng et al., 1997; Heymann & Engel, 2000; Sui et al., 2001). Each monomer constitutes a water channel and clusters with the other three to form tetramer structures in the membranes. This basic molecular pore structure remains highly conserved in most of the different aquaporin paralogues that have been discovered more recently (Kruse et al., 2006).

Early phylogenetic studies on MIPs found an ancient split between water and glycerol transporters that occurred early in the evolution of cell (Heymann & Engel, 1999; Park & Saier, 1996; Zardoya, 2005). This division was later confirmed by more recent phylogenetic studies that have also refined the classification of aquaporins in four main groups: the aquaglyceroporins (traditionally containing aquaporin classes/paralogues AQP3, AQP7, AQP9 and AQP10), AQP8-type or ammonia channel (containing class AQP8), unorthodox aquaporins (containing classes AQP11 and AQP12) and water-selective classical aquaporins (AQP0, AQP1, AQP2, AQP4, AQP5 and AQP6) (Abascal et al., 2014; Cerdà & Finn, 2010; Finn et al., 2014; Finn & Cerdà, 2011, 2015). Of these classical aquaporins, three classes (AQP2, AQP5 and AQP6) are exclusive to the tetrapod lineage (Cerdà & Finn, 2010; Finn &

Cerdà, 2011; Tingaud-Sequeira et al., 2010). More recently, four new classes of aquaporins have been reported (at least present in some vertebrate lineages): AQP13 (clustered with aquaglyceroporins), AQP14 and AQP15 (clustered with classical aquaporins) and AQP16 (clustered with AQP8-type) (Finn et al., 2014; Virkki et al., 2002).

Finn et al. (2014) postulated that aquaporins played an important role in the colonisation of terrestrial environments by tetrapods during the Late Devonian (Carrol, 2001) thanks to the appearance of the three new paralogues (AQP2, AQP5, AQP6) exclusive to the group. Water-to-land transition involves several morphological (biomechanical) and physiological (metabolic and biochemical) changes in order to confront new environmental conditions, including locomotion, vision, audition, air-breathing and desiccation problems (J. B. Graham, 1997; Johanson, 2011). In this context, aquaporins have been associated with changes in osmoregulation related to desiccation control (Finn et al., 2014). Apart from early tetrapods, the achievement of an amphibious lifestyle has evolved independently many times in vertebrates, and there are reports of over 200 extant species of fishes (spanning 17 orders) that present various degrees of terrestrial adaptation (J. B. Graham, 1997; Ord & Cooke, 2016; Wright & Turko, 2016). Among all such fish groups with amphibious behaviour (e.g. blennies, eels, bichirs), mudskippers are clearly the ones with the highest degree of terrestrialisation (Wright & Turko, 2016; You et al., 2014). Their case is likely mirroring more closely what happened with early tetrapod ancestors during their water-to-land transition in the Late Devonian. In this sense, mudskipper species with varying degrees of adaptation to an amphibious lifestyle could be comparable (although not strictly) to fossil forms of early tetrapod lineages showing intermediate forms of terrestriality (Ishimatsu & Gonzales, 2011; Wright & Turko, 2016; You et al., 2014).

Mudskippers (family Gobiidae; subfamily Oxudercinae) are amphibious fishes that are able to spend extended periods out of water. They emerge during low tide and keep very active on land (i.e. to forage, seek mates and defend territories) for some portion of their daily cycle (Jaafar & Murdy, 2017; Wright & Turko, 2016). However, their dependence on water is still important for reproduction (they possess aquatic larvae) and to minimise desiccation. For this latter, mudskippers roll on mud, immerse in small pools that remain at low tides or frequently move from land to burrows built in the mud that are submerged at high tides or retain water inside at low tides. These burrows are also used as refuges to hide from predators and for laying eggs (J. B. Graham & Lee, 2004; Ishimatsu & Gonzales, 2011; Tsuhako et al., 2003; Wright & Turko, 2016). Mudskippers live on mudflats and mangrove swamps, and have colonised peritidal habitats of tropical and subtropical Western Africa, the Indian Ocean,

and the whole Indo-West Pacific region (Murdy, 2011). Only four oxudercine genera are strictly considered to be mudskippers, namely *Boleophthalmus, Periophthalmodon*, *Periophthalmus* and *Scartelaos,* and they present different degrees of adaptation to terrestrial conditions. *Periophthalmus* and *Periophthalmodon* are the most terrestrial spending the majority of time out of water, although their degree of terrestriality gradually varies among species of these genera. In contrast, *Scartelaos* is predominantly aquatic (spending less time out of water) followed by less water-dependent *Boleophthalmus* (J. Graham et al., 2007; Ishimatsu & Gonzales, 2011; Polgar et al., 2017; You et al., 2014; Zhang et al., 2003). Mudskippers have developed specific adaptations to mudflat life in both aquatic and terrestrial conditions, including aerial respiration (through the skin and highly vascularised buccopharyngeal epithelium), higher environmental ammonia tolerance, aerial vision and terrestrial locomotion using modified pectoral fins, among others (Kok et al., 1998; Pace & Gibb, 2009; Randall et al., 2015; Zander, 2011).

If aquaporins were relevant for the water-to-land transition of tetrapods, mudskippers may represent a remarkable case to investigate whether any changes (in a relatively similar fashion to those occurred in tetrapods) in aquaporins could have possibly occurred during their terrestrialisation process as well, either in the form of appearance of new paralogues and/or change/selection at the sequence level. This study aims to investigate the diversity and molecular evolution of mudskipper aquaporins in the context of water-to-land transitions of vertebrates taking advantage of the genomic data for four mudskipper species recently made available by You et al. (2014). Our specific objectives are summarised as follows: (1) to find and catalogue the aquaporin classes and paralogues that occur in mudskipper genomes using similarity searches with vertebrate comparable data, (2) to reconstruct the phylogenetic position and relationships of mudskipper aquaporins in the general context of vertebrates, and (3) to detect branch and site positive selection in the aquaporin classes present in mudskippers.

Materials and methods

Data mining and sequence identification

We retrieved the complete genome nucleotide sequences of four mudskipper species from GenBank: *Boleophthalmus pectinirostris* (Linnaeus, 1758), *Periophthalmodon schlosseri* (Pallas, 1780), *Periophthalmus magnuspinnatus* (Lee et al., 1995), and *Scartelaos histophorus* (Valenciennes, 1837). These four genomes, sequenced by You et al. (2014),

constitute the only genomic-scale data available for mudskippers thus far. At the time of genome data download (20 January 2017), their assembly and annotation were presented at the scaffold level, meaning that reconstructed sequence contigs were connected to get larger structures still unplaced in specific genomic regions or chromosomes, and that gene identification, delimitation and location remained limited. In order to fish the different aquaporin sequences that are present in the mudskipper genomes, we conducted BLAST searches (Altschul et al., 1990) with the BLASTX tool v2.2.28 run locally using each mudskipper genome (nucleotide level) as query, and custom-made aquaporin protein databases as target. Two separate aquaporin protein databases were constructed using makeblastdb. One of them included all protein sequences returned after conducting several searches of terms related to aquaporins for vertebrates in the GenBank protein database (Benson et al., 2013): "aquaporin vertebrates", "water channel vertebrates", "aquaglyceroporin vertebrates", "major intrinsic protein vertebrates", "AQP vertebrates", as well as variants of these (i.e. singular vs. plural terms, "vertebrates" vs. "Vertebrata"). The search results were inspected by eye in order to remove duplicates and obvious incorrect hits, for example those corresponding to other proteins (not aquaporins), but that contained the term "aquaporin" elsewhere in the record page (e.g. publication title). A second database was constructed including all protein sequences returned after searching similar terms in UniProt (Consortium, 2008), and was used as double-check. All sequence data for both aquaporin protein databases were downloaded on 31 January 2017. After depuration, the number of protein sequences included in the GenBank and UniProt vertebrate aquaporin databases was 1123 and 1232, respectively. BLASTX searches were conducted using an E-value threshold $>10^{-10}$ that was deemed appropriate relative to the genome-wide size of queries and relatively small size of the databases, and to obtain an output file with all potential mudskipper aquaporins including the ones with lower sequence similarity (due to e.g. accelerated evolution).

BLASTX results from both GenBank and Uniprot databases were compared and cross-validated by eye, then arranged by query scaffold number and unified to exclude redundant matches of the same query scaffold. All BLAST best matches/hits were inspected by eye, and those that did not show clear correspondence with proper aquaporin sequences were discarded. The list of best hits for the valid scaffolds for each mudskipper genome, after the aforementioned curation process, was used to assign a preliminary aquaporin identification for such scaffold based on the respective annotation of the target sequence (Table 1). Each sequence of the valid scaffolds matching aquaporins was extracted from their

corresponding mudskipper genomes to separate files and, in each case, candidate orthologous coding DNA sequences (CDS) of teleost fishes (*Danio rerio*, *Oreochromis niloticus* and *Salmo salar* among others) and mammals (*Homo sapiens* and *Mus musculus*) were aligned against them (as references) using MAFFT v7.310 (Katoh & Standley, 2013) with the L-INS-i algorithm. These alignments were used to verify initial aquaporin identification based on BLAST searches.

Due to computational constraints, long scaffolds had to be trimmed to remove the excess of sequence data surrounding the candidate gene containing the aquaporin match (as per the BLASTX coordinates) before conducting the alignments. Geneious Pro v9.1.8 (Kearse et al., 2012) was used for trimming the sequences as well as visualising the resulting alignments and locate start and stop codons according to the corresponding start/end positions in the reference sequences. The process of scaffold trimming and reference alignment was iteratively repeated until the aquaporin gene in the scaffold had been unambiguously identified and its sequence isolated. Using the reference alignments, we also identified and manually annotated intronic and exonic regions of each mudskipper aquaporin. We used Geneious to prepare separate files (available as "Electronic Supplementary Material") for entire gene, CDS and protein (deduced amino acid) sequences for each mudskipper aquaporin that could be identified in the genome data (Table 1).

You et al. (2014) also provided information of transcriptomes of gill and liver tissue of *B. pectinirostris* and *P. magnuspinnatus*. We retrieved raw reads of such RNAseq experiments from GenBank's SRA database, and performed *de novo* assemblies for each of the two species separately (pooling together all reads for tissue samples belonging to the same species). *De novo* assemblies were performed with Trinity r20140717 (Grabherr et al., 2011) using default settings with 60 Gb of RAM (--max_memory 60G). From the reconstructed transcriptomes, candidate protein-coding genes (CDS) were identified with TransDecoder 2.0 (Grabherr et al., 2011). The CDS sequences for each species were used as query in BLASTX searches against the custom-made aquaporin protein databases (as target), and using the same steps and E-value thresholds as mentioned above for the mudskippers' genomic data. Transcriptomic information was used to complement genomic searches, either for identification of overlooked paralogues or for verification of those already identified.

Table 1. List of the 55 aquaporins identified in the four mudskippers genomes using BLASTX. GenBank accession numbers correspond to the genome scaffold assembly provided by You et al. (2014).

Genus	Species	Aquaporin	Accession number
Boleophthalmus	*pectinirostris*	0a	KN524227
Boleophthalmus	*pectinirostris*	1a	KN525001
Boleophthalmus	*pectinirostris*	3	KN521806
Boleophthalmus	*pectinirostris*	4	KN521807
Boleophthalmus	*pectinirostris*	7[a]	KN525596
Boleophthalmus	*pectinirostris*	8ab	KN523248
Boleophthalmus	*pectinirostris*	9a	KN524979
Boleophthalmus	*pectinirostris*	10a	KN523925
Boleophthalmus	*pectinirostris*	11a	KN525406
Boleophthalmus	*pectinirostris*	11b	KN522455
Boleophthalmus	*pectinirostris*	12	KN522461
Periophthalmodon	*schlosseri*	0a	JACM01053832
Periophthalmodon	*schlosseri*	0b	KN475480
Periophthalmodon	*schlosseri*	1a	KN483096
Periophthalmodon	*schlosseri*	3	KN469198
Periophthalmodon	*schlosseri*	4	JACM01058456
Periophthalmodon	*schlosseri*	7[a]	KN471526
Periophthalmodon	*schlosseri*	8ab	JACM01008777/JACM01035013
Periophthalmodon	*schlosseri*	9a	KN482823
Periophthalmodon	*schlosseri*	9b	KN469413
Periophthalmodon	*schlosseri*	10a	KN473499
Periophthalmodon	*schlosseri*	10b	KN474915
Periophthalmodon	*schlosseri*	11a	KN469785
Periophthalmodon	*schlosseri*	11b	KN480662
Periophthalmodon	*schlosseri*	12	KN477140
Periophthalmus	*magnuspinnatus*	0a	KN467902
Periophthalmus	*magnuspinnatus*	0b	KN466703
Periophthalmus	*magnuspinnatus*	1a	JACL01049526
Periophthalmus	*magnuspinnatus*	1b	KN465554
Periophthalmus	*magnuspinnatus*	3	KN462285
Periophthalmus	*magnuspinnatus*	4	KN462075
Periophthalmus	*magnuspinnatus*	7[a]	KN468109
Periophthalmus	*magnuspinnatus*	8ab	KN464676
Periophthalmus	*magnuspinnatus*	8a1	c67494 (RNAseq data)
Periophthalmus	*magnuspinnatus*	9a	KN462368
Periophthalmus	*magnuspinnatus*	9b	KN467715

Periophthalmus	*magnuspinnatus*	10a	KN468565
Periophthalmus	*magnuspinnatus*	10b	KN460469
Periophthalmus	*magnuspinnatus*	11a	KN461794
Periophthalmus	*magnuspinnatus*	11b	KN460458
Periophthalmus	*magnuspinnatus*	12	KN467541
Scartelaos	*histophorus*	0a	JACN01025822
Scartelaos	*histophorus*	0b	KN491114
Scartelaos	*histophorus*	1a	KN502285
Scartelaos	*histophorus*	1b[a]	JACN01018430
Scartelaos	*histophorus*	3	KN514423
Scartelaos	*histophorus*	4	JACN01023074
Scartelaos	*histophorus*	8ab	KN499303
Scartelaos	*histophorus*	8a1	JACN01005585
Scartelaos	*histophorus*	9b[a]	JACN01079684
Scartelaos	*histophorus*	10a	JACN01051648
Scartelaos	*histophorus*	10b	JACN01031370
Scartelaos	*histophorus*	11a[a]	JACN01020100
Scartelaos	*histophorus*	11b[a]	JACN01076962/JACN01183528
Scartelaos	*histophorus*	12	KN493770

[a]Partial sequence

Phylogenetic analyses

Amino acid sequences of all identified mudskipper aquaporins were aligned against vertebrate homologous sequences retrieved from the GenBank protein database (Supplementary Table S1) in order to produce a comprehensive protein alignment combining all vertebrate aquaporins with special emphasis on the thorough representation of fish aquaporin sequences. The alignment was performed with MAFFT using the L-INS-i algorithm, and Geneious was used to visually inspect the alignment, trim ends and correct obvious mistakes. In order to reduce the amount of missing data, we used Geneious to strip the alignment columns containing at least 95% of gaps. ProtTest v3.4 (Abascal et al., 2005; Darriba et al., 2011) was used to determine the best-fit model of amino acid substitution using the Bayesian information criterion (BIC). The model selected was JTT (Jones et al., 1992) + Γ (Yang, 1994) + I (Reeves, 1992).

The resulting combined alignment (available in TreeBASE – accession S22429) was analysed using maximum likelihood (ML; Felsenstein, 1978) and Bayesian inference (BI; Huelsenbeck & Ronquist, 2001), which are currently the standard methods for molecular

phylogenetic inference (reviewed in San Mauro & Agorreta, 2010). Analyses were run on the CIPRES Science Gateway (Miller et al., 2010). ML analyses were performed with RAxML v8.2.10 (Stamatakis, 2014) using the rapid hill-climbing algorithm, computing 1000 distinct ML trees starting from 1000 distinct randomised maximum-parsimony starting trees, and with the Gamma model of rate heterogeneity for the bootstrapping phase. BI analyses were performed with MrBayes v3.2.6 (Ronquist et al., 2012) conducting two independent MCMC runs (with four chains each) for 20 million generations, sampling every 2000 generations, and discarding the first 25% of samples as burn-in. Adequate convergence of the BI runs was judged by plots of ln *L* scores versus generation time and low standard deviation of split frequencies, as well as convergence diagnostics (Estimated Sample Size [ESS], Potential Scale Reduction Factor [PSRF]), as implemented in MrBayes. Support for internal branches was evaluated by non-parametric bootstrapping with 1000 replicates (RAxML) and posterior probabilities (MrBayes).

Positive selection analyses

Given that alignment artefacts are one main source for spurious positive selection signal (Anisimova & Yang, 2007), we did not use the combined alignment for detecting positive selection. Instead, we conducted positive selection analyses using individual CDS datasets for each aquaporin class that is present in mudskipper representatives, namely AQP0, AQP1, AQP3, AQP4, AQP7, AQP8, AQP9, AQP10, AQP11 and AQP12 (Table 1). We aimed for similar taxon sampling in the individual CDS datasets as each aquaporin subset had in the combined protein dataset (Supplementary Table S1). For each of the 10 aquaporins, CDS datasets were constructed with the respective mudskipper sequences isolated here together with candidate orthologous vertebrate sequences retrieved from GenBank, UniProt and Ensembl following the list of species used in the combined protein dataset. Each individual aquaporin CDS dataset was aligned using TranslatorX (Abascal et al., 2010) using MAFFT to compute the protein alignments. Because positive selection analyses are very sensitive to missing data, we used Geneious to strip the CDS alignment columns containing 5% or more of gaps (hence guaranteeing over 95% sequence completeness in all cases).

Each aquaporin CDS alignment (available from the corresponding author upon request) was used to perform positive selection analyses with the CODEML module of the PAML program v4 (Yang, 2007). Detection of selective pressure at the protein level can be achieved using the omega value (ω), a parameter that measures the ratio between non-

synonymous and synonymous nucleotide substitutions (dN/dS) (Yang, 2008). The value of omega can indicate purifying (negative) selection when < 1, neutral selection when = 1, and positive (adaptive) selection when > 1. In the latter case (positive selection), the fixation of advantageous mutations can potentially lead to the development of evolutionary innovations due to the emergence of novel molecular functions (Nielsen & Yang, 1998). In CODEML analyses, we used phylogenetic topologies for each individual aquaporin CDS alignment used. These were inferred using RAxML with the GTR (Tavaré, 1986) + Γ model of nucleotide substitution and the same settings as described above, and branch lengths were corrected as substitutions per codon.

In order to detect positive selection in mudskipper aquaporins, we performed the branch-site Test 2 (null model MA vs. model MA; Yang et al., 2005; Zhang et al., 2005). This test assumes that only some sites on one or more selected branches (foreground branches) of the phylogeny might have undergone positive selection, and it is recommended over previously described branch-site Test 1 for empirical data (Yang, 2007; Zhang et al., 2005). A branch-site test is deemed most adequate and robust to answer the question whether episodic positive selection can be detected on mudskipper (foreground) branches with respect to other vertebrate (background) branches (Yang & Dos Reis, 2011). The test assumes that sites on the sequences can be classified into four classes, depending on the type of selection that they undergo: sites under purifying selection ($\omega < 1$), sites under neutral evolution ($\omega = 1$), sites under positive selection ($\omega > 1$) on the foreground but under purifying (2a) or neutral (2b) selection on the remaining branches (background). For both models, some parameters are fixed: seqtype = 1, model = 2, Nsites = 2, cleandata = 0. However, for the null MA model, the ω value is fixed to 1, and, for the MA model, the ω value is estimated, starting from the default 0.4. We calculated LRTs as before (null MA as null hypothesis, MA as alternative hypothesis) and computed *P* values using a mixed χ^2 distribution, which is obtained dividing the value of the χ^2 distribution with 1 degree of freedom by 2 (Yang, 2007).

For each aquaporin, we performed several branch-site tests selecting different foreground branches in each case. In particular, selection was tested for both the stem (ancestral) branch and entire clade (stem plus all descendant branches) of the following: (1) mudskippers, (2) tetrapods, (3) mudskippers plus tetrapods, (4) *Periophthalmus* + *Periophthalmodon* clade, (5) *Periophthalmus* + *Periophthalmodon* clade plus tetrapods. Additionally, we tested the stem branch leading to the tetrapod-exclusive aquaporins (AQP2, AQP5, AQP6). In the case of those aquaporin classes with two paralogues for the

mudskippers (all except AQP3, AQP4, AQP7 and AQP12; see below), additional tests considering each paralogue either separate or together were performed. Because several tests are conducted using the same tree topology, we performed multiple-test correction of the LRT *P* values, and calculated *Q* values (corrected *P* values) for a false discovery rate (FDR) using the qvalue package of R (R Development Core Team 2016). For those tests where the LRT is significant, we used the Bayes Empirical Bayes (BEB; Yang et al., 2005) implemented in PAML, which calculates the posterior probabilities for site classes and identify sites under positive selection.

Results

Catalogue of aquaporin genes in mudskippers

BLAST searches identified a total of 55 different aquaporin candidate genes that are present in the four mudskipper species included in the study, all of which can be assigned to aquaporin classes AQP0, AQP1, AQP3, AQP4, AQP7, AQP8, AQP9, AQP10, AQP11 and AQP12 (Table 1). For all four mudskipper species, we found at least one copy of each of the aforementioned aquaporin classes, but we were unable to find AQP7 in the genome data of *S. histophorus*. In mudskippers, six aquaporins (AQP0, AQP1, AQP8, AQP9, AQP10 and AQP11) present two paralogue copies, whereas the remaining classes (AQP3, AQP4, AQP7 and AQP12) only present a single copy (Table 1 and Figs. 1 and S1). For aquaporin classes with more than one paralogue for the same species (i.e. AQP0, AQP1, AQP3, AQP8, AQP9, AQP10, AQP11), we have named them by assigning copy letters (a, b, a1, a2, ab) to the corresponding inclusive clades following the GenBank annotations of *D. rerio*, *Cyprinus carpio* and *S. salar*. In particular, we identified two paralogue copies for mudskipper AQP0 (present in all species except in *B. pectinirostris*), AQP1 and AQP8 (present only in *P. magnuspinnatus* and *S. histophorus* in both cases), AQP9 (two paralogues in *P. magnuspinnatus* and *Pn. schlosseri*, and a single paralogue in *B. pectinirostris* and *S. histophorus* each), AQP10 (two paralogues in all species except in *B. pectinirostris*) and AQP11 (both paralogues present in all species). In contrast, we did not identify any of the aquaporin classes 13 to 16 in mudskippers, nor any new paralogue exclusive to the mudskipper clade (Table 1, Figs. 1 and S1).

The complete sequences of all mudskipper aquaporin genes were identified within single scaffolds, except the case of AQP8ab of *Pn. schlosseri* and AQP11b of *S. histophorus* that were found in two consecutive scaffolds (Table 1). For *B. pectinirostris*, *Pn. schlosseri*

and *P. magnuspinnatus*, it was possible to identify the complete aquaporin gene sequences (from start to stop codon) in all cases except in AQP7 that remained incomplete at both 5′ and 3′ ends in *Pn. schlosseri* and *P. magnuspinnatus*, and incomplete at only the 5′ end in *B. pectinirostris*. The complete sequence of AQP8a1 of *P. magnuspinnatus* was exclusively extracted from the RNAseq data. In the case of *S. hitophorus*, four sequences were extracted as partial fragments only: AQP1 paralogue b (3′ end incomplete), AQP9 paralogue b (3′ incomplete) and both AQP11 paralogues (a: 5′ end incomplete, b: 3′ incomplete; Table 1).

Phylogeny of vertebrate aquaporins

The final combined protein alignment includes a total of 313 terminals and 394 amino acid positions, of which 391 are variable and 383 are parsimony-informative. Both BI (arithmetic mean of runs: -ln *L* = 70,778.05) and ML (-ln *L* = 70,506.04) methods of phylogenetic inference yield highly similar topologies with differences only in branch lengths and levels of support (Figs. 1 and S1). Differences in phylogenetic relationships between the BI and ML trees only occur in branches with low support. In general, phylogenetic resolution is higher in the BI tree. The monophyly of each vertebrate aquaporin class generally receives high support, but phylogenetic relationships among them receive lower support in some cases (Figs. 1 and S1). Several monophyletic groups are unambiguously recovered and well supported by both BI and ML: the aquaglyceroporins (clade comprising AQP3, AQP7, AQP9, AQP10 and AQP13), the AQP8-type (including AQP16), the unorthodox aquaporins (AQP11 plus AQP12), and the tetrapod-exclusive aquaporins (clade comprising AQP2, AQP5 and AQP6). The clade comprising the water-selective classical aquaporins (AQP0, AQP1, AQP4, AQP14, AQP15 and tetrapod-exclusive aquaporins) received strong support from BI only.

Phylogenetic relationships among aquaglyceroporins are in general poorly resolved. Two agnathan aquaglyceroporins (AQP3 of *Lethenteron camtschaticum* and *Petromyzon marinus*) are recovered outside the AQP3 class clade, although support is low in this part of the tree. Likewise, the phylogenetic position of AQP13 (solely represented by the sequence of *Xenopus laevis*) remains uncertain within aquaglyceroporins. In contrast, the sequence of AQP16 of *X. laevis* is recovered nested within the well-supported AQP8 clade. Phylogenetic relationships among water-selective classical aquaporins are generally well supported, in particular by BI. Within these, the first split separates AQP14 followed by AQP4, and finally a clade including AQP15 of *X. laevis*, AQP1 and AQP0 as sister to the tetrapod-exclusive aquaporins. These latter appear to be derived from AQP2-like sequences present in the

coelacanth and lungfish, although the relevant branch receives support only for BI. Regarding the phylogenetic relationships among the mudskipper sequences, in general, a *Pn. schlosseri* + *P. magnuspinnatus* clade is recovered as sister to a *B. pectinirostris* + *S. histophorus* clade, except in the case of AQP0b and AQP10a, but support is low for alternative relationships in these two latter cases (Fig. S1). Our phylogenetic results also highlight that some of the vertebrate aquaporin sequences retrieved from GenBank had an incorrect annotation. They are indicated in Table S1, and are used in our phylogeny with the corrected annotation.

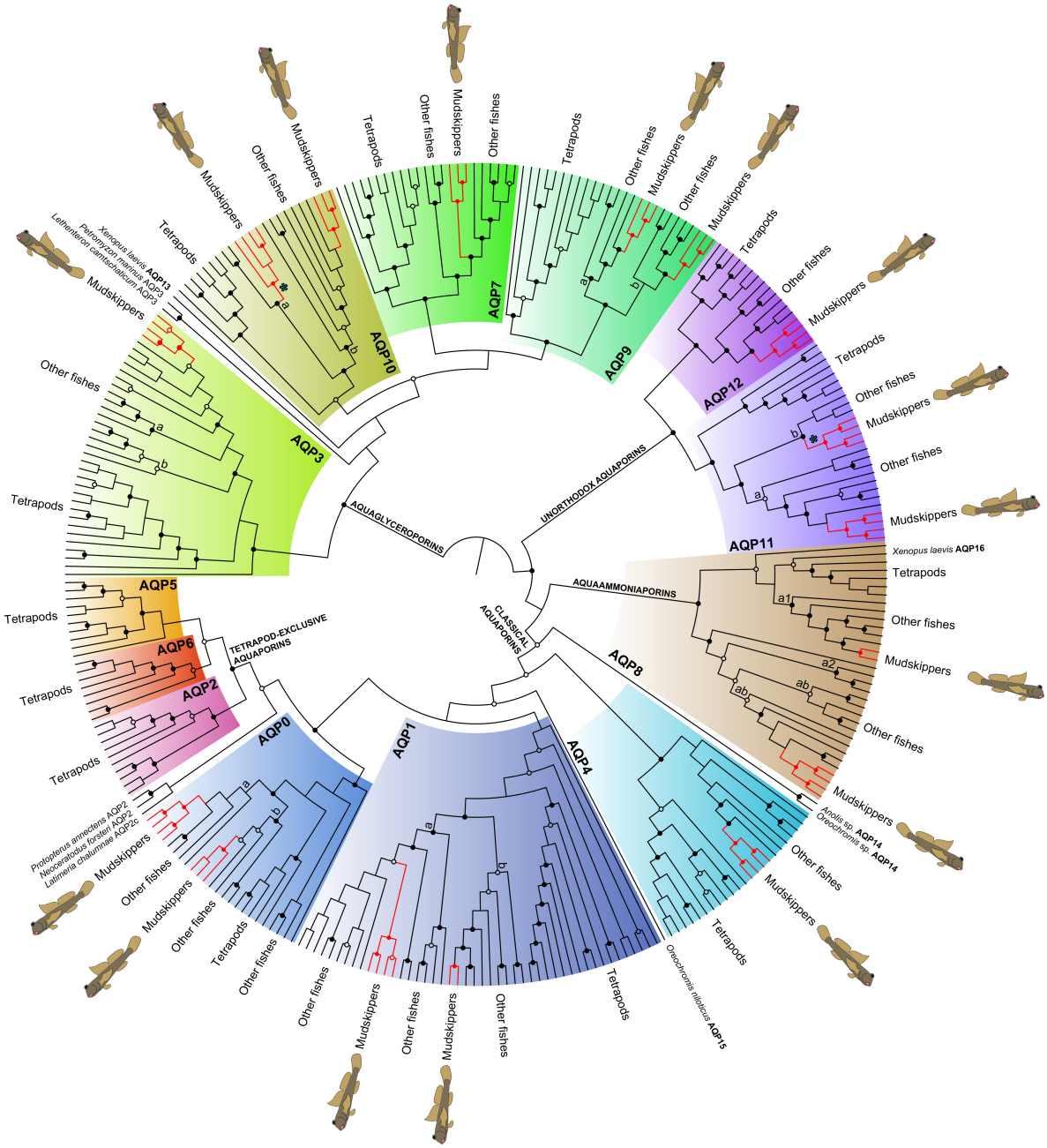


Figure 1. Bayesian inference cladogram of vertebrate aquaporins based on 394 aligned amino acid positions. Filled bullets on nodes denote support for both BI and ML methods of phylogenetic inference. Open bullets on nodes denote support for only one of the methods (either BI or ML).

Absence of bullet on a node denotes no support for any of the methods. Aquaporins classes are indicated with colour panels and class paralogues with letters (a, b, a1, a2, ab) on the corresponding branches. Mudskipper branches and associated node bullets are highlighted in red. Blue asterisks highlight the branches in which positive selection signatures are detected. A detailed, fully labelled, metric version of this phylogenetic tree is shown in Supplementary Fig. S1.

Detection of positive selection

Most of the branch-site tests (null model MA vs. model MA) conducted were not significant (*P* value of the LRT >> 0.05). However, we detected positive selection signatures (significant LRT and ω > 1 in foreground branches) in AQP10a and AQP11b of mudskippers (both entire clade and stem branch) and the *Periophthalmus* + *Periophthalmodon* clade (AQP10a) or stem (AQP11b) (Table 2). LRT *P* values that are significant (< 0.05) remain also significant after multiple-test correction (*Q* values < 0.05) in all cases. The BEB test identified specific alignment sites under positive selection in the mudskipper AQP10a clade (positions 15, 149, and 259), the mudskipper AQP10a stem (position 96), the *Periophthalmus* + *Periophthalmodon* AQP10a clade (positions 55 and 118), the mudskipper AQP11b clade (positions 38, 52, 112, 167, 183 and 205), and the mudskipper AQP11b stem (positions 72, 89, 182, 183, 190, 205 and 229). Positions under selection are referred to the AQP10 and AQP11 sequences (paralogue a) of *D. rerio*, respectively. Additionally, the branch-site test was also significant for the stem branch leading to the tetrapod-exclusive aquaporins (AQP2, AQP5, AQP6) (Table 2).

Table 2. Results of the branch-sites tests that are significant (*P* value of the LRT <0.05) and have ω>1 in foreground branches. All other tests not shown in the table are not significant and have ω<1 in foreground branches

Aquaporin	Foreground branch	LRT	*P* value[a]	*Q* value[b]	ω[c]	Prop.2a[d]	Prop.2b[e]
10a	Mudskipper clade	5.633	0.008	0.022	2.566	0.042	0.004
10a	Mudskipper clade stem	9.689	<0.001	0.005	22.609	0.033	0.003
10a	*Periophthalmus* +	4.631	0.016	0.026	6.124	0.027	0.003
11b	Mudskipper clade	3.571	0.029	0.049	1.690	0.108	0.027
11b	Mudskipper clade stem	16.189	<0.001	1.433 x	21.022	0.095	0.025
11b	*Periophthalmus* +	5.201	0.011	0.028	20.140	0.048	0.012
2, 5, 6	Tetrapod clade stem	3.199	0.037	0.037	7.440	0.059	0.001

[a]Uncorrected *P*-value of the LRT

[b]Multiple-test correction of the LRT *P*-value (False Discovery Rate)

[c]Omega (dN/dS) ratio of the foreground branch(es)

[d]Proportion of sites that are under positive selection (ω 2a > 1) on the foreground branch(es) and under purifying selection (ω < 1) on background branches

[e]Proportion of sites that are under positive selection (ω 2a > 1) on the foreground branch(es) and under neutral selection (ω = 1) on background branches

Discussion

Evolution of aquaporin classes in mudskippers

Our genomic screening returned gene matches for 10 different classes of aquaporins in mudskippers. These include the aquaglyceroporins, AQP8-type (also known as aquaammoniaporins), unorthodox aquaporins and water-selective classical aquaporins. In our analyses, only AQP7 is missing in the genome of *S. histophorus* and this may be related to the lower quality of its sequence data (You et al., 2014). The corresponding missing genes may be thus located in genomic regions outside the available sequenced scaffolds and/or in areas with low sequence quality (high proportion of N's) that might have impaired BLAST searches. We cannot completely rule out that other aquaporin genes or copies present in the four analysed mudskipper genomes have been missed by our BLAST searches. However, there are reasons that make us confident that the aquaporin candidate gene pool identified is accurate and complete given the genomic and transcriptomic data available. Besides the matches positively identified as mudskipper aquaporin genes, our BLAST searches returned hits similar to lens intrinsic membrane protein 2 (LIM2; Church & Wang, 1992), which are more distantly related to proper members of the aquaporin family. Thus, it is unlikely that additional aquaporin genes or paralogues, or even putative new/exclusive aquaporin classes of mudskippers could have been overlooked. Even the most dissimilar aquaporins have a certain degree of conservation scattered over their entire length.

Like other fishes, mudskippers lack genes for AQP2, AQP5 and AQP6 (Table 1 and Fig. 1), which are exclusive to the tetrapod lineage (Cerdà & Finn, 2010; Finn & Cerdà, 2011; Tingaud-Sequeira et al., 2010). These three aquaporin classes were likely important for the water-to-land transition of tetrapod ancestors (Finn et al., 2014; Konno et al., 2010; Saitoh et al., 2014; Suzuki et al., 2007). However, no other instance of similar emergence of novel classes (as occurred in tetrapods) has been documented for other fishes with terrestrial (amphibious) lifestyle thus far, and this is also the case of mudskippers, which lack

mudskipper-exclusive paralogues (as per this study). You et al. (2014) investigated and discussed about other genes implicated in the terrestrialisation of mudskippers, such as ammonia excretion in gills, vision or immune system, but aquaporins were not studied. On the other hand, we did not identify any of the 13 to 16 aquaporin paralogues in mudskippers either. Finn et al. (2014) reported AQP14 and AQP15 in some teleost representatives, but their ubiquity in other fish groups (including gobioids) remains to be further confirmed. Our phylogenetic tree (Fig. 1) shows that some fishes possess two or more paralogues for aquaporin classes AQP0, AQP1, AQP3, AQP8, AQP9, AQP10 and AQP11, most of them due to an extra whole genome duplication (WGD) on the stem branch of teleosts combined with tandem duplication events (Cerdà & Finn, 2010; Crow et al., 2006; Finn et al., 2014; Finn & Kristoffersen, 2007; Taylor et al., 2001; Tingaud-Sequeira et al., 2010). In the case of mudskippers, we identified two paralogue copies in at least two of the studied species (generally in three or four) in the aforementioned aquaporin classes with more than one paralogue documented, except in the case of AQP3. In this latter, only a single paralogue could be identified in each of the four mudskipper species studied (Table 1 and Fig. 1). The lack of an AQP3 paralogue in mudskippers could be due to failure in identification (but see above) or to gene loss. Indeed, the loss of aquaporin genes derived from WGD has been documented in many species of acanthomorph fishes, and incidentally such is the case of the loss of one of the AQP3 copies in *D. rerio* (Cerdà & Finn, 2010; Tingaud-Sequeira et al., 2010).

As mentioned above, the gene phylogeny of vertebrate aquaporins (Fig. 1) recovers four main groups of aquaporins (aquaglyceroporins, AQP8-type/aquaammoniaporins, unorthodox aquaporins and water-selective classical aquaporins), which is in agreement with earlier phylogenetic studies on aquaporins (Abascal et al., 2014; Cerdà & Finn, 2010; Finn et al., 2014; Finn & Cerdà, 2011, 2015; Park & Saier, 1996; Zardoya, 2005). The separation between aquaglyceroporins and all other aquaporins occurred early in the evolution of cell (Abascal et al., 2014; Heymann & Engel, 1999). The aquaglyceroporins clade contains five well-supported classes (AQP3, AQP7, AQP9, AQP10 and AQP13), but the phylogenetic relationships among them are not resolved, as in Abascal et al. (2014). In contrast, Finn et al. (2014) were able to find strong support for the sister group relationship between AQP7 and the AQP3 + AQP9 clade, and they also indicated the existence of a new paralog, AQP13. AQP8-types (including AQP16) are related with ammonia transport (Saparov et al., 2007), and AQP11 and AQP12 were named as unorthodox aquaporins because their respective functions remain the least known (Jahn et al., 2004; Wu & Beitz, 2007). The classical

aquaporins function is mainly related to water interchange, although other molecules (e.g., CO_2) may be permeable as well (Finn & Cerdà, 2015). Phylogenetic relationships among classical aquaporins are relatively well resolved (strongly supported by BI), and in agreement with the findings of previous studies (Abascal et al., 2014; Finn et al., 2014). AQP14 is the first to split from the tree, followed by AQP4, and the clade comprising AQP15 (whose phylogenetic position could not be confidently resolved by Finn et al. (2014), AQP1 and AQP0 as the sister-group of the tetrapod-exclusive aquaporins (AQP2-like first, then AQP5 and AQP6). An ancestral duplication of AQP0 led to the emergence of these three new classes unique to tetrapods (Abascal et al., 2014; Finn et al., 2014; Laforenza et al., 2016), being some coelacanth and lungfish AQP2-like paralogues the precursors of them (Finn et al., 2014). In the gene phylogeny, the position of agnathan (*L. camtschaticum* and *P. marinus*) AQP3 is not recovered as expected (Figs. 1 and S1). Assuming that class annotation reported in GenBank is correct for them, their spurious phylogenetic placement is likely related to the fact that their sequences are more plesiomorphic compared to the other vertebrates included in the tree, thus rendering phylogenetic inference more difficult for them.

Regarding the phylogenetic relationships among the four mudskipper species used in the study, most aquaporin classes and paralogues recover a *Pn. schlosseri* + *P. magnuspinnatus* clade (more terrestrial forms) as the sister group of a *B. pectinirostris* + *S. histophorus* clade (more aquatic forms). The only exceptions are AQP0b and AQP10a, where alternative branch arrangements are recovered, but all of them without support. The sister group relationship between the more terrestrial (*Pn. schlosseri* + *P. magnuspinnatus*) and more aquatic (*B. pectinirostris* + *S. histophorus*) mudskippers is in agreement with the results of recent molecular-based studies (Agorreta et al., 2013; Polgar et al., 2017; You et al., 2014), but contrasts with the earlier morphology-based cladogram of Murdy (1989) that did not recover *B. pectinirostris* and *S. histophorus* as sister groups. In any case, both molecular- and morphology-based studies agree that more terrestrial forms of mudskippers (genera *Periophthalmodon* and *Periophthalmus*) share a common ancestry, although even within these two genera there are varying degrees of adaptation to terrestriality (Ishimatsu & Gonzales, 2011).

Positive selection in mudskipper aquaporins and the water-to-land transition

The main working hypothesis of this study is that, if aquaporins were relevant for the water-to-land transition of tetrapods (Finn et al., 2014), there could be changes (in a similar

fashion) in aquaporins during the terrestrialisation process of mudskippers, either in the form of new paralogues or sequence change. Our gene search and phylogenetic analyses have shown that no new paralogues have appeared in the mudskipper lineage (e.g. by tandem duplication of existing copies) that do not occur in other vertebrate groups (Fig. 1). In contrast, we were able to detect that AQP10a and AQP11b might have been under positive selection in mudskippers, as well as in the *Periophthalmus* + *Periophthalmodon* clade (Table 2), implying different evolutionary times. This latter clade contains the two mudskipper species (of the four analysed here) that display a higher degree of terrestrial adaptation: greater portion of the daily cycle out of the water, terrestrial activity and locomotion (*P. magnuspinnatus* can even climb on mangrove trees and rocks), feeding on terrestrial prey, etc. (Ishimatsu & Gonzales, 2011; Polgar et al., 2017; You et al., 2014).

The evidence of positive selection in one of the paralogue copies of aquaporin classes 10 and 11 (AQP10a and AQP11b, respectively) suggests possible alteration in their respective molecular function caused by changes at specific protein sequence positions (i.e. those identified under positive selection by the BEB tests). The positions identified under positive selection in the mudskipper AQP10a clade and stem are not located in any of the transmembrane helices (H1–H5) or loop regions, as per comparison with other aquaporin protein sequences (following Abascal et al., 2014). However, one position under positive selection in AQP10a clade is located in transmembrane helix H6, and the position identified under positive selection in the mudskipper stem (110 in our protein alignment) is located by the 5′ end H3 and close to loop B and the NPA motif contained within (Fig. 2). In contrast, both positions identified under positive selection in the *Periophthalmus* + *Periophthalmodon* clade are located in transmembrane helix 2 (H2) and H3, respectively. The position located in H2 is only three sites away from an ar/R selectivity filter site, which delimits the narrowest part of the molecule pore and thus dictates substrate selectivity (Fu 2000). Since glycerol is a larger molecule than water, modification in the ar/R region of aquaglyceroporins with respect to other aquaporins (classical, AQP8-type, unorthodox) takes place to enlarge pore diameter and achieve glycerol transport (Laforenza et al., 2016). On the other hand, the position located in H3 is next to a P1 site, one of the five amino acid residues (named P1–P5) differentially conserved in aquaglyceroporins versus other aquaporins, and that is potentially related to glycerol selectivity (Froger et al., 1998). In the case of AQP11b, all positions under positive selection correspond to the mudskipper clade and stem (Table 2), and they are located in either helices or loops (Fig. 2). Two positions are located one and two sites away from two ar/R selectivity filter sites, respectively located in loop E and helix H2. More remarkably,

another position is found at the NPA motif of loop B itself, which is related with pore formation and thus solute specificity (Cheng et al., 1997; Heymann & Engel, 2000; Sui et al., 2001). Given that this motif is highly conserved in vertebrates, and that it is key for the function of all aquaporins, its modification in mudskippers likely had substantial impact on the protein function, suggesting adaptation.

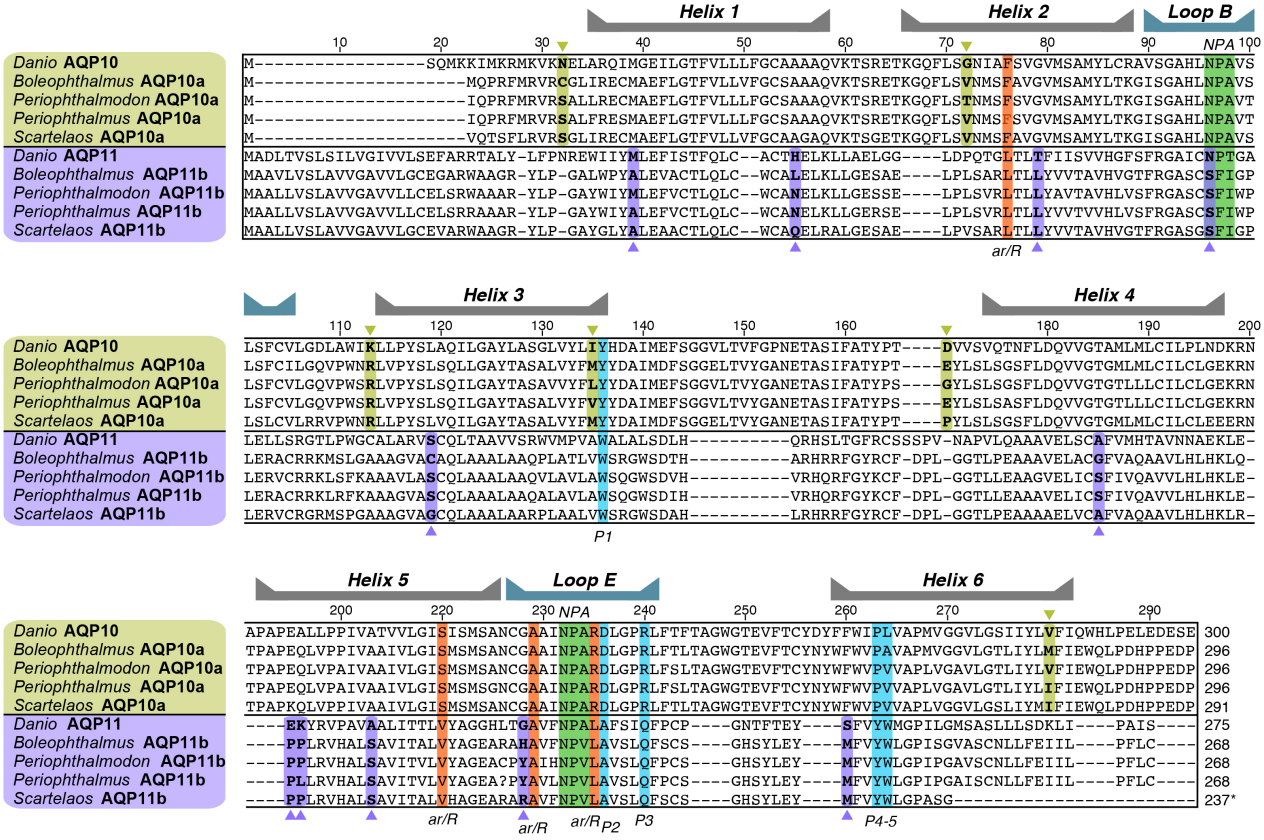


Figure 2. Sequence alignment and structural annotation of AQP10a and AQP11b showing positively selected positions in mudskippers (indicated with triangles and olive green or purple bars, respectively). The sequences of the corresponding paralogues of *D. rerio* are included as reference. Upper bars denote helices and loops. Boxes denote NPA motifs (green), the ar/R selectivity filter (orange), and sites reported to confer glycerol selectivity (P1-P5) (blue). Numbers at the end of the alignment denote the total length of each sequence (asterisk indicates partial sequence).

Understanding the precise effects on molecular structure and function of the selective changes identified at all the aforementioned protein sequence positions requires further insights from folding simulations and particularly functional studies. Nevertheless, it is reasonably plausible that selective changes that have occurred in AQP10a and AQP11b of mudskippers (magnified in more terrestrialised forms, *Periophthalmus* and

Periophthalmodon) could possibly be associated to their adaptation to an amphibious lifestyle. Aquaporins appear to have been crucial in the water-to-land transition of tetrapods (Finn et al., 2014), and have also played a role in the land-to-water transition of cetaceans (São Pedro et al., 2015). There is also evidence that aquaporins were likely important in the terrestrial evolution of green plants (Anderberg et al., 2011). It is therefore fairly reasonable to relate our positive selection evidence in mudskipper aquaporins to their terrestrial adaptation. However, without experimental validation, which is beyond the scope of this study, we cannot prove causality or otherwise fully rule out that the selective changes could be related to other processes/adaptations different from terrestrialisation or lifestyle traits in general. Notwithstanding, further support of the terrestrialisation interpretation is found in the fact that we detected signals of positive selection in aquaporins not only in the mudskipper clade and ancestral/stem branch, but also additional signals in the clade containing the most terrestrialised forms (*Periophthalmus* and *Periophthalmodon*) within mudskippers. This provides certain additional support that selection is taking place in aquaporins as terrestrialisation becomes more consolidated, and hence supporting their role in the water-to-land transition of mudskippers. Although other explanations are also possible, the interpretation in the context of the transition to an amphibious lifestyle is certainly plausible.

Interestingly, positive selection appears to be acting on aquaporin classes that have retained two paralogues from the WGD in the teleost ancestor. Positive selection in branches and sites could be evidence of neofunctionalisation (Force et al., 1999; Innan, 2009; Lynch & Force, 2000), which is a common process after a gene duplication event (Cerdà & Finn, 2010; Dehal & Boore, 2005; Escriva et al., 2006; Tingaud-Sequeira et al., 2010). Both classes of aquaporins under selection (AQP10 and AQP11) present two copies in the mudskipper genomes (a and b), being plausible that one of the copies could have evolved acquiring a new function, whereas the other maintains its ancestral function after the gene duplication. Similar cases of new function acquisition in one of the duplicated copies after a gene duplication event are well documented in many groups of organisms (e.g. Han et al., 2009; Saad et al., 2018) including cases of neofunctionalisation in teleost fishes after the WGD (Moriyama et al., 2016; Ogino et al., 2016). As in AQP10a and AQP11b of mudskippers, we also detected positive selection signatures for the stem branch leading to the tetrapod-exclusive aquaporins (Table 2). This provides evidence that, following the ancestral duplication that separated new paralogues (AQP2-like first, then AQP5 and AQP6) from the AQP0 class, positive selection favoured the acquisition of novel functions that facilitated the colonisation of terrestrial environments. This is obviously an independent case from mudskippers, but illustrates that

positive selection signatures are independently found in branches of the vertebrate aquaporin phylogenetic tree that signify neofunctionalisations (in this case likely associated to water-to-land transitions). Other branch-site tests performed for the remaining aquaporin classes (those not exclusive to tetrapods) having the tetrapod branch as foreground branch do not return signatures of positive selection. This suggests that, even in the case of tetrapod ancestors, adaptive selection could have occurred in a single branch, leading to the emergence of evolutionary innovations (AQP2, AQP5, AQP6) essential in the water-to-land transition (Finn et al., 2014). In this context, mudskippers could be seen as a similar case of that of early amphibious tetrapods before they evolved new aquaporin classes and transitioned to an exclusively terrestrial lifestyle.

AQP10 expression is high in fish gill and intestine organs because it is important for osmotic regulation in fishes (Finn & Cerdà, 2011; Madsen et al., 2015; Wright & Turko, 2016). In seawater fishes, water is lost in the gills due to high ionic concentration in the outside, but it is recovered in the intestine by drinking. In contrast, freshwater fishes gain water through the gills and the excess is eliminated with the urine (Madsen et al., 2015). Mudskippers are euryhaline fishes, which means that they can tolerate great differences in salt concentration (Ishimatsu & Gonzales, 2011; Polgar et al., 2017). It has been documented that AQP3 is important in water transport in the gills of euryhaline fishes, being down-regulated in gills when they acclimatise to seawater in order to retain higher amounts of water (Madsen et al., 2015; Martinez et al., 2005; Tipsmark et al., 2010). Given its aquaglyceroporin nature, AQP10 could play a similar role as AQP3 in the fish gills, but no specific study has confirmed this thus far. In some species, AQP10 appears to be upregulated in the intestine (Madsen et al., 2015; Martinez et al., 2005; Tipsmark et al., 2010), perhaps to recover water after active drinking. However, gills and intestine ion interchange seems to be low in amphibious fishes when they are out of the water, and it is the skin and, to lesser extent, the kidney that perform such function (Wright & Turko, 2016).

Research on *Pn. schlosseri* has shown that, in emerged mudskippers, gills are in charge of ammonia (NH_4+) excretion (Chew et al., 2007; Randall et al., 2015), which is performed with low energy and water waste. Given that AQP10 is able to transport urea (Finn & Cerdà, 2015), it certainly plays an important role in nitrogen metabolism. However, it is still unclear whether vertebrate AQP10 is permeable to ammonia (or whether this function is exclusive to other aquaporin classes, such as AQP8 type), but some studies suggest this possibility (Tingaud-Sequeira et al., 2010; Tipsmark et al., 2010). We hypothesise that changes occurring at the positively selected sites identified in mudskipper AQP10 might have

altered its normal function in a way that it becomes more permeable to ammonia instead of/in addition to urea, which is more relevant for mudskippers in terms of nitrogen metabolism and excretion (J. B. Graham, 1997; Ishimatsu & Gonzales, 2011). Apart from gills, some studies have reported that AQP10 of teleost fishes is also expressed in both skin and kidney (Finn & Cerdà, 2011; Madsen et al., 2015), which are the most important organs for ionic interchange in amphibious fishes. Mudskipper skin is modified in order to minimise desiccation (Dabruzzi et al., 2011) and it contains a special type of cells (swollen mid cells) that act as both water-loss barrier and water warehouse (Zhang et al., 2003). AQP10 could play a role in the regulation of such water storage, but this remains in need of further research. On the other hand, some studies have indicated that kidney expression of AQP10 is low compared to other aquaporins (such as AQP1), but, like AQP3, it appears upregulated in seawater conditions in *S. salar* (Madsen et al., 2015; Martinez et al., 2005; Tipsmark et al., 2010). This may be related to water recovery in a dehydrating medium, and it is reasonable that similar regulation may occur in mudskippers kidney when they are out of the water.

In the case of AQP11, its specific functions and molecular properties remain largely unknown, although, like the other aquaporin classes, a water transporter function has been documented, as well as permeability to glycerol (Madeira et al., 2014; Yakata et al., 2007, 2011). Besides, the expression profile of AQP11 is not well defined yet, but there is evidence that it could play an important role in the renal proximal tubule (Morishita et al., 2005). Due to the general lack of information about AQP11, we are still unable to provide clearer insights that relate the positive selection signatures detected in AQP11b of mudskipper sequences and the specific biochemical and physiological changes that could have possibly facilitated the water-to-land transition (e.g. implication in nitrogen metabolism and excretion, water retention in a dehydrating medium).

Because aquaporins are important for osmoregulation, the different lifestyles of the non-mudskipper fish species included in the study (whether freshwater, brackish or marine) could potentially affect the comparisons. We therefore conducted additional tests to assess whether the two aquaporins with signatures of positive selection still show the same results if only marine or only brackish/euryhaline fish outgroup species are included for the comparisons. Due to limitations of data availability, it was not possible to compare mudskipper AQP10a with any marine or brackish/euryhaline fish counterpart (only freshwater *C. carpio* and *D. rerio* sequences were available). However, it was possible to conduct the test for AQP11b since at least one outgroup fish species available was marine (*Larimichthys crocea*) and two were brackish/euryhaline (*Fundulus heteroclitus* and *Oryzias dancena*)

(Supplementary Table S1). Branch-site tests were then repeated as described above for the AQP11b stem branch and entire clade of mudskippers, as well as the *Periophthalmus* + *Periophthalmodon* clade stem. Results were very similar to those using all fish outgroup species (Supplementary Table S2), which highlights the general robustness of our positive selection results, and that they hold even when the comparisons are made against only marine or only brackish/euryhaline fish species. On the other hand, our dataset contains representatives of two other fish species for which certain amphibious behaviour (either in adults or larvae forms) has been documented, namely *Anguilla anguilla* and *F. heteroclitus* (J. B. Graham, 1997; Wright & Turko, 2016). Such amphibious behaviour is not as noteworthy as in mudskippers, but, in order to assess whether positive selection may be occurring in any aquaporin of those two species, we also performed branch-site tests for them. None of these branch-site tests were significant (*P* value of the LRT >> 0.05), hence we can discard that aquaporins are under positive selection in these two species.

In this study, we have provided a comprehensive molecular evolution and phylogenetic assessment of the aquaporin complement in mudskippers. We have focused on copy number and selective changes in aquaporin sequences that could potentially be associated with the physiology of aerial acclimation during the adaptation of mudskippers to an amphibious lifestyle. Although our positive selection results highlight the possible role of aquaporins in mudskipper terrestrial adaptation, proving causation may require further experimental research beyond the scope of this study. There is also the possibility that additional changes in mudskipper aquaporins may be related to regulatory elements rather than to the gene sequence itself, but this is something that remains little studied in other vertebrate groups as well (see e.g. Müller et al., 2006). Furthermore, protein-protein interaction analyses (of aquaporins with other nuclear receptors or cell membrane elements) may also reveal changes during mudskipper adaptation to an amphibious lifestyle. Additionally, greater representation of other gobioids closely related to mudskippers, such as less- or entirely non-amphibious oxudercines (e.g. genera *Apocryptes*, *Apocryptodon*, *Oxuderces*, *Parapocryptes*, *Pseudapocryptes* and *Zappa*) as well as amblyopines, could provide finer and more precise genomic and comparative insights. The amblyopines are paraphyletic with respect to the oxudercines, and there are both completely aquatic (the majority) and amphibious species that appear admixtured in the phylogeny (Agorreta et al., 2013). Testing whether the aquaporins of non-amphibiuos oxudercine and amblyopine species bear signatures of positive selection as well could expand the insights on how the adaptation

to amphibious lifestyles took place. All these open new exciting directions for further investigation.

Concluding remarks

Mudskippers are amphibious fishes that evolved terrestrial lifestyle independently from tetrapods, and they constitute an interesting case for investigation of water-to-land transitions. Like tetrapod ancestors, mudskippers also reached terrestrialisation due to specific modifications in locomotion, aerial respiration, and osmoregulation. For the latter, aquaporins play an important role as they are implicated in water maintenance and ionic interchange. The appearance of three new aquaporin paralogues exclusive to tetrapods appears to have facilitated their colonisation of terrestrial environments. We have found 55 aquaporin genes in four mudskipper species, all belonging to 10 known classes. AQP10a and AQP11b present positive selection signatures in mudskippers (both entire clade and the clade containing the most terrestrial species), as in the stem branch leading to the tetrapod-exclusive aquaporins. Positive selected sites in AQP10a and AQP11b are located in relatively proximity to (and one even affects directly) protein motifs involved in pore formation and substrate selectivity, and might have allowed permeability to new solutes, being a case of neofunctionalisation. Because AQP10a is important for osmotic regulation (and probably AQP11b as well), it is reasonably possible that the selective changes identified here could play a role in the adaptation of mudskippers to an amphibious lifestyle.

SUPPLEMENTARY MATERIAL

Table S1 List of vertebrate aquaporins and species (both amino acid and nucleotide sequences) used in this study. For each, accession numbers for GenBank and/or UniProt, or Ensembl are shown

Genus	Species	Aquaporin	AA accession		Nu accession	
			GenBank/Uniprot	Ensembl	GenBank	Ensembl
Bos	*taurus*	0	P06624		K02818	
Homo	*sapiens*	0	P30301		U36308	
Mus	*musculus*	0	P51180		U27502	
Gallus	*gallus*	0	P28238		AY078179	
Pseudopodoces	*humilis*	0	XP005532018		XM_005531961	
Neoceratodus	*forsteri*	0	C7G3K6		AB513619	
Protopterus	*annectens*	0	C7G3K5		AB513618	
Latimeria	*chalumnae*	0		ENSLACP00000021649		ENSLACP00000021649
Boleophthalmus	*pectinirostris*	0a			KN524227	
Cyprinus	*carpio*	0a1	A0A0K2RVW2		LC069001	
Cyprinus	*carpio*	0b1	A0A0K2RVS7		LC068999	
Danio	*rerio*	0a	Q6DEI6		BC077129	
Danio	*rerio*	0b	Q4ZJI3		DQ003080	
Fundulus[a]	*heteroclitus*	0	A0A147A698		GCES01012260	
Periophthalmodon	*schlosseri*	0a			JACM01053832	
Periophthalmodon	*schlosseri*	0b			KN475480	
Periophthalmus	*magnuspinnatus*	0a			KN467902	
Periophthalmus	*magnuspinnatus*	0b			KN466703	
Salmo	*salar*	0a1	A0A185IIP7		KM823661	

Salmo	*salar*	0b1	A0A0K0LTP0		KM677199	
Scartelaos	*histophorus*	0a			JACN01025822	
Scartelaos	*histophorus*	0b			KN491114	
Sparus	*aurata*	0a	T1T0Z6		KC589385	
Scyliorhinus	*canicula*	0	A0A0A7R6W0		KJ784515	
Homo	*sapiens*	1	P29972		AF020620	
Mus	*musculus*	1	Q02013		L02914	
Myotis	*brandtii*	1	S7MQH8		XM_005862944	
Pontoporia	*blainvillei*	1	A0A0H4LZ70		KM888076	
Sus	*scrofa*	1	Q6PQZ1		AY585335	
Tupaia	*chinensis*	1	L9KZ09		XM_006146241	
Gallus	*gallus*	1	Q2MCJ7		AM183252	
Nestor	*notabilis*	1	A0A091SYK2		KK938032	
Alligator	*mississippiensis*	1	A0A151PD70		XM_006262096	
Crotalus	*adamanteus*	1	J3S403		JU173916	
Hyla	*chrysoscelis*	1	Q2ESH3		DQ364243	
Xenopus	*tropicalis*	1	Q6DJ01		BC075384	
Protopterus	*annectens*	1	D0FZC4		KX494980	
Latimeria	*chalumnae*	1		ENSLACP00000008076		ENSLACT00000008142
Anguilla	*anguilla*	1a	Q400J9		AJ564420	
Anguilla	*anguilla*	1b	Q400J8		EF011738	
Aphyosemion	*striatum*	1a	A0A1A7YZR2		HADW01011359	
Boleophthalmus	*pectinirostris*	1a			KN525001	
Cyprinus	*carpio*	1a1	A0A0K2RVW0		LC069005	

Cyprinus	*carpio*	1a2	A0A0K2RVW6	LC069006
Danio	*rerio*	1a	Q6NZ72	BC066289
Danio	*rerio*	1b	ACA29537	EU327345
Dicentrarchus	*labrax*	1	Q06AD4	DQ924529
Fundulus	*heteroclitus*	1	C1ITJ6	EU780153
Heteropneustes	*fossilis*	1b	E7CH56	HM051492
Hippoglossus	*hippoglossus*	1aa	G0XMT4	HQ185294
Hippoglossus	*hippoglossus*	1ab	G0XMT5	HQ185295
Larimichthys	*crocea*	1	A0A0F8AIQ3	
Larimichthys	*crocea*	1	A0A0F8B377	XM_010729217
Nothobranchius	*korthausae*	1a	A0A1A8GIQ7	HAEB01017577
Oryzias	*dancena*	1	A0A023GQK1	AB759556
Osmerus	*mordax*	1	C1BJE6	BT074725
Periophthalmodon	*schlosseri*	1a		KN483096
Periophthalmus	*magnuspinnatus*	1a		JACL01049526
Periophthalmus	*magnuspinnatus*	1b		KN465554
Salmo	*salar*	1	B5X1X5	BT045044
Salmo	*salar*	1	B5X9V7	BT047826
ıos	*histophorus*	1a		KN502285
Scartelaos	*histophorus*	1b		JACN01018430
Scleropages	*formosus*	1	A0A0P7YCC1	JARO02007517
Sparus	*aurata*	1	Q32ZE5	AY626939
Sparus	*aurata*	1	Q32ZE6	AY626938
Takifugu	*obscurus*	1	G8Z2N1	GQ325617

Petromyzon	*marinus*	1	A0A0A7R6W6		KJ784520	
Callorhinchus	*milii*	1	V9L222		JW872452	
Scyliorhinus	*canicula*	1	A0A0A7RCI2		KJ784516	
Homo	*sapiens*	2	P41181		Z29491	
Mus	*musculus*	2	Q3UQD4		AK142567	
Myotis	*davidii*	2	L5MD57		KB101600	
Pontoporia	*blainvillei*	2	A0A0H4LW41		KM888077	
Sus	*scrofa*	2	B2MW86		EU636238	
Amazona	*aestiva*	2	A0A0Q3LTW3		LMAW01003137	
Pelecanus	*crispus*	2	A0A091SYG3		KK492024	
Alligator	*mississippiensis*	2	A0A151LYE2		AKHW03007029	
Chelonia	*mydas*	2	M7BDD9		KB582107	
Hyla	*chrysoscelis*	2	Q2ESH2		DQ364244	
Neoceratodus[b]	*forsteri*	2	BAH98063		AB513620	
Protopterus[b]	*annectens*	2	BAH86607		AB474277	
Latimeria[b]	*chalumnae*	2c		ENSLACT00000020645		ENSLACT00000020645
Homo	*sapiens*	3	Q92482		AB001325	
Mus	*musculus*	3	Q8R2N1		AF104416	
Myotis	*davidii*	3	L5M7H1		KB103229	
Pontoporia	*blainvillei*	3	A0A0H4MC27		KM888078	
Sus	*scrofa*	3	F1DQZ7		JARO02004730	
Tupaia	*chinensis*	3	L9L4Z4		KB320502	
Nestor	*notabilis*	3	A0A091RZT2		KK937313	
Pelecanus	*crispus*	3	A0A091SZ71		KK490429	

Alligator	*mississippiensis*	3	A0A151M1R0	AKHW03006811
Chelonia	*mydas*	3	M7BWM2	KB540241
Crotalus	*adamanteus*	3	J3S872	JU173917
Hyla	*chrysoscelis*	3	Q2ESH1	DQ364245
Xenopus	*laevis*	3	Q9YH65	AJ131847
Protopterus	*annectens*	3	D0FZC5	KX494981
Latimeria	*chalumnae*	3	H3A425	AQP3201
Anguilla	*anguilla*	3	Q7ZTH2	AJ319533
Boleophthalmus	*pectinirostris*	3		KN521806
Cyprinus	*carpio*	3a1	A0A0K2RWD5	LC069008
Cyprinus	*carpio*	3b1	A0A0K2RVW4	LC069010
Danio	*rerio*	3	Q803U6	BC044188
Danio	*rerio*	3b	D3TI80	EU341832
Dicentrarchus	*labrax*	3	Q0ZB19	DQ647191
Fundulus	*heteroclitus*	3	A0A147ACI3	EU780154
Ictalurus	*punctatus*	3	W5U622	JT405558
Ictalurus	*punctatus*	3	W5U716	JT406481
Lethenteron	*camtschaticum*	3	A0A0M4NHA4	KR054618
Nothobranchius	*furzeri*	3	A0A1A8V542	HADY01004903
Oreochromis	*niloticus*	3	A0A023ZXY1	KJ496130
Oryzias	*dancena*	3	A0A023GQJ6	AB759557
Periophthalmodon	*schlosseri*	3		KN469198
Periophthalmus	*magnuspinnatus*	3		KN462285
Scartelaos	*histophorus*	3		KN514423

Scleropages	*formosus*	3	A0A0P7UEY8	JARO02015848
Scleropages	*formosus*	3	A0A0P7WV94	JARO02004730
Sparus	*aurata*	3	T1T009	KC788197
Takifugu	*obscurus*	3	G8Z2N2	GQ325618
Tribolodon	*hakonensis*	3	Q8UWA8	AB055465
Petromyzon	*marinus*	3L1	A0A0A7R7B0	KJ784518
Homo	*sapiens*	4	P55087	D63412
Mus	*musculus*	4	P55088	U48398
Myotis	*davidii*	4	L5LSC4	KB108055
Pontoporia	*blainvillei*	4	A0A0H4M103	KM888079
Sus	*scrofa*	4	A8V978	EU165525
Tupaia	*chinensis*	4	L9KGP3	KB320866
Amazona	*aestiva*	4	A0A0Q3LYJ8	LMAW01002923
Gallus	*gallus*	4	Q65YQ3	AB190358
Alligator	*mississippiensis*	4	A0A151PGY7	AKHW03000230
Chelonia	*mydas*	4	M7APF1	KB579663
Ophiophagus	*hannah*	4	V8NT92	KK948819
Latimeria	*chalumnae*	4	H3ADX5	XM_006005561
Boleophthalmus	*pectinirostris*	4		KN521807
Cyprinus	*carpio*	4	A0A173N0R4	LC149722
Danio	*rerio*	4	Q6AZD2	BC078213
Fundulus	*heteroclitus*	4	A0A147B1H5	GCES01001665
Ictalurus	*punctatus*	4	W5UKL2	JT417843
Larimichthys	*crocea*	4	A0A0F8C781	KQ042653

Periophthalmodon	*schlosseri*	4			JACM01058456	
Periophthalmus	*magnuspinnatus*	4			KN462075	
Scartelaos	*histophorus*	4			JACN01023074	
Scleropages	*formosus*	4	A0A0P7WXD9		JARO02005672	
Petromyzon	*marinus*	4		ENSPMAT00000008952		ENSPMAT00000008952
Homo	*sapiens*	5	P55064		AH006636	
Mus	*musculus*	5	Q9WTY4		AF087654	
Myotis	*brandtii*	5	S7NC22		KE164011	
Sus	*scrofa*	5	A8W649		EU192130	
Amazona	*aestiva*	5	A0A0Q3SZZ2		LMAW01003137	
Columba	*livia*	5	R7VMR0		KB376239	
Alligator	*mississippiensis*	5	A0A151LYD6		AKHW03007029	
Chelonia	*mydas*	5	M7ALY6		KB582107	
Crotalus	*adamanteus*	5	J3RYE8		JU173918	
Xenopus	*laevis*	5	Q0PCW5		AB250090	
Homo	*sapiens*	6	Q13520		U48408	
Mus	*musculus*	6	Q8C4A0		AK082699	
Myotis	*davidii*	6	L5MDI4		KB101600	
Pontoporia	*blainvillei*	6	A0A0H4LXH5		KM888080	
Sus	*scrofa*	6	B2MUK1		EU620575	
Ophiophagus	*hannah*	6	V8NJZ3		AZIM01003608	
Pelophylax[b]	*nigromaculatus*	6	D0VY82		AB500707	
Xenopus[b]	*laevis*	6	D0VY84		AJ131847	
Homo	*sapiens*	7	O14520		AB006190	

Mus	*musculus*	7	Q5DX24	AB027516
Myotis	*davidii*	7	L5M978	KB103229
Pontoporia	*blainvillei*	7	A0A0H4LZ74	KM888081
Sus	*scrofa*	7	A9Y007	EU024116
Tupaia	*chinensis*	7	L9L6I7	KB320502
Charadrius	*vociferus*	7	A0A0A0AIW6	KL872051
Tauraco	*erythrolophus*	7	A0A093BS50	KL448390
Alligator	*mississippiensis*	7	A0A151MWE6	AKHW03004724
Chelonia[a]	*mydas*	7	M7B5U5	EMP32514
Xenopus	*tropicalis*	7	Q5FW23	BC089658
Aphyosemion	*striatum*	7	A0A1A7XWA1	HADW01020650
Boleophthalmus	*pectinirostris*	7		KN525596
Cyprinus	*carpio*	7	A0A0K2RVW8	LC069015
Danio	*rerio*	7	D3TZW4	FJ655385
Dicentrarchus	*labrax*	7	E6ZG26	FQ310506
Ictalurus	*furcatus*	7	E3TD21	GU588251
Larimichthys	*crocea*	7	A0A0F8AX88	KQ042348
Nothobranchius	*korthausae*	7	A0A1A8F9V2	HAEB01009009
Periophthalmodon	*schlosseri*	7		KN471526
Periophthalmus	*magnuspinnatus*	7		KN468109
Sparus	*aurata*	7	T1T284	KC589386
Takifugu	*obscurus*	7	H1ABZ7	AB610919
Homo	*sapiens*	8	O94778	AB013456
Mus	*musculus*	8	P56404	AF018952

Sus	*scrofa*	8	A8WCF6	JARO02000680
Tupaia	*chinensis*	8	L9KTK5	KB320654
Alligator	*mississippiensis*	8	A0A151NWY5	AKHW03001697
Latimeria	*chalumnae*	8	H3AXY4	XP_005994818
Anguilla	*japonica*	8	D5MCV8	AB378502
Anoplopoma	*fimbria*	8	C3KH11	BT082226
Boleophthalmus	*pectinirostris*	8ab		KN523248
Cyprinus	*carpio*	8a1	A0A0K2RWE3	LC069018
Cyprinus	*carpio*	8a2	A0A0K2RVU3	LC069019
Cyprinus	*carpio*	8b	A0A0K2RVX2	LC069020
Danio	*rerio*	8a1	Q66I64	BC081511
Danio	*rerio*	8ab	D3TI82	EU341834
Danio	*rerio*	8b	D3U0R1	FJ695516
Ictalurus	*punctatus*	8a1	W5UBX9	JT405474
Ictalurus	*punctatus*	8a2	W5UF90	JT414309
Larimichthys	*crocea*	8	A0A0F8BD75	KQ040999
Larimichthys	*crocea*	8	A0A0F8D0P9	KQ040999
Nothobranchius	*rachovii*	8	A0A1A8RZH6	HAEH01020383
Oryzias	*dancena*	8	A0A023GQJ5	AB759558
Periophthalmodon	*schlosseri*	8ab		JACM01008777/JACM01035013
Periophthalmus	*magnuspinnatus*	8ab		KN464676
Periophthalmus	*magnuspinnatus*	8a1		c67494
Salmo	*salar*	8aa1	S5RRC8	KC626878
Salmo	*salar*	8ab	S5R5I0	KC626879

Salmo	*salar*	8b	C0HB60	BT059566
Scartelaos	*histophorus*	8ab		KN499303
Scartelaos	*histophorus*	8a1		JACN01005585
Scleropages	*formosus*	8	A0A0P7XQS4	JARO02000680
Sparus	*aurata*	8	C7S301	DQ889225
Takifugu	*obscurus*	8	H1AC01	AB610923
Homo	*sapiens*	9	O43315	AB008775
Mus	*musculus*	9	Q9JJJ3	AB037180
Myotis	*davidii*	9	L5LSV1	KB108369
Pontoporia	*blainvillei*	9	A0A0H4LW47	KM888082
Sus	*scrofa*	9	A8WCF7	EU220427
Tupaia	*chinensis*	9	L9L8A8	KB320468
Amazona	*aestiva*	9	A0A0Q3X7K3	LMAW01000003
Alligator	*mississippiensis*	9	A0A151M9R5	AKHW03006295
Chelonia	*mydas*	9	M7CDV6	KB518028
Hyla	*chrysoscelis*	9	A0A192F0W8	KU695572
Latimeria	*chalumnae*	9	M3XHS6	XP_014345556
Boleophthalmus	*pectinirostris*	9a		KN524979
Cyprinus	*carpio*	9a1	A0A173N0K5	LC149726
Cyprinus	*carpio*	9b1	A0A0K2RVT5	LC069012
Danio	*rerio*	9a	Q498W2	BC100051
Danio	*rerio*	9b	D5FFZ1	EU341835
Fundulus	*heteroclitus*	9	A0A146N9Z0	GCES01157774
Fundulus	*heteroclitus*	9	A0A146ZLF7	GCES01019129

Larimichthys	*crocea*	9	A0A0F8CJ34	KQ041939
Nothobranchius	*furzeri*	9	A0A1A7ZJN6	HADY01004016
Periophthalmodon	*schlosseri*	9a		KN482823
Periophthalmodon	*schlosseri*	9b		KN469413
Periophthalmus	*magnuspinnatus*	9a		KN462368
Periophthalmus	*magnuspinnatus*	9b		KN467715
Scartelaos	*histophorus*	9b		JACN01079684
Sparus	*aurata*	9b	T1T061	KC589387
Homo	*sapiens*	10	Q96PS8	AF159174
Cricetulus	*griseus*	10	G3HBZ6	JH000275
Myotis	*brandtii*	10	S7PKF6	KE162497
Sus	*scrofa*	10	B2D2K0	EU582021
Aptenodytes	*forsteri*	10	A0A087R9Y3	KL226233
Alligator	*mississippiensis*	10	A0A151PJ30	AKHW03000118
Latimeria	*chalumnae*	10	XP005996556	XM_005996494
Anguilla	*japonica*	10	D5MCV9	AB378503
Anoplopoma	*fimbria*	10	C3KI76	BT082641
Boleophthalmus	*pectinirostris*	10a		KN523925
Cyprinus	*carpio*	10a1	A0A0K2RVT9	LC069014
Cyprinus	*carpio*	10b	A0A0K2RVX4	LC069016
Danio	*rerio*	10a	D3TZW7	FJ655388
Danio	*rerio*	10b	D3TI83	EU341836
Oryzias	*dancena*	10	A0A023GQJ9	AB759559
Periophthalmodon	*schlosseri*	10a		KN473499

Periophthalmodon	*schlosseri*	10b		KN474915
Periophthalmus	*magnuspinnatus*	10a		KN468565
Periophthalmus	*magnuspinnatus*	10b		KN460469
Scartelaos	*histophorus*	10a		JACN01051648
Scartelaos	*histophorus*	10b		JACN01031370
Sparus	*aurata*	10	Q67EP9	AY363261
Takifugu	*obscurus*	10a	H1ABZ8	ϑ20
Homo	*sapiens*	11	Q8NBQ7	AB028147
Mus	*musculus*	11	Q8BHH1	AB028148
Myotis	*brandtii*	11	S7PJQ3	KE162454
Sus	*scrofa*	11	A8WCF5	EU220425
Tupaia	*chinensis*	11	L9KV50	KB320703
Amazona	*aestiva*	11	A0A0Q3Q6P1	LMAW01001259
Chelonia	*mydas*	11	M7AZN7	KB562012
Aphyosemion	*striatum*	11	A0A1A7XI76	HADW01016367
Boleophthalmus	*pectinirostris*	11a		KN525406
Boleophthalmus	*pectinirostris*	11b		KN522455
Cyprinus	*carpio*	11-2	A0A0K2RVU7	LC069022
Danio	*rerio*	11	NP001314822	NM_001327893
Fundulus	*heteroclitus*	11	A0A146NB79	GCES01158036
Fundulus	*heteroclitus*	11	A0A147AUG0	GCES01004272
Ictalurus	*punctatus*	11	W5UCF5	JT411887
Larimichthys	*crocea*	11	A0A0F8AXT5	KQ042331
Nothobranchius	*kuhntae*	11	A0A1A8J9K5	HAED01019731

Oryzias	*dancena*	11	A0A023GQJ8	AB759560
Osmerus	*mordax*	11	C1BJY3	BT074912
Periophthalmodon	*schlosseri*	11b		KN480662
Periophthalmodon	*schlosseri*	11a		KN469785
Periophthalmus	*magnuspinnatus*	11a		KN461794
Periophthalmus	*magnuspinnatus*	11b		KN460458
Scartelaos	*histophorus*	11a		JACN01020100
Scartelaos	*histophorus*	11b		JACN01076962/JACN01183528
Takifugu	*obscurus*	11a	H1AC02	AB610924
Homo	*sapiens*	12a	Q8IXF9	AB040748
Camelus	*ferus*	12	S9XC49	KB016506
Mus	*musculus*	12	Q3V2I1	BC107212
Myotis	*davidii*	12	L5M4L9	KB105166
Amazona	*aestiva*	12	A0A0Q3PLJ4	LMAW01002794
Pelecanus	*crispus*	12	A0A091SMY5	KK477153
Chelonia	*mydas*	12	M7BRQ5	KB519804
Boleophthalmus	*pectinirostris*	12		KN522461
Cyprinus	*carpio*	12-2	A0A0K2RVU8	LC069024
Danio	*rerio*	12	NP001039327	NM_001045862
Fundulus	*heteroclitus*	12b	A0A146N2N5	GCES01161006
Ictalurus	*punctatus*	12	W5ULH9	JT418862
Oryzias	*dancena*	12	A0A023GQK2	AB759561
Periophthalmodon	*schlosseri*	12		KN477140
Periophthalmus	*magnuspinnatus*	12		KN467541

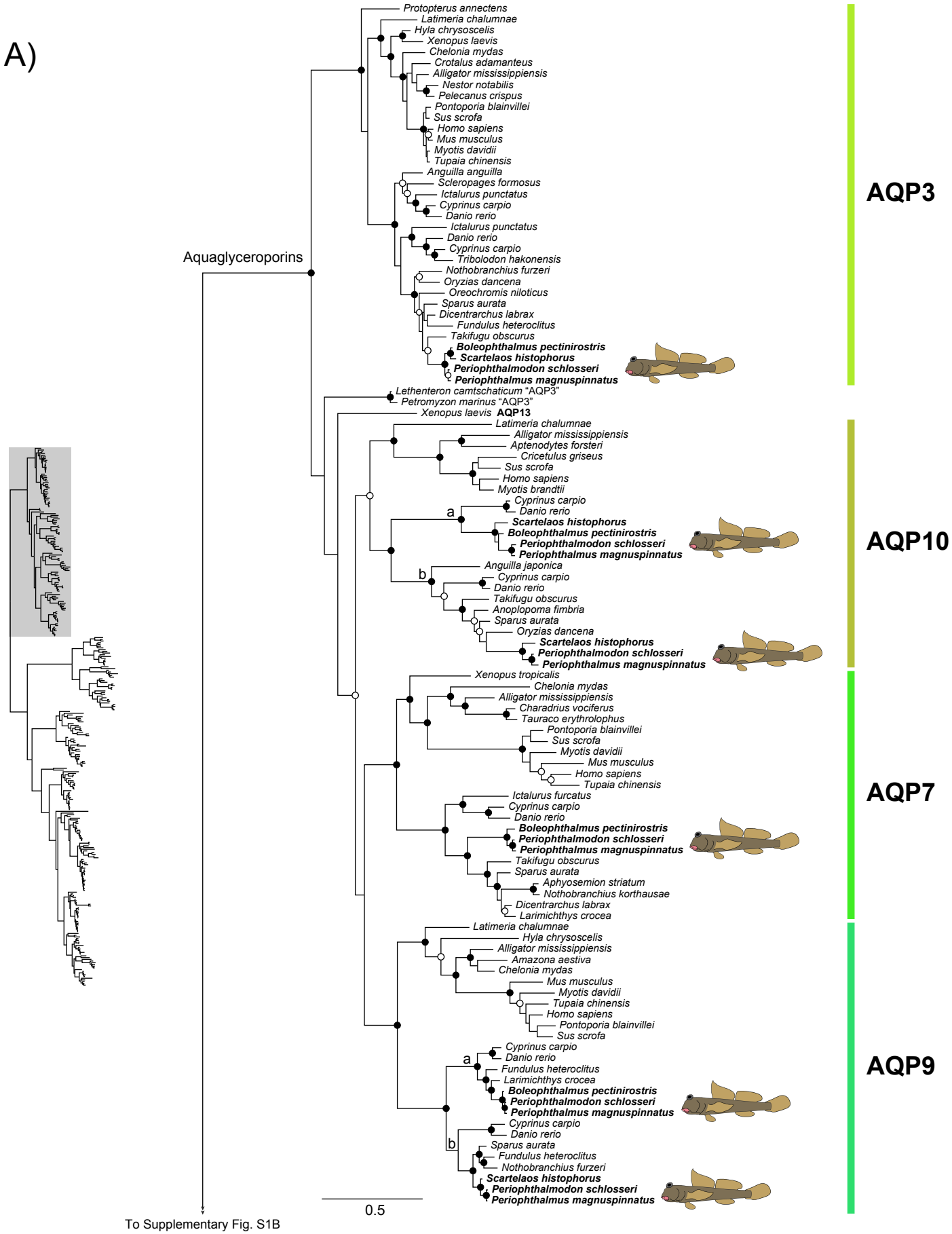
Scartelaos	*histophorus*	12			KN493770
Takifugu	*obscurus*	12	H1AC03		AB610925
Xenopus	*laevis*	13	NP_001082310		
Anolis		14		ENSACAT00000030435	
Oreochromis	sp.	14		ENSONIT00000021542	
Oreochromis	*niloticus*	15	XP_003442458		
Xenopus	*laevis*	16	NP_001089643		

[a]Aquaporin sequences with incorrect annotation in GenBank. Corrected annotation (indicated here) based on our phylogenetic results

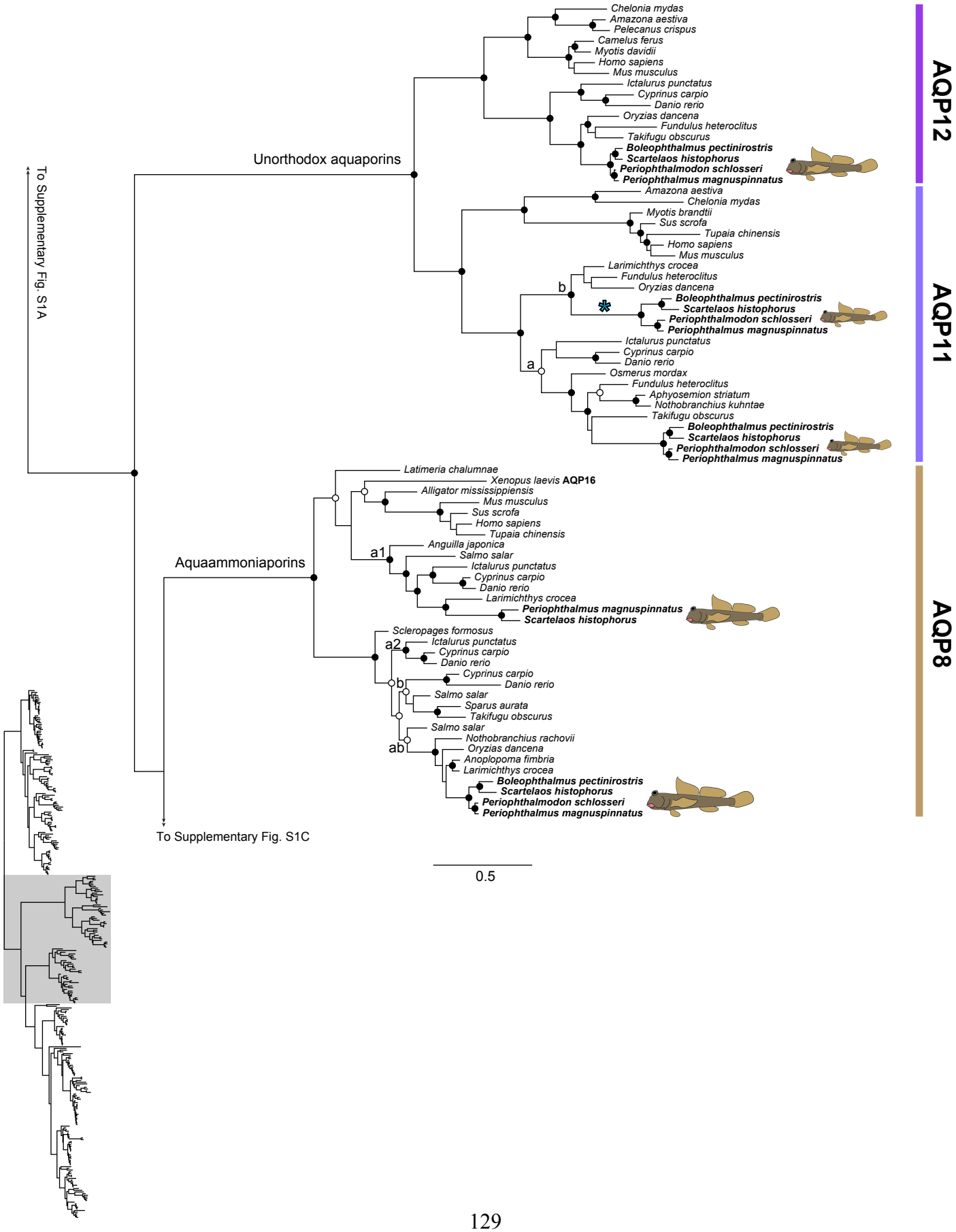
[b]Aquaporin sequences with incorrect annotation in GenBank. Corrected annotation (indicated here) based on our phylogenetic results and those by Finn et al. (2014)

Table S2. Results of the branch-site tests for AQP11b if only marine or only brackish/euryhaline fish outgroup species are included in the comparisons.

Fish outgroups	**Foreground branch**	**LRT**	***P* value**	**ω**
Marine	Mudskipper clade	2.977	0.042	1.857
Marine	Mudskipper clade stem	19.096	6.217×10^{-6}	30.965
Marine	*Periopthalmus* + *Periopthalmodon* clade stem	3.144	0.038	18.894
Brackish/euryahline	Mudskipper clade	3.416	0.032	1.608
Brackish/euryahline	Mudskipper clade stem	14.625	6.558×10^{-5}	18.202
Brackish/euryahline	Periopthalmus + Periopthalmodon clade stem	5.811	0.008	18.025


A)
Aquaglyceroporins
Protopterus annectens
Latimeria chalumnae
Hyla chrysoscelis
Xenopus laevis
Chelonia mydas
Crotalus adamanteus
Alligator mississippiensis
Nestor notabilis
Pelecanus crispus
Pontoporia blainvillei
Sus scrofa
Homo sapiens
Mus musculus
Myotis davidii
Tupaia chinensis
Anguilla anguilla
Scleropages formosus
Ictalurus punctatus
Cyprinus carpio
Danio rerio
Ictalurus punctatus
Danio rerio
Cyprinus carpio
Tribolodon hakonensis
Nothobranchius furzeri
Oryzias dancena
Oreochromis niloticus
Sparus aurata
Dicentrarchus labrax
Fundulus heteroclitus
Takifugu obscurus
Boleophthalmus pectinirostris
Scartelaos histophorus
Periophthalmodon schlosseri
Periophthalmus magnuspinnatus
Lethenteron camtschaticum "AQP3"
Petromyzon marinus "AQP3"
Xenopus laevis AQP13
Latimeria chalumnae
Alligator mississippiensis
Aptenodytes forsteri
Cricetulus griseus
Sus scrofa
Homo sapiens
Myotis brandtii
a
Cyprinus carpio
Danio rerio
Scartelaos histophorus
Boleophthalmus pectinirostris
Periophthalmodon schlosseri
Periophthalmus magnuspinnatus
b
Anguilla japonica
Cyprinus carpio
Danio rerio
Takifugu obscurus
Anoplopoma fimbria
Sparus aurata
Oryzias dancena
Scartelaos histophorus
Periophthalmodon schlosseri
Periophthalmus magnuspinnatus
Xenopus tropicalis
Chelonia mydas
Alligator mississippiensis
Charadrius vociferus
Tauraco erythrolophus
Pontoporia blainvillei
Sus scrofa
Myotis davidii
Mus musculus
Homo sapiens
Tupaia chinensis
Ictalurus furcatus
Cyprinus carpio
Danio rerio
Boleophthalmus pectinirostris
Periophthalmodon schlosseri
Periophthalmus magnuspinnatus
Takifugu obscurus
Sparus aurata
Aphyosemion striatum
Nothobranchius korthausae
Dicentrarchus labrax
Larimichthys crocea
Latimeria chalumnae
Hyla chrysoscelis
Alligator mississippiensis
Amazona aestiva
Chelonia mydas
Mus musculus
Myotis davidii
Tupaia chinensis
Homo sapiens
Pontoporia blainvillei
Sus scrofa
Cyprinus carpio
Danio rerio
a
Fundulus heteroclitus
Larimichthys crocea
Boleophthalmus pectinirostris
Periophthalmodon schlosseri
Periophthalmus magnuspinnatus
Cyprinus carpio
Danio rerio
b
Sparus aurata
Fundulus heteroclitus
Nothobranchius furzeri
Scartelaos histophorus
Periophthalmodon schlosseri
Periophthalmus magnuspinnatus
0.5
To Supplementary Fig. S1B
AQP3
AQP10
AQP7
AQP9

B)


Chelonia mydas
Amazona aestiva
Pelecanus crispus
Camelus ferus
Myotis davidii
Homo sapiens
Mus musculus
Ictalurus punctatus
Cyprinus carpio
Danio rerio
Oryzias dancena
Fundulus heteroclitus
Takifugu obscurus
Boleophthalmus pectinirostris
Scartelaos histophorus
Periophthalmodon schlosseri
Periophthalmus magnuspinnatus
Unorthodox aquaporins
To Supplementary Fig. S1A
Amazona aestiva
Chelonia mydas
Myotis brandtii
Sus scrofa
Tupaia chinensis
Homo sapiens
Mus musculus
Larimichthys crocea
Fundulus heteroclitus
Oryzias dancena
b
Boleophthalmus pectinirostris
Scartelaos histophorus
Periophthalmodon schlosseri
Periophthalmus magnuspinnatus
Ictalurus punctatus
Cyprinus carpio
Danio rerio
a
Osmerus mordax
Fundulus heteroclitus
Aphyosemion striatum
Nothobranchius kuhntae
Takifugu obscurus
Boleophthalmus pectinirostris
Scartelaos histophorus
Periophthalmodon schlosseri
Periophthalmus magnuspinnatus
Latimeria chalumnae
Xenopus laevis AQP16
Alligator mississippiensis
Mus musculus
Sus scrofa
Homo sapiens
Tupaia chinensis
Anguilla japonica
a1
Salmo salar
Aquaammoniaporins
Ictalurus punctatus
Cyprinus carpio
Danio rerio
Larimichthys crocea
Periophthalmus magnuspinnatus
Scartelaos histophorus
Scleropages formosus
a2
Ictalurus punctatus
Cyprinus carpio
Danio rerio
Cyprinus carpio
b
Danio rerio
Salmo salar
Sparus aurata
Takifugu obscurus
Salmo salar
Nothobranchius rachovii
Oryzias dancena
ab
Anoplopoma fimbria
Larimichthys crocea
Boleophthalmus pectinirostris
Scartelaos histophorus
Periophthalmodon schlosseri
Periophthalmus magnuspinnatus
To Supplementary Fig. S1C
0.5
AQP12
AQP11
AQP8

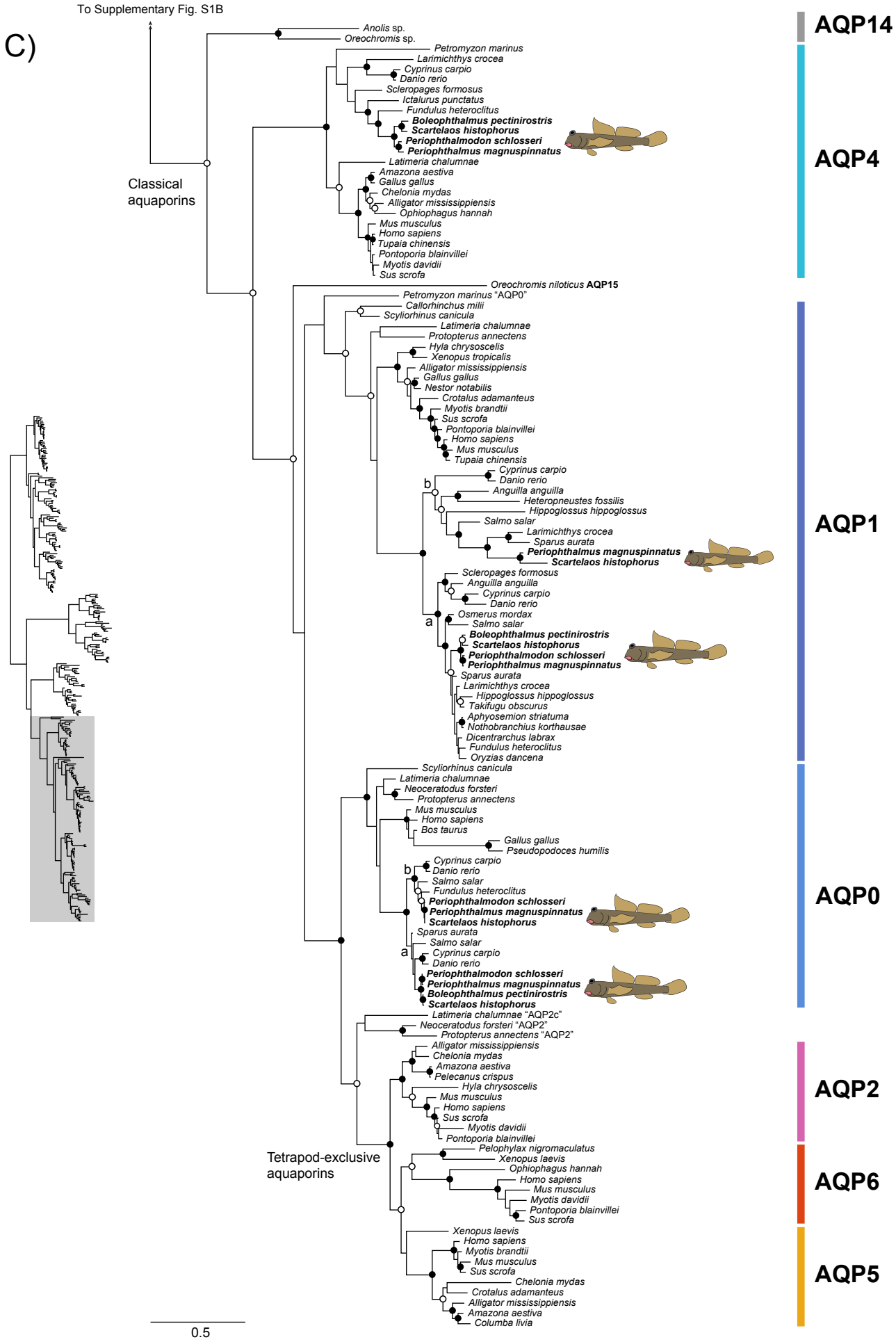

C)
To Supplementary Fig. S1B
Classical aquaporins
Tetrapod-exclusive aquaporins
Anolis sp.
Oreochromis sp.
Petromyzon marinus
Larimichthys crocea
Cyprinus carpio
Danio rerio
Scleropages formosus
Ictalurus punctatus
Fundulus heteroclitus
Boleophthalmus pectinirostris
Scartelaos histophorus
Periophthalmodon schlosseri
Periophthalmus magnuspinnatus
Latimeria chalumnae
Amazona aestiva
Gallus gallus
Chelonia mydas
Alligator mississippiensis
Ophiophagus hannah
Mus musculus
Homo sapiens
Tupaia chinensis
Pontoporia blainvillei
Myotis davidii
Sus scrofa
Oreochromis niloticus AQP15
Petromyzon marinus "AQP0"
Callorhinchus milii
Scyliorhinus canicula
Protopterus annectens
Hyla chrysoscelis
Xenopus tropicalis
Nestor notabilis
Crotalus adamanteus
Myotis brandtii
Anguilla anguilla
Heteropneustes fossilis
Hippoglossus hippoglossus
Salmo salar
Sparus aurata
Osmerus mordax
Takifugu obscurus
Aphyosemion striatuma
Nothobranchius korthausae
Dicentrarchus labrax
Oryzias dancena
Neoceratodus forsteri
Bos taurus
Pseudopodoces humilis
Latimeria chalumnae "AQP2c"
Neoceratodus forsteri "AQP2"
Protopterus annectens "AQP2"
Pelecanus crispus
Pelophylax nigromaculatus
Xenopus laevis
Columba livia
a
b
0.5
AQP14
AQP4
AQP1
AQP0
AQP2
AQP6
AQP5

Figure S1. Bayesian inference phylogram of vertebrate aquaporins based on 394 aligned amino acid positions. Filled bullets on nodes denote support for both BI and ML methods of phylogenetic inference. Open bullets on nodes denote support for only one of the methods (either BI or ML). Absence of bullet on a node denotes no support for any of the methods. Aquaporin classes are indicated on the right and class paralogues with letters (a, b, a1, a2, ab) on the corresponding branches. Mudskipper branches are highlighted in bold type. Blue asterisks highlight the branches in which positive selection signatures are detected. Scale bar represents amino acid substitutions per site

CHAPTER IV

Multiple instances of adaptive evolution in aquaporins of amphibious fishes

---

Múltiples casos de evolución adaptativa en las acuaporinas de peces anfibios

This chapter is a reproduction of the following manuscript, currently in preparation:

Lorente-Martínez, H., Agorreta, A., Irisarri, I., Zardoya, R., Edwards, S.V., & San Mauro, D. Multiple instances of adaptive evolution in aquaporins of amphibious fishes.

Abstract

Aquaporins (AQPs) are a highly diverse family of transmembrane proteins involved in osmotic regulation that played an important role in the conquest of land by tetrapods. However, little is known about their possible implication in the development of an amphibious lifestyle in actinopterygian fishes. In this work, we have investigated the molecular evolution of aquaporins in 22 putative amphibious fishes across actinopterigians by assembling a comprehensive dataset that was subjected to adaptive selection analyses. We found evidence of adaptive selection in 19 aquaporins (including AQP1, AQP3, AQP8, AQP10, AQP11, and AQP12 subfamilies). However, we only identified specific positively-selected sites in 13 of these branches. Interestingly, almost half of the branches and sites that are under selection are found in the AQP11 subfamily. The sequence changes detected suggest modifications of the molecular function and/or structure of the aquaporins, which can be related with the adaptation to an amphibious lifestyle and the subsequent development of terrestrial traits in some of the most extreme cases. Our results also suggest that AQP11 orthologues are the most promising candidates to have been involved in the processes of water-to-land transition in amphibious fishes. Besides, the signature of adaptive selection found in the AQP11b stem branch of the Gobiidae clade (occurring before the appearance and evolution of the mudskippers lineage) suggests a possible case of exaptation in this clade.

Keywords: aquaporin, amphibious fishes, adaptive selection, emersion.

Resumen

Las acuaporinas (AQPs) son una familia muy diversa de proteínas transmembrana involucradas en la regulación osmótica que jugaron un papel importante en la conquista de la tierra por parte de los tetrápodos. Sin embargo, se conoce poco sobre su posible implicación en el desarrollo de un estilo de vida anfibio en los peces actinopterigios. En este trabajo, hemos investigado la evolución molecular de las acuaporinas en 22 posibles peces anfibios del grupo de los actinopterigios mediante el ensamblaje de un amplio y exhaustivo conjunto de datos que se sometió a análisis de selección adaptativa. Encontramos evidencia de selección adaptativa en 19 acuaporinas (incluidas las subfamilias AQP1, AQP3, AQP8, AQP10, AQP11 y AQP12). Sin embargo, solo detectamos sitios específicos seleccionados positivamente en 13 de estas ramas. Es interesante destacar que casi la mitad de las ramas y los sitios que están bajo selección se encuentran en la subfamilia de la AQP11. Los cambios de secuencia detectados sugieren modificaciones de la función molecular y/o de la estructura de las acuaporinas que pueden estar relacionados con la adaptación a un estilo de vida anfibio y el subsecuente desarrollo de la terrestrialidad en algunos de los casos más extremos. Nuestros resultados también sugieren que los ortólogos de la AQP11 son los candidatos más prometedores a haber estado implicados en los procesos de transición agua-tierra en los peces anfibios. Además, la señal de selección adaptativa en la rama ancestral de la AQP11b del clado Gobiidae (que es anterior a la aparición y evolución del linaje de los saltarines del fango) sugiere un posible caso de exaptación en este clado.

Palabras clave: acuaporina, peces anfibios, selección adaptativa, transición agua-tierra.

Introduction

Adaptation to new, unexploited environments is generally challenging, but provides most advantageous conditions for natural selection to trigger evolution and diversification (J. B. Graham, 1997; Nosil, 2012; Schluter, 2000). In particular, water-to-land transitions appear to be the most recurrent remarkable and challenging adaptations in the history of life (Laurin, 2010; Long & Gordon, 2004). Water provides a buoyant medium where osmoregulation is a straightforward process, whereas terrestrial animals must confront higher gravitational pressure and desiccation conditions. Consequently, emersion from water is a complex evolutionary process that involves numerous morphological (biomechanical) and physiological (metabolic and biochemical) changes, mostly associated with locomotion, vision, audition, respiration, and desiccation (J. B. Graham, 1997; Laurin, 2010; Sayer, 2005). Within vertebrates, the multiple convergent cases of terrestrialization that have occurred during the evolutionary history of teleost fishes provide an excellent model system for studying and comparing the tempo and mode of these complex adaptations (reviewed in J. B. Graham, 1997; Ord & Cooke, 2016; Wright & Turko, 2016). These, dubbed amphibious fishes, typically inhabit intertidal areas, taking refuge in small pools during low tides (Bridges, 2015; J. B. Graham, 1997; Ord & Cooke, 2016) and present several adaptations for emersion. Many of them are air-breathers (J. B. Graham, 1997), as is the case of mudskippers, which can gulp air (J. Graham et al., 2007) and killifishes, which can use their skin as a gas exchanger (Frick & Wright, 2002a, 2002b; LeBlanc et al., 2010). Likewise, higher ammonia tolerance and the ability to actively excrete this compound appear to be widespread adaptations in amphibious fishes (e.g. Chew et al., 2005; Ip, Chew, et al., 2004; Ip, Randall, et al., 2004; Randall et al., 2015). There are also outstanding examples of terrestrial locomotion as in the climbing perch (*Anabas testudineus*) and the walking catfish (*Clarias batrachus*) also (Davenport & Matin, 1990; Pace & Gibb, 2014). However, despite water loss and desiccation in the aerial environment are major challenges faced by amphibious fishes, little it is known about the mechanisms underpinning water recovery, maintenance and homeostasis during emersion in these fishes. Konno et al. (2010) described a strategy for active water recovery in lungfish kidneys that is activated in terrestrial conditions. This mechanism involves a water channel named aquaporin, and it was firstly discovered in tetrapods (Konno et al., 2009, 2010; van Balkom et al., 2002). In fact, the emergence of three new

paralogues (AQP2, 5 and 6) of these proteins in this lineage could have been relevant for the conquest of land (Finn et al., 2014).

Aquaporins or AQPs (also known as major intrinsic proteins, MIPs) are transmembrane channels that can carry water and small uncharged solutes (Agre et al., 1993; Agre & Kozono, 2003; Connolly et al., 1998). Besides water, AQPs can transport a plethora of compounds such as glycerol, urea, ammonia, CO_2, reactive oxygen species (ROS), and hydrogen peroxide (Bestetti et al., 2020; Finn & Cerdà, 2015), suggesting a broadly relevance in different physiological mechanisms. Up to 17 vertebrate different aquaporins subfamilies or classes have been described, which can be clustered into four main groups: (1) the aquaglyceroporins or GLPs; (2) the water-selective classical aquaporins; (3) the unorthodox or superaquaporins; and (4) the AQP8-type or aquaammoniaporins (Abascal et al., 2014; Finn et al., 2014; Zardoya, 2005). Aquaporins are particularly abundant in the main organs for water recovery in fishes, including gills, intestine, and kidney; and they have been broadly associated with the osmoregulatory process in fishes and with fish acclimatisation to different salinities (reviewed in Madsen et al., 2015). In tetrapods (i.e., amphibians, sauropsids and mammals), which have a terrestrial lifestyle, several AQPs have been involved directly in the mechanistic basis of water conservation (reviewed in Finn et al., 2014). Hence, it can be postulated that some AQPs could have been recruited to be involved in the physiological adaptation needed during emersion of amphibious fishes. For instance, Ip et al. (2013) postulated that the upregulation of an aquaporin in the gills and skin of the climbing perch could be related with a higher ammonia excretion.

Here, the molecular evolution of aquaporins in 22 fish genomes in the context of water-to-land adaptation is investigated. This study follows up on our earlier work on the role of aquaporins in the amphibious behaviour of mudskippers (Gobiidae) based on four genomes (Lorente-Martínez et al., 2018), and takes advantage of the recently available genomic data from of other amphibious fishes, thus assembling a more comprehensive dataset. With these expanded data, we conduct cataloguing of aquaporin classes and paralogues, reconstruction of phylogenetic frameworks within vertebrates, and investigation of molecular evolution and adaptive selection at the sequence level. Our results pinpoint a wide range of adaptive evolution events in different aquaporins across the studied amphibious fish lineages, suggesting the possible role of these proteins in their adaptation to an amphibious lifestyle.

Results and Discussion

Diversity of amphibious fish aquaporins

We assemble our previous dataset (Lorente-Martínez et al., 2018), plus several new genomes from species that either showed some degree of amphibious development or have undergone some degree of amphibiousness although not sufficient as to be considered truly amphibious (reviewed in J. B. Graham, 1997; Ord & Cooke, 2016; Wright & Turko, 2016). Up to 22 genomes of amphibious fishes along with four sequences of lungfishes (of classes AQP0 and AQP2-like) were retrieved (Fig. 1). Additionally, 42 vertebrates (Table S1) with special emphasis in fishes were included. Redundant and misleading sequences were pruned. Our genomic screening led us to a comprehensive protein alignment of 1006 terminals and 441 amino acid positions, of which 439 are variable and 434 are parsimony informative. This dataset was subject to maximum likelihood (ML) under RAxML (−ln L = 172,803.455) and IQ-TREE (−ln L = 172,474.585), both yielding highly similar topologies with slight differences mainly concentrated on low-supported branches (Figs. 2 and S1). The reconstructed trees recovered 16 AQP classes grouped into four main clades with strong statistical support: (1) the aquaglyceroporins (AQP3, 7, 9, 10); (2) the water-selective classical aquaporins (AQP0, 1, 2, 4, 5, 6, 14, 15); (3) the unorthodox or superaquaporins (AQP11 and 12); and (4) the AQP8-type or aquaammoniaporins (AQP8 and 16; Fig. 2 and Fig. S1; Abascal et al., 2014; Finn et al., 2014; Zardoya, 2005). Many of the AQP paralogues were generated through two rounds of whole genome duplication (WGD) events that occurred early in the evolution of vertebrates (Finn et al., 2014; Yilmaz et al., 2020). In addition, a more recent WGD event occurred at the stem branch of teleost fishes provided them a wider repertorie of AQP genes (Cerdà & Finn, 2010; Taylor et al., 2001).

Figure 1. Aquaporin catalogue of the studied amphibious fishes. Red triangles denote complete sequences retrieved from whole-genome shotgun data. Orange triangles denote aquaporins retrieved in Lorente-Martínez et al. (2018). Blue triangles denoted sequences directedly downloaded from the GenBank nucleotide database. Half grey triangles denote partial sequences. *The existence of *Anguilla anguilla* AQP16 is unclear. **The duplication of these classes (dubbed 'a' and 'b') mainly occurred on the branch of teleost fishes, therefore *Erpetoichthys calabaricus, Polypterus senegalus*, *Neoceratodus forsteri* and *Protopterus annectens* only possess one copy of each paralogue, except for *E. calabaricus* AQP10 and *P. senegalus* AQP8 and 10. ***Proposed as AQP2-like (see Finn et al., 2014).

Focusing on amphibious fish, a total of 356 putative AQP genes were recovered (Fig. 1; including nine more mudskipper AQPs than previously reported Lorente-Martínez et al., 2018). These aquaporins could be classified into 13 different aquaporin classes. For most amphibious fish species, one paralogue per class was found, no gene expansions were detected, and only some putative cases of gene loss could be spotted. In particular, few AQP15 orthologues were identified and several species lacked this paralogue (Figs. 1, 2 and S1). This gene loss was previously described by Finn et al. (2014) and associated to a genome reduction (Wolf & Koonin, 2013). However, its potential physiological impact of this gene loss is not well understood. On one hand,

several studies have shown examples of overlapping functions among different aquaporin classes (reviewed in Finn & Cerdà, 2015). On the other hand, the patterns of expression of these genes are highly diverse and even under similar physiological conditions, there are reports of different expression patterns among species and organs (see Madsen et al., 2015). Altogether, these results suggest a complex evolutionary pattern beyond birth and death gene family processes that may not be disentangled simply using presence-absence approaches. However, a very recent study has found patterns of gene loss and duplications in arthropod aquaporins that may be related to the acclimation to non-marine environments and ultimately with the water-to-land transition in arthropods (Martínez-Redondo et al., 2023)

Surprisingly, we found a match for a putative AQP16 orthologue within the *Anguilla anguilla* genome although no other AQP16 orthologue was described in any fish genome (Finn et al., 2014). However, our tree does not recover a supported clade for AQP16 orthologues (Fig. 2 and S1), so the poor phylogenetic resolution in this part of the tree does not allow a proper classification of this sequence and led us to dismiss it for further analysis. As expected, we did not find any AQP13 orthologue as this class has only been described in the genomes of a few amphibians, hagfishes, and the platypus (Yilmaz et al., 2020). Therefore, based on this result, and aiming for a better resolution within the aquaglyceroporins clade, we decided to remove all the AQP13 sequences from our dataset. Additionally, in concordance with previous studies (Abascal et al., 2014; Finn et al., 2014; Finn & Cerdà, 2011; Tingaud-Sequeira et al., 2010), no AQP2, 5 and 6 was found in any of the studied actinopterygian fish genome, and we were not able to describe any novel aquaporin class that could have been related with the amphibious development of these fishes. This suggests that if aquaporins were relevant for terrestrialisation of these fishes this could have been achieved by changes at the molecular level—that could entail conformational changes involved in, for instance, differences in function or expression—on aquaporins subfamilies that are also present in their fully-aquatic relatives.

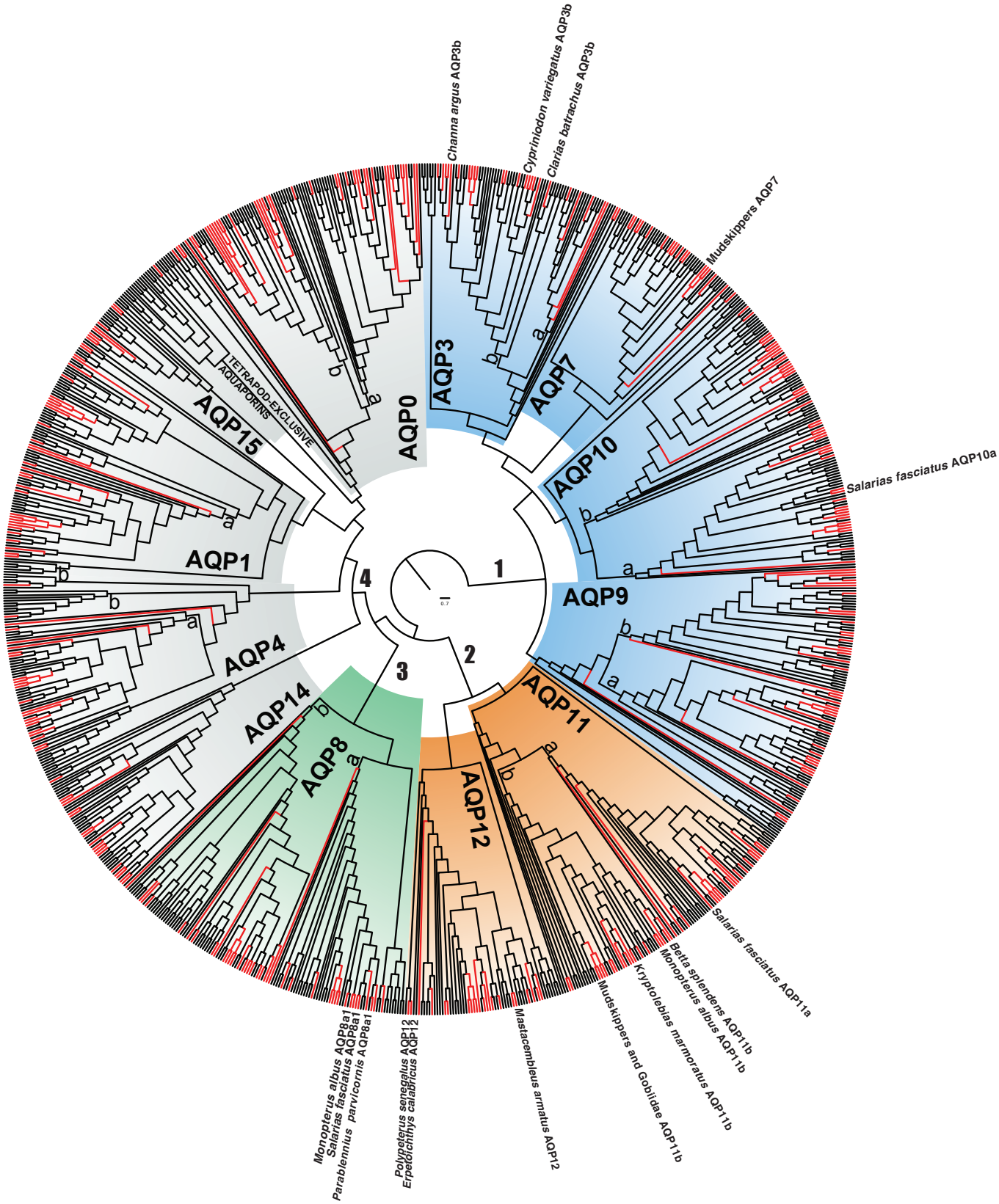


Figure 2. Maximum likelihood (IQ-TREE) cladogram of vertebrate aquaporins based on 441 aligned amino acid positions. Aquaporin classes are indicated with colour panels and class paralogues with letters (a, b) on the corresponding branches. Branches of amphibious fishes are highlighted in red. The names of the branches under adaptive selection are written near their

terminal location in the tree. A detailed, fully labelled, metric version of this phylogenetic tree is shown in Supplementary Fig. S1.

Adaptive selection in amphibious fish aquaporins

The footprints of positive or adaptive selection can be detected in gene sequences by estimating the ratio between non-synonymous and synonymous nucleotide substitutions (dN/dS), usually known as the omega (ω) value (Yang, 2008). We investigated this parameter to see whether there were any differentially selected positions on aquaporin protein sequences of the amphibious fishes that could be linked to their amphibiousness. When positive selection occurs, the fixation of advantageous mutations can potentially lead to the development of evolutionary innovations due to the emergence of novel molecular functions (Nielsen & Yang, 1998). Two recent studies detected positively selected sites in AQPs, in aquaglyceroporins of squamates, which could be related to adaptation to life in dry habitats (São Pedro et al., 2015), and during the cetacean land-to-water transition (Zang et al., 2019). Similarly, branch-site tests were conducted to search for positions under adaptive selection in each paralogue in those branches of the reconstructed tree that led to amphibious fish species, and thus could be related to adaptation to terrestrial environments.

Table 1. Results of the branch-site tests that are significant (*Q*-value of the LRT <0.05) and have ω>1 in foreground branches. All other tests not shown in the table are not significant and have ω≤1 in foreground branches although one branch of *Fundulus heteroclitus* and one of *Clarias bactrachus* (see Results and Discussion)

Aquaporin	Foreground branch	LRT	*P*-value[a]	*Q*-value[b]	ω[c]	Prop. 2a[d]	Prop. 2b[e]	Selected sites
1a	*Anableps* (Cyprinodontiformes)	7.711	0.003	0.0298	999	0.006	0.001	0
3a	*Channa* (Anabantiformes)	16.551	2.367 x 10-5	2.603 x 10-4	31.599	0.032	0.008	2
3a	*Clarias* (Siluriformes)	19.478	5.087×10^{-6}	1.119×10^{-4}	49.655	0.037	0.009	3
3a	*Cyprinodon* (Cyprinodontiformes)	9.906	8.233×10^{-4}	0.006	67.061	0.013	0.003	1
7	Mudskipper clade stem (Gobiiformes)	8.639	0.002	0.026	999	0.015	0.004	0
8a1	*Monopterus* (Synbranchiformes)	12.334	2.217×10^{-4}	0.007	12.806	0.068	0.013	6
8a1	Mudskipper clade stem (Gobiiformes)	7.295	0.003	0.027	36.256	0.019	0.004	0
8a1	*Parablennius* (Blenniiformes)	7.997	0.002	0.028	23.468	0.085	0.017	3
8b1	*Betta* (Anabantiformes)	7.460	0.003	0.028	8.669	0.039	0.008	0
10b	*Salarias* (Blenniiformes)	13.001	1.557×10^{-4}	0.003	59.219	0.019	0.003	1
11b	*Betta* (Anabantiformes)	15.126	5.029×10^{-5}	8.298×10^{-4}	43.954	0.030	0.007	3
11b	*Kryptolebias* (Cyprinodontiformes)	6.510	0.005	0.029	38.602	0.022	0.005	2
11b	*Monopterus* (Synbranchiformes)	13.291	1.334×10^{-4}	0.001	18.128	0.038	0.009	3
11b	Mudskipper clade stem (Gobiiformes)	9.607	9.690×10^{-4}	0.006	999	0.029	0.006	0
11a	*Salarias* (Blenniiformes)	12.613	1.916×10^{-4}	0.002	16.503	0.047	0.011	2
12	*Erpetoichthys* (Polypteriformes)	10.325	6.560×10^{-4}	0.004	708.322	0.021	0.005	1
12	*Mastacembelus* (Synbranchiformes)	10.907	4.790×10^{-4}	0.004	998.999	0.014	0.003	1
12	*Polypterus* (Polypteriformes)	11.961	2.717×10^{-4}	0.004	999	0.016	0.004	0
15	*Anableps* (Cyprinodontiformes)	9.594	9.761×10^{-4}	0.008	1	0.057	0.017	0

[a]Uncorrected *P*-value of the LRT

[b]Multiple-test correction of the LRT *P*-value (False Discovery Rate). For AQP11 sequences we also included the stem branch of the Gobiidae clade for correction. See Results and Discussion and Table S2.

[c]Omega (dN/dS) ratio of the foreground branch(es)

[d]Proportion of sites that are under positive selection (ω 2a > 1) on the foreground branch(es) and under purifying selection (ω < 1) on background branches

[e]Proportion of sites that are under positive selection (ω 2a > 1) on the foreground branch(es) and under neutral selection (ω = 1) on background branches

A total of 19 different branches of the studied amphibious fishes, which expand across seven different fish orders, showed footprints of adaptive selection (Table 1). However, specific positions under adaptive selection could only be identified in 13 out of these 19 branches (Fig. 3). This discrepancy could indicate that in some branches the signal of adaptive selection was cumulative and not strong enough at any particular site. Moreover, the branch-site test used for detecting which positions could have undergone adaptive selection is highly conservative and sometimes may lack enough statistical power (see Methods and Yang et al., 2005). Another potential caveat may be the presence in some AQPs of highly variable regions (which could reflect either true fast evolutionary rates or more likely, low quality sequences), as of the employed tests heavily rely on robust alignments (Zhang et al., 2005). Even though, we carefully clean and performed a by-eye inspection of all of them (see Methods), still some very variable regions were retained in a few aquaporins. Although this can truly reflect the molecular evolution of these proteins it is more likely that these highly variable regions are due to low quality sequences. For this reason, we do not have confidence in the significant results (not shown) of the AQP10b branch of *Fundulus heteroclitus* and the AQP1b1 of *C. batrachus* and we decided to discard them. Finally, a seemingly contradictory result was found on the *Anableps anableps* AQP15 branch (Table 1). The branch-site test suggested a statistically significant event of adaptive selection but the associated omega value indicated neutral selection ($\omega=1$) and no position was found under adaptive selection (Table 1, Fig. 3). This incongruence may be directly related to the small number of AQP15 orthologues in the analysed dataset, as branch-site tests are known to not perform well with such small-size alignments (Zhang et al., 2005).

The branch-site test detected signatures of adaptive selection in mudskipper AQP11b, and on the AQP8a1, and AQP7 but could not assing them to any particular site (Table 1 and Fig.3). In contrast, our previous study (Lorente-Martínez et al., 2018) could identify several positively selected sites in mudskipper AQP10a and AQP11b. Then, we suggested the possible implication of this modifications in the terrestrialisation of these amphibious fishes. Of special interest was the positively selected site located in one of the AQP11b NPA (Asn–Pro–Ala) motifs which was substituted by a SFI (Ser-Phe-Ile) one. The two opposite NPA motifs form the pore and bind with the water molecule (Murata et al., 2000). Besides, they have been directly associated with aquaporin function because it determines which solutes can pass across the pore (Ikeda et al., 2011). We suggest that these differences can be mainly due to

taxon sampling limitations. For this study, we have assembled a more comprehensive dataset, thanks to the application of a thorough protocol (Lorente-Martínez et al., 2022) which may have helped in the finding of aquaporin paralogues and copies previously overlooked (Fig. 1). This dataset includes two additional species of gobiies (Gobioidei: Gobiidae), *Neogobius melanostomus* and *Lesueurigobius sanzi* as well as a representative of Apogonidae (*Sphaeramia orbicularis*) as their sister group (Agorreta et al., 2013; Simonovic et al., 2001). Therefore, to trace the selection more precisely over the AQP10a and AQP11b orthologues we tested the stem branch of all the Gobiidae family (as represented in our dataset). We did not find any signal of adaptive selection on the AQP10a stem branch of the Gobiidae family, suggesting that our earlier result appears to have been spurious or artefactual, likely influenced by the lower density of the taxon sampling. In the 2018 dataset, only two sequences of fish AQP10a could be included (apart from the mudskippers themselves), and as we discussed before the reliability of these results is fairly low. The same explanation could be applied to the AQP7 class (with adaptive selection detected here, but not in the 2018 study) as it has much denser sampling in the present study. In the case of the AQP8 class, adaptive selection had not been detected in the 2018 study (as it is here) because this paralogue could not be catalogued in mudskippers before (Fig. 1). On the other hand, our result indicates that the Gobiidae AQP11b stem branch have experimented adaptive selection and the same NPA motif as in our previous study reported signals of adaptive selection (Table S2 and Fig. 3). This result suggests that the modification of the NPA motif occurred before the origin of the mudskipper clade, being an apomorphy of the entire Gobiidae (or even Gobioidei) clade that is not present in the sister group Apogonidae. Even though further research is needed, this might be a case of exaptation (Gould & Vrba, 1982. In fact, mudskippers terrestrialisation seems to have occurred in two different events during their evolution (Steppan et al., 2022). Maybe the characteristics that allowed their amphibious lifestyle could have arisen before their adaptation to mudflats facilitating these convergent events.

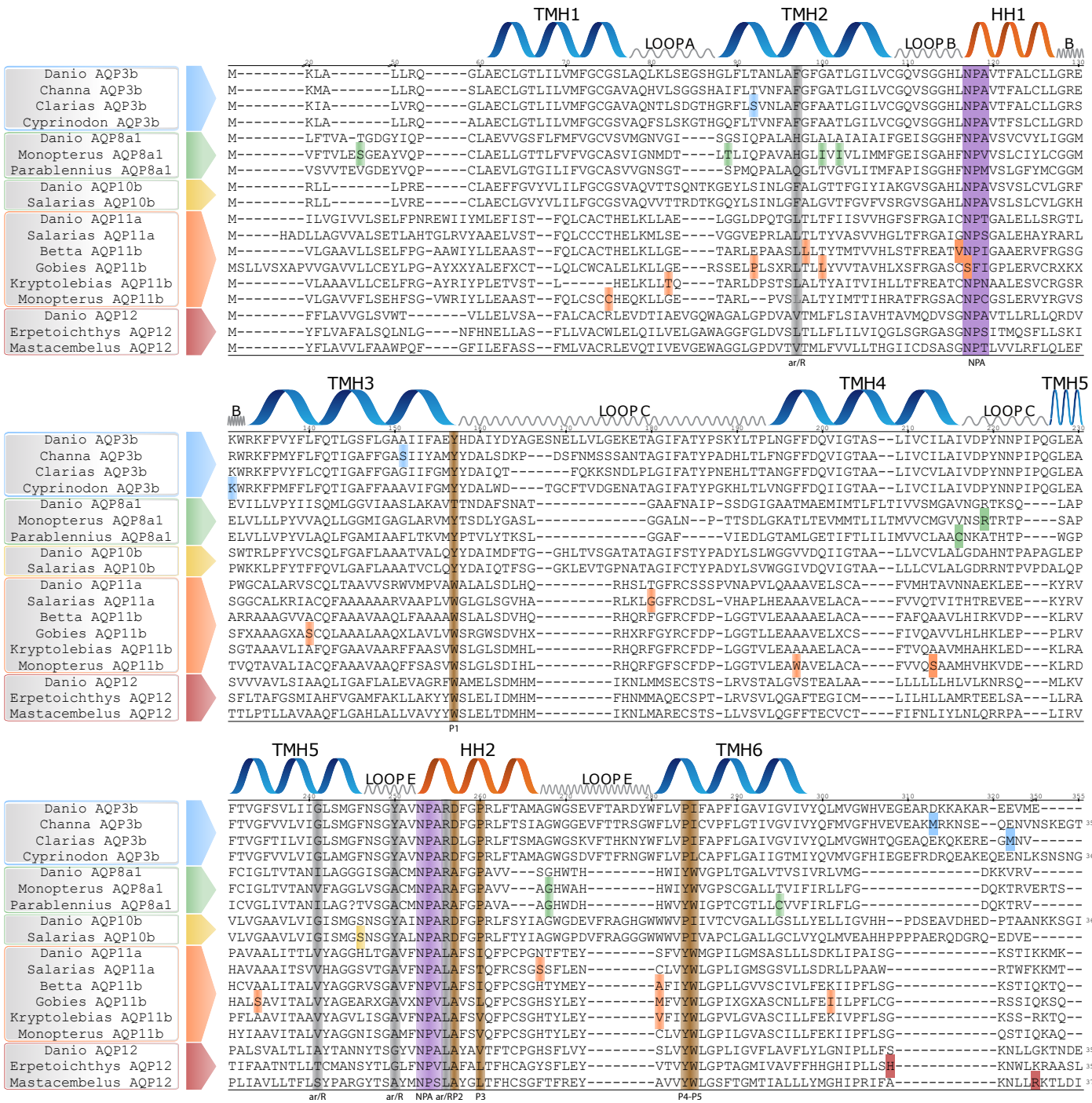


Figure 3. Sequence alignment and structural annotation of positively selected aquaporins. The sequences of the corresponding paralogs of *Danio rerio* are included as reference. The transmembrane helix (TM1-6; blue), the hemi-helices (HH1-HH2; orange) and loops A-E (grey), are annotated for *D. rerio* Aqp10b based on a molecular sequence wrap of the crystallographically resolved structure. NPA motifs (purple), ar/R selectivity filters (grey) and sites reported to confer glycerol selectivity (P1–P5) (blue) are highlighted following (Abascal et al., 2014).

Outside the Gobiidae clade, up to 28 different AQP positions could have undergone adaptive selection in 12 different branches (Table 1 and Fig. 3). Among these, another five AQP11 branches were detected as under adaptive selection making this AQP class a highly promising candidate for being involved in the adaptation to an amphibious lifestyle of the ancestors of the studied fishes. AQP11 forms, together with the AQP12, the superaquaporins or unorthodox clade (Figs. 2 and S1; Abascal et al., 2014; Finn et al., 2014). It has been proved that AQP11 is fully capable of water transport (Yakata et al., 2011) and that can be related with osmoregulation during seawater acclimation (e.g. Kim et al., 2014; Ma et al., 2020). Furthermore, AQP11 can transport H_2O_2 and has been associated with the reduction of cellular stress in the endoplasmic reticulum (Bestetti et al., 2020; Ishibashi et al., 2021; Yakata et al., 2011). However, whether AQP11 can deal with the rise of ROS, and therefore oxidative stress, when fishes adapt to land and air-breathing conditions (Pelster & Wood, 2018) and their specific function or the pattern of expression of these sequences remains unknown.

The molecular structure of aquaporins comprises six α-helices connected by five loops (Fig. 4; Cheng et al., 1997; Walz et al., 1997). These proteins tetramerise and form pores five pores (one each monomer plus the central one) in the cell membranes (Heymann & Engel, 1999; Walz et al., 1997). Based on this it is plausible to suggest that the conformational changes around or in the NPA motifs could have a relevant influence in the final structure, and therefore in the function of this proteins. In terms of solute recognition, the aromatic/arginine (ar/R) selectivity filters and the differentially conserved amino acids in aquaglyceroporins (P1–P5, Fig. 3) have been described so far as the most important motifs (Abascal et al., 2014; Heymann & Engel, 2000; Sui et al., 2001). Most of these positions map in the external half of the aquaporin molecule suggesting that this region is mainly involved in solute specificity. Our results did not find any signal of adaptive selection at any of above-mentioned specific sites, but another two positively selected sites were located on this half of the molecule (*Salarias fasciatus* AQP10b and 11a; Fig. 4).

It is on the cytoplasmatic half of the AQP molecule where most of the regulation processes occur (reviewed in Törnroth-Horsefield et al., 2010). Aquaporin function can be regulated by gating and trafficking especially on the cytoplasmatic c-terminal region (reviewed in Kreida & Törnroth-Horsefield, 2015). A few positively selected sites in this region were detected on two different branches (AQP3 and 12), and could be related with changes in AQP regulation. Altogether, all these results suggest that point

mutations under adaptive selection could have modified AQP function and regulation, as well as solute recognition during the evolutionary history of different lineages of teleosts and have facilitated the convergent transition to amphibious lifestyle. However, how these substitutions, especially the one detected in the first NPA motif of the Gobiidae clade, may have modified the molecular structure of these proteins and thus, their function, needs further research and a better understanding of the 3D structure of every AQP class.

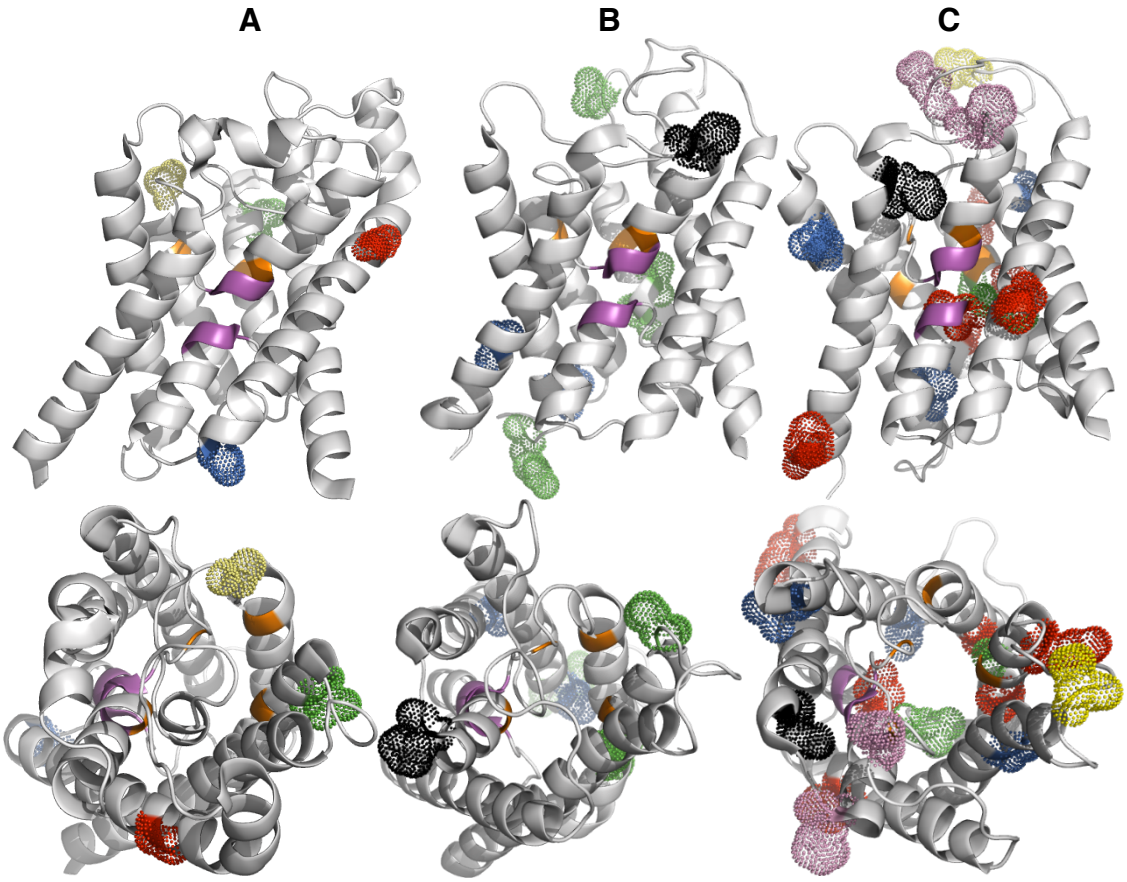


Figure 4. Structural view of positively -selected positions of the studied amphibious fishes. NPA boxes (purple) and ar/R filter sites (orange) (A) AQP3 and AQP10 paralogues sites are mapped onto the three-dimensional structure of *Homo sapiens* AQP10 (PDB ID: 6F7H). Blue dots correspond to *Cyprinodon variegatus* AQP3b, green to *Clarias batrachus* AQP3b branches, reds to *Channa argus* AQP3b, and yellow to *Salarias fasciatus* AQP10b. (B) AQP8 paralogue sites are mapped onto the three-dimensional structure of *H. sapiens* AQP1 (PDB ID: 1H6I). Blue dots correspond to *Parablennius parvicornis* AQP8a1 and green to *Monopterus albus* AQP8a1 branches. Black dots correspond to a shared position between *P. parvicornis* and *M. albus* AQP8a1 branches. (C) Sites of AQP11 paralogue mapped onto the three-dimensional structure of *H. sapiens* AQP1 (PDB ID: 1H6I). Blue dots correspond to the following branches:

M. albus AQP11b, green to *Betta splendens* AQP11b, pink to *S. fasciatus* AQP11a, red to the Gobiidae clade AQP11b, and yellow to *Kryptolebias marmoratus* AQP11b. Black dots correspond to a shared position among *B. splendens*, *K. marmoratus* and the Gobiidae clade AQP11b branches.

Few studies have focused on how AQPs function during emersion in teleost fishes. In *F. heteroclitus* embryos an AQP3 sequence seems to be downregulated during aerial exposure, likely to reduce water loss (Tingaud-Sequeira et al., 2009). Similarly, in lungfishes, a kidney AQP2-like (homologous to the tetrapod-exclusive AQP2) seems to be upregulated during aestivation (Finn et al., 2014; Konno et al., 2009, 2010). In emersed *A. testudineus* an AQP1 sequence is upregulated in the gills (Ip et al., 2013), although this paralogue has not been associated with osmoregulation but with nitrogen metabolism. Even though our dataset comprised sequences from these two paralogues, no signal of adaptive selection was detected on these branches. In the case of lungfishes, in our previous study we detected adaptive selection on the Sarcopterygii AQP2-like stem branch (Actinistia sequences plus tetrapods AQP2, 5 and 6 subfamilies). Therefore, it seems that the apomorphy, that may have facilitated the conquest of land in this group, could have occurred very early in the evolution of sarcopterygian fishes. In this sense, Finn et al. (2014) even suggested that the emergence of these AQP2-like paralogues could have occurred even before the split among Actinopterygii and Sarcopterygii fishes due to a WGD event. Finally, some amphibious fish species did not report any signal of adaptive selection on their branches (Tables 1 and S1), and this may be related to the different degrees of amphibiousness considered. For example, we did not find any branch under selection neither in *Gambusia affinis*, a species that can be considered a fully-aquatic fish although it is proven that it can leave water for predator avoidance for some minutes (Boumis et al., 2014), nor in fishes able to expend hours out of water, such as *A. anguilla* (J. B. Graham, 1997), suggesting a more complex evolutionary scenario regarding AQPs and the different instances of amphibious behaviour, which demands further research.

Conclusions

The multiple instances of adaptive selection that may have modified the molecular function and/or structure of the AQPs of amphibious fishes suggest the possible implication of these proteins in adaptation during water-to-land transition. In particular, AQP11 orthologues are the most promising candidates for further investigation within the amphibious fish framework, as almost half of the reported branches and positions under adaptive selection correspond to sequences of this class. In this regard, the adaptive selection sites in the NPA motif detected on the AQP11b stem branch of the Gobiidae clade could represent a case of exaptation. Here, a robust phylogenetic framework was reconstructed to provide a thorough bioinformatic cataloguing of AQP paralogues and to detect instances of adaptive selection. Further confirmation of the implication and actual role of detected sequence changes under selection in the adaptation to terrestrial conditions in these amphibious fishes requires the integration of our findings with physiological, biochemical, and even ecological insights.

Methods

Genomic mining and phylogenetic reconstruction

Genome and AQP sequence retrievals were performed following the protocol detailed in (Lorente-Martínez et al., 2022) against GenBank, as of June 2021. In short, BLAST searches (Altschul et al., 1990) using the BLASTX tool v2.2.28 were run locally to fish all sequence fragments that could be identified as AQPs and to isolate them from either whole genome or transcript entries, depending on the source data and level of genomic assembly and annotation (scaffold or chromosome). When available, the BLASP tool was run on protein files too, as a double-check. Once the potential AQP sequences were extracted, they were aligned at the amino acid level using the the L-INS-i algorithm in MAFFT v7.505 (Katoh et al., 2019). For sequences of genomic origin, nucleotide sequences were curated and exons were identified and translated into amino acids using Geneious Pro v9.1.8 (Kearse et al., 2012).

Best-fit models of amino acid substitution were determined using the Bayesian information criterion (BIC) in ProtTest v3.4. (Abascal et al., 2005; Darriba et al., 2011). The selected model was the site-homogeneous model JTT (Jones et al., 1992) + Γ

(Yang, 1994)+ I (Reeves, 1992). The final AQP dataset was subjected to maximum likelihood (ML) analysis using: (1) IQ-TREE with 1000 ultrafast boostrapping (UFBoot) and SH-aLRT pseudo-replicates each (Guindon et al., 2010; Hoang et al., 2018; Nguyen et al., 2015), and (2) the rapid hill-climbing algorithm from RAxML v8.2.10 (Stamatakis, 2014) computing 1000 distinct ML trees starting from 1000 distinct randomised maximum-parsimony starting trees, and with the Gamma model of rate heterogeneity for the bootstrapping phase, under 1000 rapid bootstrap replicates. The RAxML analysis was run on the CIPRES Science Gateway (Miller et al., 2010).

Positive selection analyses

Individual CDS alignments for each of the aquaporin subfamily present in our dataset were created and aligned using TranslatorX (Abascal et al., 2010) with the MAFFT algorithm. Highly partial sequences were removed, as well as alignment columns containing 5 % or more of gaps using Geneious Pro v9.1.8 (Kearse et al., 2012). Additionally, sequenes were inspected manually and highly variable regions due to low quality sequences were removed. We inferred the phylogenetic topologies of these datasets using RAxML with the GTR (Tavaré, 1986) + Γ model of nucleotide substitution and the same settings as described above, and branch lengths were corrected as substitutions per codon. In some cases, we manually corrected obvious phylogenetic artefacts to standardise the topologies according to the bony fish classification of Hughes et al. (2018).

Trees and alignments were analysed with the CODEML module of PAML v4 (Yang, 2007) Positive selection tests were performed using the branch-site Test 2 (null model MA vs. model MA; Zhang et al., 2005). We calculated Likelihood Ratio Tests (LRTs) having null MA as null hypothesis and MA as alternative hypothesis, and computed *p* values using a mixed $\chi 2$ distribution, which was obtained dividing by two the value of the $\chi 2$ distribution with one degree of freedom (Self & Liang, 1987; Yang, 2007). Due to the several tests conducted on the same tree topology, we performed multiple-test correction of the LRT *p* values, and calculated *q* values (corrected *p* values) for a false discovery rate (FDR) using the qvalue package of R (R Development Core Team, 2016). For those tests where the LRT was significant, we calculated the posterior probabilities for site classes using the Bayes Empirical Bayes (BEB; Yang et

al., 2005) implemented in PAML, identifying the specific sites under adaptive selection in every branch.

3D Structure figure

The 3D structures of the human AQP10 (6F7H) and AQP1 (1H6I) were retrieved from the Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB-PDB). The program PyMOL v1.8 (Schrödinger LLC, 2015) was used for mapping all positions under adaptive selection into the 3D structure.

Acknowledgments

This work received partial financial support from the Ministry of Science and Innovation of Spain (grant PID2020-115481GB-I00 to D.SM.). H.L-M. was sponsored by a predoctoral contract of the Complutense University of Madrid in partnership with the Real Colegio Complutense at Harvard University (RCC-UCM). Support was also provided by the Spanish network of research laboratories working on adaptation genomics (AdaptNET).

Authors' contributions

A.A. and D.S.M. conceived the study. H.L.-M. retrieved genomic data and conducted bioinformatic analyses, H.L-M. performed phylogenetic and molecular evolution analyses with the help of A.A. and D.S.M., H.L-M. wrote the first draft of the manuscript. All the authors contributed with insightful comments and approved the final manuscript for publication.

SUPPLEMENTARY MATERIAL

Table S1. List of vertebrate aquaporins and species used in this study. For each aquaporin class, the accession numbers of either GenBank (GB) or Uniprot (U) for both the amino acid sequences (AA) and the nucleotide sequences (Nu) are shown. For fish species (Actinopterygii), when applicable (AQP0, AQP1, AQP3, AQP4, AQP8, AQP9, AQP10, and AQP11), the type of aquaporin subunits are indicated as well. Species names with * indicate the species that are new for this study. Taxa highlighted in grey indicate species with some grade of amphibiousness. The assignment of fish species to Orders and Families follows Hughes et al. 2018.

Genus	Species	Class	Order	Family	AQP	GB/U AA acc.	GB/U Nu acc.
Erpetoichthys	*calabaricus**	Actinopterygii	Polypteriformes	Polypteridae	0	XP_028652034	XM_028796201
Polypterus	*senegalus**	Actinopterygii	Polypteriformes	Polypteridae	0	XP_039605857	XM_039749923
Acipenser	*ruthenus**	Actinopterygii	Acipenseriformes	Acipenseridae	0	XP_033857198	XM_034001307
Lepisosteus	*oculatus**	Actinopterygii	Lepisosteiformes	Lepisosteidae	0	XP_006629345	XM_006629282
Anguilla	*anguilla**	Actinopterygii	Anguiliformes	Anguilidae	0a	XP_035238772	XM_035382881
Scleropages	*formosus**	Actinopterygii	Osteoglossiformes	Osteoglossidae	0a	XP_018615871	XM_018760355
Astyanax	*mexicanus**	Actinopterygii	Characiformes	Characidae	0a	XP_007257935	XM_007257873
Cyprinus	*carpio*	Actinopterygii	Cypriniformes	Cyprinidae	0a	A0A0K2RVW2	LC069001
Danio	*rerio*	Actinopterygii	Cypriniformes	Danionidae	0a	Q6DEI6	BC077129
Clarias	*batrachus**	Actinopterygii	Siluriformes	Clariidae	0a		QMIH01001154
Ictalurus	*punctatus**	Actinopterygii	SIluriformes	Ictaluridae	0a	XP_017342612	XM_017487123
Pangasianodon	*hypophthalmus**	Actinopterygii	Siluriformes	Pangasiidae	0a	XP_026776844	XM_026921043
Salmo	*salar*	Actinopterygii	Salmoniformes	Salmonidae	0a	A0A185IIP7	KM823661
Gadus	*morhua**	Actinopterygii	Gadiformes	Gadidae	0a	XP_030206126	XM_030350266
Myripristis	*murdjan**	Actinopterygii	Holocentriformes	Holocentridae	0a	XP_029912510	XM_030056650

Contd...

Contd...

Contd...

Lesueurigobius	*sanzi**	Actinopterygii	Gobiiformes	Gobiidae	0a		OMNZ01055008
Neogobius	*melanostomus**	Actinopterygii	Gobiiformes	Gobiidae	0a		VHKM01000037
Boleophthalmus	*pectinirostris*	Actinopterygii	Gobiiformes	Oxudercidae	0a	XP_020785352	XM_020929693
Periophthalmodon	*schlosseri*	Actinopterygii	Gobiiformes	Oxudercidae	0a		JACM01053832
Periophthalmus	*magnuspinnatus*	Actinopterygii	Gobiiformes	Oxudercidae	0a	XP_033825049	XM_033969158
Scartelaos	*histophorus*	Actinopterygii	Gobiiformes	Oxudercidae	0a		JACN01025822
Sphaeramia	*orbicularis**	Actinopterygii	Kurtiformes	Apogonidae	0a	XP_029995313	XM_030139453
Anabas	*testudineus**	Actinopterygii	Anabantiformes	Anabantidae	0a	XP_026223245	XM_026367460
Channa	*argus**	Actinopterygii	Anabantiformes	Channidae	0a		CM015723
Helostoma	*temminki**	Actinopterygii	Anabantiformes	Helostomatidae	0a		OMLM01020935
Betta	*splendens**	Actinopterygii	Anabantiformes	Osphronemidae	0a	XP_029012598	XM_029156765
Cynoglossus	*semilaevis**	Actinopterygii	Pleuronectiformes	Cynoglossidae	0a	XP_008316434	XM_008318212
Scophthalmus	*maximus**	Actinopterygii	Pleuronectiformes	Scophthalmidae	0a	AWP09633	
Mastacembelus	*armatus**	Actinopterygii	Synbranchiformes	Mastacembelidae	0a	XP_026189508	XM_026333723
Monopterus	*albus**	Actinopterygii	Synbranchiformes	Synbranchidae	0a	XP_020474970	XM_020619314
Parablennius	*parvicornis**	Actinopterygii	Blenniiformes	Blenniidae	0a		OMNM01015415
Salarias	*fasciatus**	Actinopterygii	Blenniiformes	Blenniidae	0a	XP_029974945	XM_030119085
Gouania	*willdenowi**	Actinopterygii	Gobiesociformes	Gobiesocidae	0a	XP_028307855	XM_028452054
Oryzias	*latipes*	Actinopterygii	Beloniformes	Adrianichthyidae	0a	XP_004085170	XM_004085122
Maylandia	*zebra**	Actinopterygii	Cichliformes	Cichlidae	0a	XP_004558762	XM_004558705
Oreochromis	*niloticus**	Actinopterygii	Cichliformes	Cichlidae	0a	XP_003448200	XM_003448152

Contd...

Contd...

Genus	Species	Class	Order	Family	AQP	GB/U AA acc.	GB/U Nu acc.
Anableps	*anableps**	Actinopterygii	Cyprinodontiformes	Anablepidae	0a		CM026071
Cyprinodon	*variegatus**	Actinopterygii	Cyprinodontiformes	Cyprinodontidae	0a	XP_015246630	XM_015391144
Fundulus	*heteroclitus**	Actinopterygii	Cyprinodontiformes	Fundulidae	0a	NP_001296937	NM_001310008
Aphyosemion	*australe**	Actinopterygii	Cyprinodontiformes	Nothobranchiidae	0a		SSNS01000297
Nothobranchius	*furzeri**	Actinopterygii	Cyprinodontiformes	Nothobranchiidae	0a	XP_015823291	XM_015967805
Gambusia	*affinis**	Actinopterygii	Cyprinodontiformes	Poeciliidae	0a		NHOQ01000318
Poecilia	*reticulata**	Actinopterygii	Cyprinodontiformes	Poeciliidae	0a	XP_008411795	XM_008413573
Austrofundulus	*limnaeus**	Actinopterygii	Cyprinodontiformes	Rivulidae	0a	XP_013874938	XM_014019484
Kryptolebias	*marmoratus**	Actinopterygii	Cyprinodontiformes	Rivulidae	0a	XP_017286405	XM_017430916
Gymnodraco	*acuticeps**	Actinopterygii	Perciformes	Bathydraconidae	0a	XP_034065150	XM_034209259
Cyclopterus	*lumpus**	Actinopterygii	Perciformes	Cyclopteridae	0a	XP_034392950	XM_034537059
Notothenia	*coriiceps**	Actinopterygii	Perciformes	Nototheniidae	0a	XP_010795326	XM_010797024
Perca	*flavescens**	Actinopterygii	Perciformes	Percidae	0a	XP_028438837	XM_028583036
Sparus	*aurata*	Actinopterygii	Spariformes	Sparidae	0a	T1T0Z6	KC589385
Takifugu	*rubripes**	Actinopterygii	Tetraodontiformes	Tetraodontidae	0a	XP_003963140	XM_003963091
Anarrhichthys	*ocellatus**	Actinopterygii		Anarhichadidae	0a	XP_031700887	XM_031845027
Larimichthys	*crocea**	Actinopterygii		Sciaenidae	0a	XP_010746587	XM_010748285
Cebidichthys	*violaceus**	Actinopterygii		Stichaeidae	0a		NJBE01000438
Latimeria	*chalumnae**	Sarcopterygii			0	XP_005986314	XM_005986252
Neoceratodus	*forsteri*	Sarcopterygii			0	C7G3K6	AB513619

Contd...

Contd...

Protopterus	*annectens*	Sarcopterygii			0	C7G3K5	AB513618
Alligator	*mississippiensis*	Reptilia			0	XP_006267718	XM_006267656
Chelonia	*mydas*	Reptilia			0	XP_007064998	XM_007064936
Gallus	*gallus*	Aves			0	P28238	AY078179
Homo	*sapiens*	Mammalia			0	P30301	U36308
Mus	*musculus*	Mammalia			0	P51180	U27502
Anguilla	*anguilla**	Actinopterygii	Anguiliformes	Anguilidae	0b	XP_035244846	XM_035388955
Scleropages	*formosus**	Actinopterygii	Osteoglossiformes	Osteoglossidae	0b	XP_018580792	XM_018725276
Astyanax	*mexicanus**	Actinopterygii	Characiformes	Characidae	0b	XP_007233064	XM_007233002
Cyprinus	*carpio*	Actinopterygii	Cypriniformes	Cyprinidae	0b	A0A0K2RVS7	LC068999
Danio	*rerio*	Actinopterygii	Cypriniformes	Danionidae	0b	Q4ZJI3	DQ003080
Salmo	*salar*	Actinopterygii	Salmoniformes	Salmonidae	0b	A0A0K0LTP0	KM677199
Gadus	*morhua**	Actinopterygii	Gadiformes	Gadidae	0b	XP_030208548	XM_030352688
Myripristis	*murdjan**	Actinopterygii	Holocentriformes	Holocentridae	0b	XP_029911600	XM_030055740
Lesueurigobius	*sanzi**	Actinopterygii	Gobiiformes	Gobiidae	0b		OMNZ01100408
Neogobius	*melanostomus**	Actinopterygii	Gobiiformes	Gobiidae	0b		VHKM01000223
Boleophthalmus	*pectinirostris*	Actinopterygii	Gobiiformes	Oxudercidae	0b	XP_020785343	XM_020929684
Periophthalmodon	*schlosseri*	Actinopterygii	Gobiiformes	Oxudercidae	0b		KN475480
Periophthalmus	*magnuspinnatus*	Actinopterygii	Gobiiformes	Oxudercidae	0b	XP_033826145	XM_033970254
Scartelaos	*histophorus*	Actinopterygii	Gobiiformes	Oxudercidae	0b		KN491114
Sphaeramia	*orbicularis**	Actinopterygii	Kurtiformes	Apogonidae	0b	XP_029995109	XM_030139249

Contd...

Contd...

Genus	Species	Class	Order	Family	AQP	GB/U AA acc.	GB/U Nu acc.
Anabas	*testudineus**	Actinopterygii	Anabantiformes	Anabantidae	0b	XP_026224172	XM_026368387
Helostoma	*temminki**	Actinopterygii	Anabantiformes	Helostomatidae	0b		OMLM01005575
Betta	*splendens**	Actinopterygii	Anabantiformes	Osphronemidae	0b	XP_029012285	XM_029156452
Cynoglossus	*semilaevis**	Actinopterygii	Pleuronectiformes	Cynoglossidae	0b	XP_008317523	XM_008319301
Scophthalmus	*maximus**	Actinopterygii	Pleuronectiformes	Scophthalmidae	0b	AWP10244	
Mastacembelus	*armatus**	Actinopterygii	Synbranchiformes	Mastacembelidae	0b	XP_026187154_1	XM_026331369
Monopterus	*albus**	Actinopterygii	Synbranchiformes	Synbranchidae	0b	XP_020466424	XM_020610768
Parablennius	*parvicornis**	Actinopterygii	Blenniiformes	Blenniidae	0b		OMNM01004137
Salarias	*fasciatus**	Actinopterygii	Blenniiformes	Blenniidae	0b	XP_029974481	XM_030118621
Gouania	*willdenowi**	Actinopterygii	Gobiesociformes	Gobiesocidae	0b	XP_028308729	XM_028452928
Oryzias	*latipes*	Actinopterygii	Beloniformes	Adrianichthyidae	0b	XP_004070852	XM_004070804
Maylandia	*zebra**	Actinopterygii	Cichliformes	Cichlidae	0b	XP_004546271	XM_004546214
Oreochromis	*niloticus**	Actinopterygii	Cichliformes	Cichlidae	0b	XP_003439024	XM_003438976
Anableps	*anableps**	Actinopterygii	Cyprinodontiformes	Anablepidae	0b	CM026071	
Cyprinodon	*variegatus**	Actinopterygii	Cyprinodontiformes	Cyprinodontidae	0b	XP_015239028	XM_015383542
Fundulus	*heteroclitus**	Actinopterygii	Cyprinodontiformes	Fundulidae	0b	XP_012721399	XM_012865945
Aphyosemion	*australe**	Actinopterygii	Cyprinodontiformes	Nothobranchiidae	0b		SSNS01000521
Nothobranchius	*furzeri**	Actinopterygii	Cyprinodontiformes	Nothobranchiidae	0b	XP_015823775	XM_015968289
Gambusia	*affinis**	Actinopterygii	Cyprinodontiformes	Poeciliidae	0b		NHOQ01000244
Poecilia	*reticulata**	Actinopterygii	Cyprinodontiformes	Poeciliidae	0b	XP_008412384	XM_008414162

Contd...

Contd...

Austrofundulus	*limnaeus**	Actinopterygii	Cyprinodontiformes	Rivulidae	0b	XP_013862725	XM_014007271
Kryptolebias	*marmoratus**	Actinopterygii	Cyprinodontiformes	Rivulidae	0b	XP_017264869	XM_017409380
Gymnodraco	*acuticeps**	Actinopterygii	Perciformes	Bathydraconidae	0b	XP_034066832	XM_034210941
Cyclopterus	*lumpus**	Actinopterygii	Perciformes	Cyclopteridae	0b	XP_034392941	XM_034537050
Notothenia	*coriiceps**	Actinopterygii	Perciformes	Nototheniidae	0b	XP_010776695	XM_010778393
Perca	*flavescens**	Actinopterygii	Perciformes	Percidae	0b	XP_028438860	XM_028583059
Takifugu	*rubripes**	Actinopterygii	Tetraodontiformes	Tetraodontidae	0b	XP_003963475	XM_003963426
Anarrhichthys	*ocellatus**	Actinopterygii		Anarhichadidae	0b	XP_031700269	XM_031844409
Larimichthys	*crocea**	Actinopterygii		Sciaenidae	0b	XP_010730003	XM_010731701
Cebidichthys	*violaceus**	Actinopterygii		Stichaeidae	0b		NJBE01000438
Erpetoichthys	calabaricus*	Actinopterygii	Polypteriformes	Polypteridae	1	XP_028660019	XM_028804186
Polypterus	senegalus*	Actinopterygii	Polypteriformes	Polypteridae	1	XP_039609606	XM_039753672
Acipenser	ruthenus*	Actinopterygii	Acipenseriformes	Acipenseridae	1	XP_03385640	XM_034000517
Lepisosteus	oculatus*	Actinopterygii	Lepisosteiformes	Lepisosteidae	1	XP_006634313	XM_006634250
Anguilla	anguilla*	Actinopterygii	Anguiliformes	Anguilidae	1a	XP_035287273	XM_035431382
Scleropages	formosus*	Actinopterygii	Osteoglossiformes	Osteoglossidae	1a	A0A0P7YCC1	JARO02007517
Astyanax	*mexicanus**	Actinopterygii	Characiformes	Characidae	1a	XP_007242700	XM_007242638
Cyprinus	*carpio*	Actinopterygii	Cypriniformes	Cyprinidae	1a	A0A0K2RVW0	LC069005
Danio	*rerio*	Actinopterygii	Cypriniformes	Danionidae	1a	NP_996942	NM_207059
Clarias	*batrachus**	Actinopterygii	Siluriformes	Clariidae	1a		QMIH01000959
Ictalurus	*punctatus**	Actinopterygii	SIluriformes	Ictaluridae	1a	XP_017326545	XM_017471056

Contd...

Contd...

Genus	Species	Class	Order	Family	AQP	GB/U AA acc.	GB/U Nu acc.
Pangasianodon	*hypophthalmus**	Actinopterygii	Siluriformes	Pangasiidae	1a	XP_026766072	XM_026910271
Esox	*lucius**	Actinopterygii	Esociformes	Esocidae	1a	NP_001291145	NM_001304216
Salmo	*salar*	Actinopterygii	Salmoniformes	Salmonidae	1a	XP_013996040	XM_014140565
Gadus	*morhua**	Actinopterygii	Gadiformes	Gadidae	1a	XP_030219980	XM_030364120
Myripristis	*murdjan**	Actinopterygii	Holocentriformes	Holocentridae	1a	XP_029931156	XM_030075296
Lesueurigobius	*sanzi**	Actinopterygii	Gobiiformes	Gobiidae	1a		OMNZ01034392
Neogobius	*melanostomus**	Actinopterygii	Gobiiformes	Gobiidae	1a		VHKM01000016
Boleophthalmus	*pectinirostris*	Actinopterygii	Gobiiformes	Oxudercidae	1a	XP_020791255	XM_020935596
Periophthalmodon	*schlosseri*	Actinopterygii	Gobiiformes	Oxudercidae	1a		KN483096
Periophthalmus	*magnuspinnatus*	Actinopterygii	Gobiiformes	Oxudercidae	1a	XP_033838364	XM_033982473
Scartelaos	*histophorus*	Actinopterygii	Gobiiformes	Oxudercidae	1a		KN502285
Sphaeramia	*orbicularis**	Actinopterygii	Kurtiformes	Apogonidae	1a	XP_030016897	XM_030161037
Anabas	*testudineus**	Actinopterygii	Anabantiformes	Anabantidae	1a	XP_0262311881	XM_026375403
Channa	*argus**	Actinopterygii	Anabantiformes	Channidae	1a		CM015725
Helostoma	*temminki**	Actinopterygii	Anabantiformes	Helostomatidae	1a		OMLM01009264
Betta	*splendens**	Actinopterygii	Anabantiformes	Osphronemidae	1a	XP_028988674	XM_029132841
Cynoglossus	*semilaevis**	Actinopterygii	Pleuronectiformes	Cynoglossidae	1a	XP_008332666	XM_008334444
Scophthalmus	*maximus**	Actinopterygii	Pleuronectiformes	Scophthalmidae	1a	AWP03918	
Mastacembelus	*armatus**	Actinopterygii	Synbranchiformes	Mastacembelidae	1a	XP_026168860	XM_026313075
Monopterus	*albus**	Actinopterygii	Synbranchiformes	Synbranchidae	1a	XP_020447141	XM_020591485

Contd...

Contd...

Parablennius	*parvicornis**	Actinopterygii	Blenniiformes	Blenniidae	1a		OMNM01055971
Salarias	*fasciatus**	Actinopterygii	Blenniiformes	Blenniidae	1a	XP_029970825	XM_030114965
Gouania	*willdenowi**	Actinopterygii	Gobiesociformes	Gobiesocidae	1a	XP_028328824	XM_028473023
Oryzias	*latipes*	Actinopterygii	Beloniformes	Adrianichthyidae	1a	XP_011485314	XM_011487012
Maylandia	*zebra**	Actinopterygii	Cichliformes	Cichlidae	1a	XP_004542904	XM_004542847
Oreochromis	*niloticus**	Actinopterygii	Cichliformes	Cichlidae	1a	XP_003438133	XM_003438085
Anableps	*anableps**	Actinopterygii	Cyprinodontiformes	Anablepidae	1a		CM026077
Cyprinodon	*variegatus**	Actinopterygii	Cyprinodontiformes	Cyprinodontidae	1a	XP_015225343	XM_015369857
Fundulus	*heteroclitus**	Actinopterygii	Cyprinodontiformes	Fundulidae	1a	NP_001296903	NM_001309974
Aphyosemion	*australe**	Actinopterygii	Cyprinodontiformes	Nothobranchiidae	1a		SSNS01000142
Nothobranchius	*furzeri**	Actinopterygii	Cyprinodontiformes	Nothobranchiidae	1a	XP_015812162	XM_015956676
Gambusia	*affinis**	Actinopterygii	Cyprinodontiformes	Poeciliidae	1a		NHOQ01001318
Poecilia	*reticulata**	Actinopterygii	Cyprinodontiformes	Poeciliidae	1a	XP_008432534	XM_008434312
Austrofundulus	*limnaeus**	Actinopterygii	Cyprinodontiformes	Rivulidae	1a	XP_013884868	XM_014029414
Kryptolebias	*marmoratus**	Actinopterygii	Cyprinodontiformes	Rivulidae	1a	XP_0172814311	XM_017425942
Gymnodraco	*acuticeps**	Actinopterygii	Perciformes	Bathydraconidae	1a	XP_034090637	XM_034234746
Cyclopterus	*lumpus**	Actinopterygii	Perciformes	Cyclopteridae	1a	XP_034411769	XM_034555878
Notothenia	*coriiceps**	Actinopterygii	Perciformes	Nototheniidae	1a	XP_010765944	XM_010767642
Perca	*flavescens**	Actinopterygii	Perciformes	Percidae	1a	XP_028448701	XM_028592900
Sparus	*aurata*	Actinopterygii	Spariformes	Sparidae	1a	Q32ZE5	AY626939
Takifugu	*rubripes**	Actinopterygii	Tetraodontiformes	Tetraodontidae	1a	XP_003975375	XM_003975326

Contd...

Contd...

Genus	Species	Class	Order	Family	AQP	GB/U AA acc.	GB/U Nu acc.
Anarrhichthys	*ocellatus**	Actinopterygii		Anarhichadidae	1a	XP_031721097	XM_031865237
Larimichthys	*crocea**	Actinopterygii		Sciaenidae	1a	XP_010727519	XM_010729217
Cebidichthys	*violaceus**	Actinopterygii		Stichaeidae	1a		NJBE01000440
Latimeria	*chalumnae**	Sarcopterygii			1	XP_006005961	XM_006005899
Xenopus	*laevis*	Amphibia			1	NP_001085391	NM_001091922
Alligator	*mississippiensis*	Reptilia			1	XP_006262158	XM_006262096
Chelonia	*mydas*	Reptilia			1	XP_027684664	XM_027828863
Gallus	*gallus*	Aves			1	NP_001034542	NM_001039453
Homo	*sapiens*	Mammalia			1	NP_932766	NM_198098
Mus	*musculus*	Mammalia			1	NP_031498	NM_007472
Anguilla	*anguilla**	Actinopterygii	Anguiliformes	Anguilidae	1b1	XP_035287271	XM_035431380
Clarias	*batrachus**	Actinopterygii	Siluriformes	Clariidae	1b1		QMIH01000959
Ictalurus	*punctatus**	Actinopterygii	SIluriformes	Ictaluridae	1b1	XP_017326543	XM_017471054
Pangasianodon	*hypophthalmus**	Actinopterygii	Siluriformes	Pangasiidae	1b1	XP_026766026	XM_026910225
Myripristis	*murdjan**	Actinopterygii	Holocentriformes	Holocentridae	1b1	XP_029930897	XM_030075037
Cynoglossus	*semilaevis**	Actinopterygii	Pleuronectiformes	Cynoglossidae	1b1	XP_008332667	XM_025066020
Maylandia	*zebra**	Actinopterygii	Cichliformes	Cichlidae	1b1	XP_004542906	XM_004542849
Oreochromis	*niloticus**	Actinopterygii	Cichliformes	Cichlidae	1b1	XP_003438132	XM_003438084
Austrofundulus	*limnaeus**	Actinopterygii	Cyprinodontiformes	Rivulidae	1b1	XP_013884876	XM_014029422
Kryptolebias	*marmoratus**	Actinopterygii	Cyprinodontiformes	Rivulidae	1b1	XP_0172815591	XM_017426070

Contd...

Contd...

Anarrhichthys	*ocellatus**	Actinopterygii		Anarhichadidae	1b1	XP_031721141	XM_031865281
Cyprinus	*carpio*	Actinopterygii	Cypriniformes	Cyprinidae	1b2	A0A0K2RVW6	LC069006
Danio	*rerio*	Actinopterygii	Cypriniformes	Danionidae	1b2	NP_001129154	NM_001135682
Clarias	*batrachus**	Actinopterygii	Siluriformes	Clariidae	1b2		QMIH01000959
Ictalurus	*punctatus**	Actinopterygii	SIluriformes	Ictaluridae	1b2	XP_017326544	XM_017471055
Esox	*lucius**	Actinopterygii	Esociformes	Esocidae	1b2	XP_010884724	XM_010886422
Salmo	*salar*	Actinopterygii	Salmoniformes	Salmonidae	1b2	NP_001133472	NM_001140000
Myripristis	*murdjan**	Actinopterygii	Holocentriformes	Holocentridae	1b2	XP_029930896	XM_030075036
Lesueurigobius	*sanzi**	Actinopterygii	Gobiiformes	Gobiidae	1b2		OMNZ01051794
Neogobius	*melanostomus**	Actinopterygii	Gobiiformes	Gobiidae	1b2		VHKM01000016
Boleophthalmus	*pectinirostris*	Actinopterygii	Gobiiformes	Oxudercidae	1b2	XP_020791256	XM_020935597
Periophthalmodon	*schlosseri*	Actinopterygii	Gobiiformes	Oxudercidae	1b2		KN483096
Periophthalmus	*magnuspinnatus*	Actinopterygii	Gobiiformes	Oxudercidae	1b2	XP_033838366	XM_033982475
Scartelaos	*histophorus*	Actinopterygii	Gobiiformes	Oxudercidae	1b2		JACN01018430
Sphaeramia	*orbicularis**	Actinopterygii	Kurtiformes	Apogonidae	1b2	XP_030016676	XM_030160816
Anabas	*testudineus**	Actinopterygii	Anabantiformes	Anabantidae	1b2	XP_0262309711	XM_026375186
Helostoma	*temminki**	Actinopterygii	Anabantiformes	Helostomatidae	1b2		OMLM01009264
Salarias	*fasciatus**	Actinopterygii	Blenniiformes	Blenniidae	1b2	XP_029970824	XM_030114964
Maylandia	*zebra**	Actinopterygii	Cichliformes	Cichlidae	1b2	XP_004542905	XM_004542848
Oreochromis	*niloticus**	Actinopterygii	Cichliformes	Cichlidae	1b2	XP_003438131	XM_003438083
Sparus	*aurata*	Actinopterygii	Spariformes	Sparidae	1b2	Q32ZE6	AY626938

Contd...

Contd...

Genus	Species	Class	Order	Family	AQP	GB/U AA acc.	GB/U Nu acc.
Takifugu	*rubripes**	Actinopterygii	Tetraodontiformes	Tetraodontidae	1b2	XP_003975420	XM_003975371
Larimichthys	*crocea**	Actinopterygii		Sciaenidae	1b2	XP_010727520	XM_010729218
Latimeria	*chalumnae**	Sarcopterygii			2	XP_005986215	XM_005986153
Neoceratodus	*forsteri*	Sarcopterygii			2	BAH98063	AB513620
Protopterus	*annectens*	Sarcopterygii			2	BAH86607	AB474277
Xenopus	*laevis*	Amphibia			2	NP_001079331	NM_001085862
Alligator	*mississippiensis*	Reptilia			2	XP_006275751	XM_006275689
Chelonia	*mydas*	Reptilia			2	XP_007069977	XM_007069915
Gallus	*gallus*	Aves			2	NP_001279001	NM_001292072
Homo	*sapiens*	Mammalia			2	NP_000477	NM_000486
Mus	*musculus*	Mammalia			2	NP_033829	NM_009699
Erpetoichthys	*calabaricus**	Actinopterygii	Polypteriformes	Polypteridae	3	XP_028660985	XM_028805152
Polypterus	*senegalus**	Actinopterygii	Polypteriformes	Polypteridae	3	XP_039615087	XM_039759153
Acipenser	*ruthenus**	Actinopterygii	Acipenseriformes	Acipenseridae	3	XP_033862580	XM_034006689
Lepisosteus	*oculatus**	Actinopterygii	Lepisosteiformes	Lepisosteidae	3	XP_006626763	XM_015363296
Anguilla	*anguilla**	Actinopterygii	Anguiliformes	Anguilidae	3a	XP_035245855	XM_035389964
Astyanax	*mexicanus**	Actinopterygii	Characiformes	Characidae	3a	XP_007238017	XM_007237955
Cyprinus	*carpio*	Actinopterygii	Cypriniformes	Cyprinidae	3a	A0A0K2RVW4	
Danio	*rerio*	Actinopterygii	Cypriniformes	Danionidae	3a	NP_001159593	NM_001166121
Clarias	*batrachus**	Actinopterygii	Siluriformes	Clariidae	3a		QMIH01000866

Contd...

Contd...

Ictalurus	*punctatus**	Actinopterygii	SIluriformes	Ictaluridae	3a	XP_017348706	XM_017493217
Pangasianodon	*hypophthalmus**	Actinopterygii	Siluriformes	Pangasiidae	3a	XP_026784818	XM_026929017
Salmo	*salar*	Actinopterygii	Salmoniformes	Salmonidae	3a	XP_013992403	XM_014136928
Gadus	*morhua**	Actinopterygii	Gadiformes	Gadidae	3a	XP_030209598	XM_030353738
Myripristis	*murdjan**	Actinopterygii	Holocentriformes	Holocentridae	3a	XP_029921125	XM_030065265
Cynoglossus	*semilaevis**	Actinopterygii	Pleuronectiformes	Cynoglossidae	3a	XP_008323950	XM_008325728
Salarias	*fasciatus**	Actinopterygii	Blenniiformes	Blenniidae	3a	XP_029940702	XM_030084842
Notothenia	*coriiceps**	Actinopterygii	Perciformes	Nototheniidae	3a	XP_010778437	XM_010780135
Anarrhichthys	*ocellatus**	Actinopterygii		Anarhichadidae	3a	XP_031715974	XM_031860114
Larimichthys	*crocea**	Actinopterygii		Sciaenidae	3a	XP_019133927	XM_019278382
Cebidichthys	*violaceus**	Actinopterygii		Stichaeidae	3a		NJBE01000445
Latimeria	*chalumnae**	Sarcopterygii			3	XP_006004793	XM_006004731
Protopterus	*annectens*	Sarcopterygii			3	D0FZC5	KX494981
Xenopus	*laevis*	Amphibia			3	Q9YH65	AJ131847
Alligator	*mississippiensis*	Reptilia			3	XP_006265318	XM_006265256
Chelonia	*mydas*	Reptilia			3	XP_007063360	XM_007063298
Gallus	*gallus*	Aves			3	XP_424500	XM_424500
Homo	*sapiens*	Mammalia			3	NP_004916	NM_004925
Mus	*musculus*	Mammalia			3	NP_057898	NM_016689
Anguilla	*anguilla**	Actinopterygii	Anguiliformes	Anguilidae	3b	XP_035234834	XM_035378943
Scleropages	*formosus**	Actinopterygii	Osteoglossiformes	Osteoglossidae	3b	XP_018592159	XM_018736643

Contd...

Contd...

Genus	Species	Class	Order	Family	AQP	GB/U AA acc.	GB/U Nu acc.
Astyanax	*mexicanus**	Actinopterygii	Characiformes	Characidae	3b	XP_007258373	XM_007258311
Cyprinus	*carpio*	Actinopterygii	Cypriniformes	Cyprinidae	3b	XP_018939297	XM_019083752
Danio	*rerio*	Actinopterygii	Cypriniformes	Danionidae	3b	NP_998633	NM_213468
Clarias	*batrachus**	Actinopterygii	Siluriformes	Clariidae	3b		QMIH01000977
Ictalurus	*punctatus**	Actinopterygii	SIluriformes	Ictaluridae	3b	XP_017307685	XM_017452196
Pangasianodon	*hypophthalmus**	Actinopterygii	Siluriformes	Pangasiidae	3b	XP_026787255	XM_026931454
Esox	*lucius**	Actinopterygii	Esociformes	Esocidae	3b	XP_010903750	XM_010905448
Salmo	*salar*	Actinopterygii	Salmoniformes	Salmonidae	3b	XP_014016368	XM_014160893
Gadus	*morhua**	Actinopterygii	Gadiformes	Gadidae	3b	XP_030215680	XM_030359820
Myripristis	*murdjan**	Actinopterygii	Holocentriformes	Holocentridae	3b	XP_029916894	XM_030061034
Lesueurigobius	*sanzi**	Actinopterygii	Gobiiformes	Gobiidae	3b		OMNZ01043962
Neogobius	*melanostomus**	Actinopterygii	Gobiiformes	Gobiidae	3b		VHKM01000006
Boleophthalmus	*pectinirostris*	Actinopterygii	Gobiiformes	Oxudercidae	3b	XP_020791696	XM_020936037
Periophthalmodon	*schlosseri*	Actinopterygii	Gobiiformes	Oxudercidae	3b		KN469198
Periophthalmus	*magnuspinnatus*	Actinopterygii	Gobiiformes	Oxudercidae	3b	XP_033827909	XM_033972018
Scartelaos	*histophorus*	Actinopterygii	Gobiiformes	Oxudercidae	3b		KN514423
Sphaeramia	*orbicularis**	Actinopterygii	Kurtiformes	Apogonidae	3b	XP_030000487	XM_030144627
Anabas	*testudineus**	Actinopterygii	Anabantiformes	Anabantidae	3b	XP_0261995741	XM_026343789
Channa	*argus**	Actinopterygii	Anabantiformes	Channidae	3b	CM015722	
Helostoma	*temminki**	Actinopterygii	Anabantiformes	Helostomatidae	3b		OMLM01009369

Contd...

Contd...

Betta	*splendens**	Actinopterygii	Anabantiformes	Osphronemidae	3b	XP_029019612	XM_029163779
Scophthalmus	*maximus**	Actinopterygii	Pleuronectiformes	Scophthalmidae	3b		AWP06911
Mastacembelus	*armatus**	Actinopterygii	Synbranchiformes	Mastacembelidae	3b	XP_026178122	XM_026322337
Monopterus	*albus**	Actinopterygii	Synbranchiformes	Synbranchidae	3b	XP_020450150	XM_020594494
Parablennius	*parvicornis**	Actinopterygii	Blenniiformes	Blenniidae	3b		OMNM01035729
Salarias	*fasciatus**	Actinopterygii	Blenniiformes	Blenniidae	3b	XP_029961271	XM_030105411
Gouania	*willdenowi**	Actinopterygii	Gobiesociformes	Gobiesocidae	3b	XP_028313777	XM_028457976
Oryzias	*latipes*	Actinopterygii	Beloniformes	Adrianichthyidae	3b	XP_004072505	XM_004072457
Maylandia	*zebra**	Actinopterygii	Cichliformes	Cichlidae	3b1	XP_004544455	XM_004544398
Oreochromis	*niloticus**	Actinopterygii	Cichliformes	Cichlidae	3b1	NP_001298262	NM_001311333
Maylandia	*zebra**	Actinopterygii	Cichliformes	Cichlidae	3b2	XP_014269509	XM_014414023
Oreochromis	*niloticus**	Actinopterygii	Cichliformes	Cichlidae	3b2	XP_013120054	XM_013264600
Anableps	*anableps**	Actinopterygii	Cyprinodontiformes	Anablepidae	3b		CM026066
Cyprinodon	*variegatus**	Actinopterygii	Cyprinodontiformes	Cyprinodontidae	3b	XP_015259478	XM_015403992
Fundulus	*heteroclitus**	Actinopterygii	Cyprinodontiformes	Fundulidae	3b	NP_001296892	NM_001309963
Aphyosemion	*australe**	Actinopterygii	Cyprinodontiformes	Nothobranchiidae	3b		SSNS01000094
Nothobranchius	*furzeri**	Actinopterygii	Cyprinodontiformes	Nothobranchiidae	3b	XP_015827941	XM_015972455
Gambusia	*affinis**	Actinopterygii	Cyprinodontiformes	Poeciliidae	3b		NHOQ01002094
Poecilia	*reticulata**	Actinopterygii	Cyprinodontiformes	Poeciliidae	3b	XP_008417140	XM_008418918
Austrofundulus	*limnaeus**	Actinopterygii	Cyprinodontiformes	Rivulidae	3b	XP_013863824	XM_014008370
Kryptolebias	*marmoratus**	Actinopterygii	Cyprinodontiformes	Rivulidae	3b	XP_017283398	XM_017427909

Contd...

Contd...

Genus	Species	Class	Order	Family	AQP	GB/U AA acc.	GB/U Nu acc.
Gymnodraco	*acuticeps**	Actinopterygii	Perciformes	Bathydraconidae	3b	XP_034096186	XM_034240295
Cyclopterus	*lumpus**	Actinopterygii	Perciformes	Cyclopteridae	3b	XP_034398004	XM_034542113
Notothenia	*coriiceps**	Actinopterygii	Perciformes	Nototheniidae	3b	XP_010776959	XM_010778657
Perca	*flavescens**	Actinopterygii	Perciformes	Percidae	3b	XP_028433841	XM_028578040
Sparus	*aurata*	Actinopterygii	Spariformes	Sparidae	3b	T1T009	KC788197
Takifugu	*rubripes**	Actinopterygii	Tetraodontiformes	Tetraodontidae	3b	XP_003975282	XM_003975233
Anarrhichthys	*ocellatus**	Actinopterygii		Anarhichadidae	3b	XP_031705548	XM_031849688
Larimichthys	*crocea**	Actinopterygii		Sciaenidae	3b	XP_010728020	XM_010729718
Cebidichthys	*violaceus**	Actinopterygii		Stichaeidae	3b		NJBE01000444
Erpetoichthys	*calabaricus**	Actinopterygii	Polypteriformes	Polypteridae	4	XP_028659469	XM_028803636
Polypterus	*senegalus**	Actinopterygii	Polypteriformes	Polypteridae	4	XP_039610614	XM_039754680
Acipenser	*ruthenus**	Actinopterygii	Acipenseriformes	Acipenseridae	4	XP_033857011	XM_034001119
Lepisosteus	*oculatus**	Actinopterygii	Lepisosteiformes	Lepisosteidae	4	XP_006634102	XM_006634039
Anguilla	*anguilla**	Actinopterygii	Anguiliformes	Anguilidae	4a	XP_035284243	XM_035428352
Scleropages	*formosus**	Actinopterygii	Osteoglossiformes	Osteoglossidae	4a	XP_018613023	XM_018757507
Astyanax	*mexicanus**	Actinopterygii	Characiformes	Characidae	4a	XP_022533869	XM_022678148
Ictalurus	*punctatus**	Actinopterygii	SIluriformes	Ictaluridae	4a	XP_017326842	XM_017471353
Pangasianodon	*hypophthalmus**	Actinopterygii	Siluriformes	Pangasiidae	4a	XP_026769902	XM_026914101
Salmo	*salar*	Actinopterygii	Salmoniformes	Salmonidae	4a	XP_013995700	XM_014140225
Gadus	*morhua**	Actinopterygii	Gadiformes	Gadidae	4a	XP_030220294	XM_030364434

Contd...

Contd...

Myripristis	*murdjan**	Actinopterygii	Holocentriformes	Holocentridae	4a	XP_029930393	XM_030074533
Lesueurigobius	*sanzi**	Actinopterygii	Gobiiformes	Gobiidae	4a		OMNZ01061975
Neogobius	*melanostomus**	Actinopterygii	Gobiiformes	Gobiidae	4a		VHKM01000009
Boleophthalmus	*pectinirostris*	Actinopterygii	Gobiiformes	Oxudercidae	4a	XP_020777769	XM_020922110
Periophthalmodon	*schlosseri*	Actinopterygii	Gobiiformes	Oxudercidae	4a		JACM01058456
Periophthalmus	*magnuspinnatus*	Actinopterygii	Gobiiformes	Oxudercidae	4a	XP_033838967	XM_033983076
Scartelaos	*histophorus*	Actinopterygii	Gobiiformes	Oxudercidae	4a		JACN01023074
Sphaeramia	*orbicularis**	Actinopterygii	Kurtiformes	Apogonidae	4a	XP_030017218	XM_030161358
Anabas	*testudineus**	Actinopterygii	Anabantiformes	Anabantidae	4a	XP_0262304541	XM_026374669
Channa	*argus**	Actinopterygii	Anabantiformes	Channidae	4a		CM015725
Helostoma	*temminki**	Actinopterygii	Anabantiformes	Helostomatidae	4a		OMLM01000915
Cynoglossus	*semilaevis**	Actinopterygii	Pleuronectiformes	Cynoglossidae	4a	XP_008332600	XM_008334378
Scophthalmus	*maximus**	Actinopterygii	Pleuronectiformes	Scophthalmidae	4a	AWP04023	
Mastacembelus	*armatus**	Actinopterygii	Synbranchiformes	Mastacembelidae	4a	XP_026169295	XM_026313510
Monopterus	*albus**	Actinopterygii	Synbranchiformes	Synbranchidae	4a	XP_020451481	XM_020595825
Parablennius	*parvicornis**	Actinopterygii	Blenniiformes	Blenniidae	4a		OMNM01007192
Salarias	*fasciatus**	Actinopterygii	Blenniiformes	Blenniidae	4a	XP_029970302	XM_030114442
Gouania	*willdenowi**	Actinopterygii	Gobiesociformes	Gobiesocidae	4a	XP_028329320	XM_028473519
Oryzias	*latipes*	Actinopterygii	Beloniformes	Adrianichthyidae	4a	XP_004079411	XM_004079363
Maylandia	*zebra**	Actinopterygii	Cichliformes	Cichlidae	4a	XP_004542791	XM_004542734
Oreochromis	*niloticus**	Actinopterygii	Cichliformes	Cichlidae	4a	XP_005476412	XM_005476355

Contd...

Contd...

Genus	Species	Class	Order	Family	AQP	GB/U AA acc.	GB/U Nu acc.
Anableps	*anableps**	Actinopterygii	Cyprinodontiformes	Anablepidae	4a		CM026077
Cyprinodon	*variegatus**	Actinopterygii	Cyprinodontiformes	Cyprinodontidae	4a	XP_015227982	XM_015372496
Fundulus	*heteroclitus**	Actinopterygii	Cyprinodontiformes	Fundulidae	4a	XP_012730861	XM_012875407
Aphyosemion	*australe**	Actinopterygii	Cyprinodontiformes	Nothobranchiidae	4a		SSNS01000681
Nothobranchius	*furzeri**	Actinopterygii	Cyprinodontiformes	Nothobranchiidae	4a	XP_015812316	XM_015956830
Gambusia	*affinis**	Actinopterygii	Cyprinodontiformes	Poeciliidae	4a		NHOQ01001318
Poecilia	*reticulata**	Actinopterygii	Cyprinodontiformes	Poeciliidae	4a	XP_008432290	XM_008434068
Austrofundulus	*limnaeus**	Actinopterygii	Cyprinodontiformes	Rivulidae	4a	XP_013886050	XM_014030596
Kryptolebias	*marmoratus**	Actinopterygii	Cyprinodontiformes	Rivulidae	4a	XP_0172778891	XM_017422400
Gymnodraco	*acuticeps**	Actinopterygii	Perciformes	Bathydraconidae	4a	XP_034090488	XM_034234597
Cyclopterus	*lumpus**	Actinopterygii	Perciformes	Cyclopteridae	4a	XP_034411603	XM_034555712
Notothenia	*coriiceps**	Actinopterygii	Perciformes	Nototheniidae	4a	XP_010783623	XM_010785321
Perca	*flavescens**	Actinopterygii	Perciformes	Percidae	4a	XP_028448786	XM_028592985
Takifugu	*rubripes**	Actinopterygii	Tetraodontiformes	Tetraodontidae	4a	XP_011617137	XM_011618835
Anarrhichthys	*ocellatus**	Actinopterygii		Anarhichadidae	4a	XP_031711173	XM_031855313
Larimichthys	*crocea**	Actinopterygii		Sciaenidae	4a	XP_010728313	XM_010730011
Cebidichthys	*violaceus**	Actinopterygii		Stichaeidae	4a		NJBE01000440
Latimeria	*chalumnae**	Sarcopterygii			4	XP_006005620	XM_006005558
Xenopus	*laevis*	Amphibia			4	NP_001304790	NM_001130949
Alligator	*mississippiensis*	Reptilia			4	XP_006264124	XM_019487427

Contd...

Contd...

Chelonia	*mydas*	Reptilia			4	XP_007069741	XM_007069679
Gallus	*gallus*	Aves			4	NP_001304756	NM_001004765
Homo	*sapiens*	Mammalia			4	NP_001641	NM_001317384
Mus	*musculus*	Mammalia			4	NP_001295573	NM_001308641
Anguilla	*anguilla**	Actinopterygii	Anguiliformes	Anguilidae	4b	XP_035270217	XM_035414326
Scleropages	*formosus**	Actinopterygii	Osteoglossiformes	Osteoglossidae	4b	XP_018602986	XM_018747470
Astyanax	*mexicanus**	Actinopterygii	Characiformes	Characidae	4b	XP_007233079	XM_007233017
Cyprinus	*carpio*	Actinopterygii	Cypriniformes	Cyprinidae	4b	A0A173N0R4	LC149722
Danio	*rerio*	Actinopterygii	Cypriniformes	Danionidae	4b	NP_001003749	NM_001003749
Clarias	*batrachus**	Actinopterygii	Siluriformes	Clariidae	4b		QMIH01000781
Ictalurus	*punctatus**	Actinopterygii	SIluriformes	Ictaluridae	4b	XP_017310898	XM_017455409
Pangasianodon	*hypophthalmus**	Actinopterygii	Siluriformes	Pangasiidae	4b	XP_026800384	XM_026944583
Esox	*lucius**	Actinopterygii	Esociformes	Esocidae	4b	XP_010870619	XM_010872317
Salmo	*salar*	Actinopterygii	Salmoniformes	Salmonidae	4b	XP_014024653	XM_014169178
Gadus	*morhua**	Actinopterygii	Gadiformes	Gadidae	4b	XP_030227976	XM_030372116
Myripristis	*murdjan**	Actinopterygii	Holocentriformes	Holocentridae	4b	XP_029905997	XM_030050137
Sphaeramia	*orbicularis**	Actinopterygii	Kurtiformes	Apogonidae	4b	XP_029989069	XM_030133209
Oreochromis	*niloticus**	Actinopterygii	Cichliformes	Cichlidae	4b	XP_003452583	XM_003452535
Gymnodraco	*acuticeps**	Actinopterygii	Perciformes	Bathydraconidae	4b	XP_034093351	XM_034237460
Cyclopterus	*lumpus**	Actinopterygii	Perciformes	Cyclopteridae	4b	XP_034386201	XM_034530310
Notothenia	*coriiceps**	Actinopterygii	Perciformes	Nototheniidae	4b	XP_010777188	XM_010778886

Contd...

Contd...

Genus	Species	Class	Order	Family	AQP	GB/U AA acc.	GB/U Nu acc.
Anarrhichthys	*ocellatus**	Actinopterygii		Anarhichadidae	4b	XP_031734673	XM_031878813
Larimichthys	*crocea**	Actinopterygii		Sciaenidae	4b	XP_010733986	XM_010735684
Cebidichthys	*violaceus**	Actinopterygii		Stichaeidae	4b		NJBE01000466
Xenopus	*laevis*	Amphibia			5	XP_018104235	XM_018248746
Alligator	*mississippiensis*	Reptilia			5	XP_006275752	XM_006275690
Chelonia	*mydas*	Reptilia			5	M7ALY6	KB582107
Gallus	*gallus*	Aves			5	XP_001231781	XM_001231780
Homo	*sapiens*	Mammalia			5	NP_001642	NM_001651
Mus	*musculus*	Mammalia			5	NP_033831	NM_009701
Xenopus	*laevis*	Amphibia			6	NP_001163923	NM_001170452
Chelonia	*mydas*	Reptilia			6	XP_007064776	XM_007064714
Homo	*sapiens*	Mammalia			6	NP_001643	NM_001652
Mus	*musculus*	Mammalia			6	NP_780296	NM_175087
Polypterus	*senegalus**	Actinopterygii	Polypteriformes	Polypteridae	7	XP_039613889	XM_039757955
Acipenser	*ruthenus**	Actinopterygii	Acipenseriformes	Acipenseridae	7	XP_033862155	XM_034006264
Anguilla	*anguilla**	Actinopterygii	Anguiliformes	Anguilidae	7	XP_035248109	XM_035392218
Astyanax	*mexicanus**	Actinopterygii	Characiformes	Characidae	7	XP_007246076	XM_007246014
Cyprinus	*carpio*	Actinopterygii	Cypriniformes	Cyprinidae	7	A0A0K2RVW8	LC069015
Danio	*rerio*	Actinopterygii	Cypriniformes	Danionidae	7	NP_956204	NM_199910
Clarias	*batrachus**	Actinopterygii	Siluriformes	Clariidae	7		QMIH01000185

Contd...

Contd...

Ictalurus	*punctatus**	Actinopterygii	SIluriformes	Ictaluridae	7	NP_001188011	NM_001201082
Pangasianodon	*hypophthalmus**	Actinopterygii	Siluriformes	Pangasiidae	7	XP_026784646	XM_026928845
Salmo	*salar*	Actinopterygii	Salmoniformes	Salmonidae	7	XP_013987498	XM_014132023
Gadus	*morhua**	Actinopterygii	Gadiformes	Gadidae	7	XP_030229722	XM_030373862
Myripristis	*murdjan**	Actinopterygii	Holocentriformes	Holocentridae	7	XP_029907011	XM_030051151
Neogobius	*melanostomus**	Actinopterygii	Gobiiformes	Gobiidae	7		VHKM01000729
Boleophthalmus	*pectinirostris*	Actinopterygii	Gobiiformes	Oxudercidae	7		KN525596
Periophthalmodon	*schlosseri*	Actinopterygii	Gobiiformes	Oxudercidae	7		KN471526
Periophthalmus	*magnuspinnatus*	Actinopterygii	Gobiiformes	Oxudercidae	7	XP_033822890	XM_033966999
Sphaeramia	*orbicularis**	Actinopterygii	Kurtiformes	Apogonidae	7	XP_029990298	XM_030134438
Anabas	*testudineus**	Actinopterygii	Anabantiformes	Anabantidae	7	XP_0262050591	XM_026349274
Helostoma	*temminki**	Actinopterygii	Anabantiformes	Helostomatidae	7		OMLM01000905
Betta	*splendens**	Actinopterygii	Anabantiformes	Osphronemidae	7	XP_029006757	XM_029150924
Cynoglossus	*semilaevis**	Actinopterygii	Pleuronectiformes	Cynoglossidae	7	XP_008318637	XM_008320415
Scophthalmus	*maximus**	Actinopterygii	Pleuronectiformes	Scophthalmidae	7	AWP03195	
Mastacembelus	*armatus**	Actinopterygii	Synbranchiformes	Mastacembelidae	7	XP_026161742	XM_026305957
Monopterus	*albus**	Actinopterygii	Synbranchiformes	Synbranchidae	7	XP_020448275	XM_020592619
Parablennius	*parvicornis**	Actinopterygii	Blenniiformes	Blenniidae	7		OMNM01020894
Salarias	*fasciatus**	Actinopterygii	Blenniiformes	Blenniidae	7	XP_029947169	XM_030091309
Gouania	*willdenowi**	Actinopterygii	Gobiesociformes	Gobiesocidae	7	XP_028301671	XM_028445870
Oryzias	*latipes*	Actinopterygii	Beloniformes	Adrianichthyidae	7	XP_020558956	XM_020703297

Contd...

Contd...

Genus	Species	Class	Order	Family	AQP	GB/U AA acc.	GB/U Nu acc.
Maylandia	*zebra**	Actinopterygii	Cichliformes	Cichlidae	7	XP_004545281	XM_004545224
Oreochromis	*niloticus**	Actinopterygii	Cichliformes	Cichlidae	7	XP_003441614	XM_003441566
Anableps	*anableps**	Actinopterygii	Cyprinodontiformes	Anablepidae	7		CM026069
Cyprinodon	*variegatus**	Actinopterygii	Cyprinodontiformes	Cyprinodontidae	7	XP_015235357	XM_015379871
Fundulus	*heteroclitus**	Actinopterygii	Cyprinodontiformes	Fundulidae	7	XP_012726204	XM_012870750
Aphyosemion	*australe**	Actinopterygii	Cyprinodontiformes	Nothobranchiidae	7		SSNS01000035
Nothobranchius	*furzeri**	Actinopterygii	Cyprinodontiformes	Nothobranchiidae	7	XP_015814885	XM_015959399
Poecilia	*reticulata**	Actinopterygii	Cyprinodontiformes	Poeciliidae	7	XP_008407704	XM_008409482
Austrofundulus	*limnaeus**	Actinopterygii	Cyprinodontiformes	Rivulidae	7	XP_013872483	XM_014017029
Kryptolebias	*marmoratus**	Actinopterygii	Cyprinodontiformes	Rivulidae	7	XP_0172784241	XM_017422935
Gymnodraco	*acuticeps**	Actinopterygii	Perciformes	Bathydraconidae	7	XP_034095378	XM_034239487
Cyclopterus	*lumpus**	Actinopterygii	Perciformes	Cyclopteridae	7	XP_034389320	XM_034533429
Notothenia	*coriiceps**	Actinopterygii	Perciformes	Nototheniidae	7	XP_010764522	XM_010766220
Perca	*flavescens**	Actinopterygii	Perciformes	Percidae	7	XP_028432788	XM_028576987
Sparus	*aurata*	Actinopterygii	Spariformes	Sparidae	7	T1T284	KC589386
Takifugu	*rubripes**	Actinopterygii	Tetraodontiformes	Tetraodontidae	7	XP_003973612	XM_003973563
Anarrhichthys	*ocellatus**	Actinopterygii		Anarhichadidae	7	XP_031701732	XM_031845872
Larimichthys	*crocea**	Actinopterygii		Sciaenidae	7	XP_010731808	XM_010733506
Cebidichthys	*violaceus**	Actinopterygii		Stichaeidae	7		NJBE01000433
Latimeria	*chalumnae**	Sarcopterygii			7	XP_006004792	

Contd...

Contd...

Xenopus	*laevis*	Amphibia			7	Q5FW23	BC089658
Alligator	*mississippiensis*	Reptilia			7	XP_006268313	XM_006268251
Chelonia	*mydas*	Reptilia			7	XP_007063361	XM_007063299
Gallus	*gallus*	Aves			7	XP_015132918	XM_015277432
Homo	*sapiens*	Mammalia			7	NP_001161	NM_001170
Mus	*musculus*	Mammalia			7	NP_031499	XM_006537560
Polypterus	*senegalus**	Actinopterygii	Polypteriformes	Polypteridae	8a	XP_039630512	XM_039774578
Acipenser	*ruthenus**	Actinopterygii	Acipenseriformes	Acipenseridae	8a	XP_033885726	XM_034029835
Lepisosteus	*oculatus**	Actinopterygii	Lepisosteiformes	Lepisosteidae	8a	XP_006637203	XM_006637140
Anguilla	*anguilla**	Actinopterygii	Anguiliformes	Anguilidae	8a1	XP_035258615	XM_035402724
Scleropages	*formosus**	Actinopterygii	Osteoglossiformes	Osteoglossidae	8a1	XP_018610193	XM_018754677
Astyanax	*mexicanus**	Actinopterygii	Characiformes	Characidae	8a1	XP_007240584	XM_007240522
Cyprinus	*carpio*	Actinopterygii	Cypriniformes	Cyprinidae	8a1	XP_018968618	XM_019113073
Danio	*rerio*	Actinopterygii	Cypriniformes	Danionidae	8a1	NP_001004661	NM_001004661
Clarias	*batrachus**	Actinopterygii	Siluriformes	Clariidae	8a1		QMIH01000112
Ictalurus	*punctatus**	Actinopterygii	SIluriformes	Ictaluridae	8a1	XP_017339872	XM_017484383
Pangasianodon	*hypophthalmus**	Actinopterygii	Siluriformes	Pangasiidae	8a1	XP_026770446	XM_026914645
Esox	*lucius**	Actinopterygii	Esociformes	Esocidae	8a1	XP_010903710	XM_010905408
Salmo	*salar*	Actinopterygii	Salmoniformes	Salmonidae	8a1	XP_014035158	XM_014179683
Gadus	*morhua**	Actinopterygii	Gadiformes	Gadidae	8a1	XP_030195136	XM_030339276
Myripristis	*murdjan**	Actinopterygii	Holocentriformes	Holocentridae	8a1	XP_029933908	XM_030078048

Contd...

Contd...

Genus	Species	Class	Order	Family	AQP	GB/U AA acc.	GB/U Nu acc.
Lesueurigobius	*sanzi**	Actinopterygii	Gobiiformes	Gobiidae	8a1		OMNZ01071362
Boleophthalmus	*pectinirostris*	Actinopterygii	Gobiiformes	Oxudercidae	8a1	XP_020780042	XM_020924383
Periophthalmodon	*schlosseri*	Actinopterygii	Gobiiformes	Oxudercidae	8a1		JACM01008777
Periophthalmus	*magnuspinnatus*	Actinopterygii	Gobiiformes	Oxudercidae	8a1	XP_033840801	XM_033984910
Scartelaos	*histophorus*	Actinopterygii	Gobiiformes	Oxudercidae	8a1		JACN01005585
Sphaeramia	*orbicularis**	Actinopterygii	Kurtiformes	Apogonidae	8a1	XP_029978537	XM_030122677
Anabas	*testudineus**	Actinopterygii	Anabantiformes	Anabantidae	8a1	XP_0262075341	XM_026351749
Helostoma	*temminki**	Actinopterygii	Anabantiformes	Helostomatidae	8a1		OMLM01011930
Betta	*splendens**	Actinopterygii	Anabantiformes	Osphronemidae	8a1	XP_028991073	XM_029135240
Mastacembelus	*armatus**	Actinopterygii	Synbranchiformes	Mastacembelidae	8a1	XP_026172224	XM_026316439
Monopterus	*albus**	Actinopterygii	Synbranchiformes	Synbranchidae	8a1	XP_020471596	XM_020615940
Parablennius	*parvicornis**	Actinopterygii	Blenniiformes	Blenniidae	8a1		OMNM01003462
Salarias	*fasciatus**	Actinopterygii	Blenniiformes	Blenniidae	8a1	XP_029954519	XM_030098659
Maylandia	*zebra**	Actinopterygii	Cichliformes	Cichlidae	8a1	XP_004562341	XM_004562284
Oreochromis	*niloticus**	Actinopterygii	Cichliformes	Cichlidae	8a1	XP_003438763	XM_003438715
Cyprinodon	*variegatus**	Actinopterygii	Cyprinodontiformes	Cyprinodontidae	8a1	XP_015232546	XM_015377060
Gymnodraco	*acuticeps**	Actinopterygii	Perciformes	Bathydraconidae	8a1	XP_034056689	XM_034200798
Notothenia	*coriiceps**	Actinopterygii	Perciformes	Nototheniidae	8a1	XP_010791351	XM_010793049
Perca	*flavescens**	Actinopterygii	Perciformes	Percidae	8a1	XP_028424040	XM_028568239
Anarrhichthys	*ocellatus**	Actinopterygii		Anarhichadidae	8a1	XP_031699747	XM_031843887

Contd...

Contd...

Larimichthys	*crocea**	Actinopterygii		Sciaenidae	8a1	XP_010729101	XM_010730799
Cebidichthys	*violaceus**	Actinopterygii		Stichaeidae	8a1		NJBE01000436
Latimeria	*chalumnae**	Sarcopterygii			8	XP_005994818	XM_005994756
Xenopus	*laevis*	Amphibia			8	XP_018094770	XM_018239281
Alligator	*mississippiensis*	Reptilia			8	XP_006268506	XM_006268444
Chelonia	*mydas*	Reptilia			8	XP_007058667	XM_007058605
Gallus	*gallus*	Aves			8	XP_414866	XM_414866
Homo	*sapiens*	Mammalia			8	NP_001160	NM_001169
Mus	*musculus*	Mammalia			8	NP_031500	NM_007474
Clarias	*batrachus**	Actinopterygii	Siluriformes	Clariidae	8a2		QMIH01000583
Ictalurus	*punctatus**	Actinopterygii	SIluriformes	Ictaluridae	8a2	XP_017341894	XM_017486405
Pangasianodon	*hypophthalmus**	Actinopterygii	Siluriformes	Pangasiidae	8a2	XP_026772388	XM_026916587
Gadus	*morhua**	Actinopterygii	Gadiformes	Gadidae	8a2	XP_030230176	XM_030374316
Myripristis	*murdjan**	Actinopterygii	Holocentriformes	Holocentridae	8a2	XP_029914309	XM_030058449
Polypterus	*senegalus**	*Actinopterygii*	*Polypteriformes*	*Polypteridae*	8b	XP_039632310	XM_039776376
Acipenser	*ruthenus**	Actinopterygii	Acipenseriformes	Acipenseridae	8b	XP_033909017	XM_034926537
Lepisosteus	*oculatus**	Actinopterygii	Lepisosteiformes	Lepisosteidae	8b	XP_006637204	XM_015360095
Anguilla	*anguilla**	Actinopterygii	Anguiliformes	Anguilidae	8b1	XP_035261913	XM_035406022
Cyprinus	*carpio*	Actinopterygii	Cypriniformes	Cyprinidae	8b1	A0A0K2RVU3	LC069019
Danio	*rerio*	Actinopterygii	Cypriniformes	Danionidae	8b1	NP_001073651	NM_001080182
Clarias	*batrachus**	Actinopterygii	Siluriformes	Clariidae	8b1		QMIH01000112

Contd...

Contd...

Genus	Species	Class	Order	Family	AQP	GB/U AA acc.	GB/U Nu acc.
Ictalurus	*punctatus**	Actinopterygii	SIluriformes	Ictaluridae	8b1	XP_017339873	XM_017484384
Pangasianodon	*hypophthalmus**	Actinopterygii	Siluriformes	Pangasiidae	8b1	XP_026770553	XM_026914752
Esox	*lucius**	Actinopterygii	Esociformes	Esocidae	8b1	XP_010872559	XM_010874257
Salmo	*salar*	Actinopterygii	Salmoniformes	Salmonidae	8b1	XP_014035160	XM_014179685
Gadus	*morhua**	Actinopterygii	Gadiformes	Gadidae	8b1	XP_030195134	XM_030339274
Myripristis	*murdjan**	Actinopterygii	Holocentriformes	Holocentridae	8b1	XP_029933943	XM_030078083
Lesueurigobius	*sanzi**	Actinopterygii	Gobiiformes	Gobiidae	8b1		OMNZ01044540
Neogobius	*melanostomus**	Actinopterygii	Gobiiformes	Gobiidae	8b1		VHKM01000156
Boleophthalmus	*pectinirostris*	Actinopterygii	Gobiiformes	Oxudercidae	8b1	XP_020780043	XM_020924384
Periophthalmodon	*schlosseri*	Actinopterygii	Gobiiformes	Oxudercidae	8b1		JACM0100877/01035013
Periophthalmus	*magnuspinnatus*	Actinopterygii	Gobiiformes	Oxudercidae	8b1	XP_033840925	XM_033985034
Scartelaos	*histophorus*	Actinopterygii	Gobiiformes	Oxudercidae	8b1		KN499303
Sphaeramia	*orbicularis**	Actinopterygii	Kurtiformes	Apogonidae	8b1	XP_029978149	XM_030122289
Anabas	*testudineus**	Actinopterygii	Anabantiformes	Anabantidae	8b1	XP_0262074741	XM_026351689
Channa	*argus**	Actinopterygii	Anabantiformes	Channidae	8b1		CM015731
Helostoma	*temminki**	Actinopterygii	Anabantiformes	Helostomatidae	8b1		OMLM01011930
Betta	*splendens**	Actinopterygii	Anabantiformes	Osphronemidae	8b1	XP_028990930	XM_029135097
Scophthalmus	*maximus**	Actinopterygii	Pleuronectiformes	Scophthalmidae	8b1	AWP17402	
Mastacembelus	*armatus**	Actinopterygii	Synbranchiformes	Mastacembelidae	8b1	XP_026171359	XM_026315574
Monopterus	*albus**	Actinopterygii	Synbranchiformes	Synbranchidae	8b1	XP_020471597	XM_020615941

Contd...

Contd...

Salarias	*fasciatus**	Actinopterygii	Blenniiformes	Blenniidae	8b1	XP_029954721	XM_030098861
Gouania	*willdenowi**	Actinopterygii	Gobiesociformes	Gobiesocidae	8b1	XP_028332066	XM_028476265
Oryzias	*latipes*	Actinopterygii	Beloniformes	Adrianichthyidae	8b1	XP_023805199	XM_023949431
Maylandia	*zebra**	Actinopterygii	Cichliformes	Cichlidae	8b1	XP_004562340	XM_004562283
Oreochromis	*niloticus**	Actinopterygii	Cichliformes	Cichlidae	8b1	XP_003438762	XM_003438714
Anableps	*anableps**	Actinopterygii	Cyprinodontiformes	Anablepidae	8b1		CM026077
Cyprinodon	*variegatus**	Actinopterygii	Cyprinodontiformes	Cyprinodontidae	8b1	XP_015232536	XM_015377050
Fundulus	*heteroclitus**	Actinopterygii	Cyprinodontiformes	Fundulidae	8b1	XP_012721780	XM_012866326
Aphyosemion	*australe**	Actinopterygii	Cyprinodontiformes	Nothobranchiidae	8b1		SSNS01000586
Nothobranchius	*furzeri**	Actinopterygii	Cyprinodontiformes	Nothobranchiidae	8b1	XP_015819003	XM_015963517
Gambusia	*affinis**	Actinopterygii	Cyprinodontiformes	Poeciliidae	8b1		NHOQ01000968
Poecilia	*reticulata**	Actinopterygii	Cyprinodontiformes	Poeciliidae	8b1	XP_008434661	XM_008436439
Austrofundulus	*limnaeus**	Actinopterygii	Cyprinodontiformes	Rivulidae	8b1	XP_013874723	XM_014019269
Kryptolebias	*marmoratus**	Actinopterygii	Cyprinodontiformes	Rivulidae	8b1	XP_017274015	XM_017418526
Gymnodraco	*acuticeps**	Actinopterygii	Perciformes	Bathydraconidae	8b1	XP_034056695	XM_034200804
Cyclopterus	*lumpus**	Actinopterygii	Perciformes	Cyclopteridae	8b1	XP_034414128	XM_034558237
Notothenia	*coriiceps**	Actinopterygii	Perciformes	Nototheniidae	8b1	XP_010791353	XM_010793051
Perca	*flavescens**	Actinopterygii	Perciformes	Percidae	8b1	XP_028423592	XM_028567791
Larimichthys	*crocea**	Actinopterygii		Sciaenidae	8b1	XP_010729100	XM_010730798
Cebidichthys	*violaceus**	Actinopterygii		Stichaeidae	8b1		NJBE01000436
Anguilla	*anguilla**	Actinopterygii	Anguiliformes	Anguilidae	8b2	XP_035254918	XM_035399027

Contd...

Contd...

Genus	Species	Class	Order	Family	AQP	GB/U AA acc.	GB/U Nu acc.
Scleropages	*formosus**	Actinopterygii	Osteoglossiformes	Osteoglossidae	8b2	XP_018610213	XM_018754697
Astyanax	*mexicanus**	Actinopterygii	Characiformes	Characidae	8b2	XP_022533364	XM_022677643
Cyprinus	*carpio*	Actinopterygii	Cypriniformes	Cyprinidae	8b2	A0A0K2RVX2	LC069020
Danio	*rerio*	Actinopterygii	Cypriniformes	Danionidae	8b2	NP_001108382	NM_001114910
Salmo	*salar*	Actinopterygii	Salmoniformes	Salmonidae	8b2	NP_001167386	NM_001173915
Myripristis	*murdjan**	Actinopterygii	Holocentriformes	Holocentridae	8b2	XP_029914305	XM_030058445
Sphaeramia	*orbicularis**	Actinopterygii	Kurtiformes	Apogonidae	8b2	XP_029997540	XM_030141680
Cynoglossus	*semilaevis**	Actinopterygii	Pleuronectiformes	Cynoglossidae	8b2	XP_016896014	XM_017040525
Scophthalmus	*maximus**	Actinopterygii	Pleuronectiformes	Scophthalmidae	8b2	AWP18872	
Gymnodraco	*acuticeps**	Actinopterygii	Perciformes	Bathydraconidae	8b2	XP_034076043	XM_034220152
Cyclopterus	*lumpus**	Actinopterygii	Perciformes	Cyclopteridae	8b2	XP_034395904	XM_034540013
Notothenia	*coriiceps**	Actinopterygii	Perciformes	Nototheniidae	8b2	XP_010791963	XM_010793661
Sparus	*aurata*	Actinopterygii	Spariformes	Sparidae	8b2	C7S301	DQ889225
Takifugu	*rubripes**	Actinopterygii	Tetraodontiformes	Tetraodontidae	8b2	XP_003964594	XM_003964545
Anarrhichthys	*ocellatus**	Actinopterygii		Anarhichadidae	8b2	XP_031728107	XM_031872247
Larimichthys	*crocea**	Actinopterygii		Sciaenidae	8b2	XP_010728980	XM_010730678
Cebidichthys	*violaceus**	Actinopterygii		Stichaeidae	8b2		NJBE01000011
Erpetoichthys	*calabaricus**	Actinopterygii	Polypteriformes	Polypteridae	9	XP_028679713	XM_028823880
Polypterus	*senegalus**	Actinopterygii	Polypteriformes	Polypteridae	9	XP_039626162	XM_039770228
Acipenser	*ruthenus**	Actinopterygii	Acipenseriformes	Acipenseridae	9	XP_034763413	XM_034048312

Contd...

Contd...

Lepisosteus	*oculatus**	Actinopterygii	Lepisosteiformes	Lepisosteidae	9	XP_006628786	XM_006628723
Anguilla	*anguilla**	Actinopterygii	Anguiliformes	Anguilidae	9a	XP_035252445	XM_035396554
Scleropages	*formosus**	Actinopterygii	Osteoglossiformes	Osteoglossidae	9a	XP_018595090	XM_018739574
Astyanax	*mexicanus**	Actinopterygii	Characiformes	Characidae	9a	XP_007244411	XM_007244349
Cyprinus	*carpio*	Actinopterygii	Cypriniformes	Cyprinidae	9a	A0A173N0K5	LC149726
Danio	*rerio*	Actinopterygii	Cypriniformes	Danionidae	9a	NP_001028268	NM_001033096
Clarias	*batrachus**	Actinopterygii	Siluriformes	Clariidae	9a		QMIH01000630
Ictalurus	*punctatus**	Actinopterygii	SIluriformes	Ictaluridae	9a	XP_017341657	XM_017486168
Pangasianodon	*hypophthalmus**	Actinopterygii	Siluriformes	Pangasiidae	9a	XP_034163046	XM_034307155
Esox	*lucius**	Actinopterygii	Esociformes	Esocidae	9a	XP_010881995	XM_010883693
Salmo	*salar*	Actinopterygii	Salmoniformes	Salmonidae	9a	XP_014025743	XM_014170268
Gadus	*morhua**	Actinopterygii	Gadiformes	Gadidae	9a	XP_030221325	XM_030365465
Myripristis	*murdjan**	Actinopterygii	Holocentriformes	Holocentridae	9a	XP_029910403	XM_030054543
Neogobius	*melanostomus**	Actinopterygii	Gobiiformes	Gobiidae	9a		VHKM01000258
Boleophthalmus	*pectinirostris*	Actinopterygii	Gobiiformes	Oxudercidae	9a	XP_020790536	XM_020934877
Periophthalmodon	*schlosseri*	Actinopterygii	Gobiiformes	Oxudercidae	9a	KN482823	
Periophthalmus	*magnuspinnatus*	Actinopterygii	Gobiiformes	Oxudercidae	9a	XP_033823925	XM_033968034
Sphaeramia	*orbicularis**	Actinopterygii	Kurtiformes	Apogonidae	9a	XP_029992516	XM_030136656
Anabas	*testudineus**	Actinopterygii	Anabantiformes	Anabantidae	9a	XP_026222073	XM_026366288
Channa	*argus**	Actinopterygii	Anabantiformes	Channidae	9a		CM015715
Helostoma	*temminki**	Actinopterygii	Anabantiformes	Helostomatidae	9a		OMLM01005269

Contd...

Contd...

Genus	Species	Class	Order	Family	AQP	GB/U AA acc.	GB/U Nu acc.
Betta	*splendens**	Actinopterygii	Anabantiformes	Osphronemidae	9a	XP_029010150	XM_029154317
Cynoglossus	*semilaevis**	Actinopterygii	Pleuronectiformes	Cynoglossidae	9a	XP_008310708	XM_008312486
Scophthalmus	*maximus**	Actinopterygii	Pleuronectiformes	Scophthalmidae	9a	AWP08290	
Mastacembelus	*armatus**	Actinopterygii	Synbranchiformes	Mastacembelidae	9a	XP_026167196	XM_026311411
Monopterus	*albus**	Actinopterygii	Synbranchiformes	Synbranchidae	9a	XP_020480340	XM_020624684
Parablennius	*parvicornis**	Actinopterygii	Blenniiformes	Blenniidae	9a		OMNM0100286/01052775
Salarias	*fasciatus**	Actinopterygii	Blenniiformes	Blenniidae	9a	XP_029951172	XM_030095312
Gouania	*willdenowi**	Actinopterygii	Gobiesociformes	Gobiesocidae	9a	XP_028304598	XM_028448797
Oryzias	*latipes*	Actinopterygii	Beloniformes	Adrianichthyidae	9a	XP_004069619	XM_004069571
Maylandia	*zebra**	Actinopterygii	Cichliformes	Cichlidae	9a	XP_004567715	XM_004567658
Oreochromis	*niloticus**	Actinopterygii	Cichliformes	Cichlidae	9a	XP_003443906	XM_003443858
Anableps	*anableps**	Actinopterygii	Cyprinodontiformes	Anablepidae	9a		CM026073
Cyprinodon	*variegatus**	Actinopterygii	Cyprinodontiformes	Cyprinodontidae	9a	XP_015260792	XM_015405306
Fundulus	*heteroclitus**	Actinopterygii	Cyprinodontiformes	Fundulidae	9a	XP_012735786	XM_012880332
Aphyosemion	*australe**	Actinopterygii	Cyprinodontiformes	Nothobranchiidae	9a		SSNS01000408
Nothobranchius	*furzeri**	Actinopterygii	Cyprinodontiformes	Nothobranchiidae	9a	XP_015819717	XM_015964231
Gambusia	*affinis**	Actinopterygii	Cyprinodontiformes	Poeciliidae	9a		NHOQ01000384
Poecilia	*reticulata**	Actinopterygii	Cyprinodontiformes	Poeciliidae	9a	XP_008410230	XM_008412008
Austrofundulus	*limnaeus**	Actinopterygii	Cyprinodontiformes	Rivulidae	9a	XP_013876960	XM_014021506
Kryptolebias	*marmoratus**	Actinopterygii	Cyprinodontiformes	Rivulidae	9a	XP_0172610821	XM_017405593

Contd...

Contd...

Gymnodraco	*acuticeps**	Actinopterygii	Perciformes	Bathydraconidae	9a	XP_034060795	XM_034204904
Cyclopterus	*lumpus**	Actinopterygii	Perciformes	Cyclopteridae	9a	XP_034390472	XM_034534581
Notothenia	*coriiceps**	Actinopterygii	Perciformes	Nototheniidae	9a	XP_010786336	XM_010788034
Perca	*flavescens**	Actinopterygii	Perciformes	Percidae	9a	XP_028441172	XM_028585371
Takifugu	*rubripes**	Actinopterygii	Tetraodontiformes	Tetraodontidae	9a	XP_003967758	XM_003967709
Anarrhichthys	*ocellatus**	Actinopterygii		Anarhichadidae	9a	XP_031721786	XM_031865926
Larimichthys	*crocea**	Actinopterygii		Sciaenidae	9a	A0A0F8CJ34	KQ041939
Cebidichthys	*violaceus**	Actinopterygii		Stichaeidae	9a		NJBE01000432
Latimeria	*chalumnae**	Sarcopterygii			9	XP_014345555	XM_005998732
Xenopus	*laevis*	Amphibia			9	XP_018110968	XM_018255479
Alligator	*mississippiensis*	Reptilia			9	XP_014461591	XM_014606105
Chelonia	*mydas*	Reptilia			9	XP_007056447	XM_007056385
Gallus	*gallus*	Aves			9	NP_001280167	NM_001293238
Homo	*sapiens*	Mammalia			9	NP_066190	NM_020980
Mus	*musculus*	Mammalia			9	NP_001258772	NM_001271843
Anguilla	*anguilla**	Actinopterygii	Anguiliformes	Anguilidae	9b	XP_035275681	XM_035419790
Astyanax	*mexicanus**	Actinopterygii	Characiformes	Characidae	9b	XP_007230524	XM_007230462
Cyprinus	*carpio*	Actinopterygii	Cypriniformes	Cyprinidae	9b	XP_018956851	XM_019101306
Danio	*rerio*	Actinopterygii	Cypriniformes	Danionidae	9b	NP_001171215	NM_001177744
Clarias	*batrachus**	Actinopterygii	Siluriformes	Clariidae	9b		QMIH01000350
Ictalurus	*punctatus**	Actinopterygii	SIluriformes	Ictaluridae	9b	XP_017314149	XM_017458660

Contd...

Contd...

Genus	Species	Class	Order	Family	AQP	GB/U AA acc.	GB/U Nu acc.
Pangasianodon	*hypophthalmus**	Actinopterygii	Siluriformes	Pangasiidae	9b	XP_034155531	XM_034299640
Esox	*lucius**	Actinopterygii	Esociformes	Esocidae	9b	XP_010878699	XM_010880397
Salmo	*salar*	Actinopterygii	Salmoniformes	Salmonidae	9b	XP_013982375	XM_014126900
Gadus	*morhua**	Actinopterygii	Gadiformes	Gadidae	9b	XP_030233880	XM_030378020
Myripristis	*murdjan**	Actinopterygii	Holocentriformes	Holocentridae	9b	XP_029900464	XM_030044604
Lesueurigobius	*sanzi**	Actinopterygii	Gobiiformes	Gobiidae	9b		OMNZ01051098
Neogobius	*melanostomus**	Actinopterygii	Gobiiformes	Gobiidae	9b		VHKM01000131
Boleophthalmus	*pectinirostris*	Actinopterygii	Gobiiformes	Oxudercidae	9b	XP_020796975	XM_020941316
Periophthalmodon	*schlosseri*	Actinopterygii	Gobiiformes	Oxudercidae	9b		KN469413
Periophthalmus	*magnuspinnatus*	Actinopterygii	Gobiiformes	Oxudercidae	9b	XP_033843944	XM_033988053
Scartelaos	*histophorus*	Actinopterygii	Gobiiformes	Oxudercidae	9b		JACN01079684
Sphaeramia	*orbicularis**	Actinopterygii	Kurtiformes	Apogonidae	9b	XP_029979212	XM_030123352
Anabas	*testudineus**	Actinopterygii	Anabantiformes	Anabantidae	9b	XP_026234701	XM_026378916
Channa	*argus**	Actinopterygii	Anabantiformes	Channidae	9b		CM015713
Helostoma	*temminki**	Actinopterygii	Anabantiformes	Helostomatidae	9b		OMLM01020921
Betta	*splendens**	Actinopterygii	Anabantiformes	Osphronemidae	9b	XP_029000352	XM_029144519
Cynoglossus	*semilaevis**	Actinopterygii	Pleuronectiformes	Cynoglossidae	9b	XP_008335183	XM_008336961
Mastacembelus	*armatus**	Actinopterygii	Synbranchiformes	Mastacembelidae	9b	XP_026175791	XM_026320006
Monopterus	*albus**	Actinopterygii	Synbranchiformes	Synbranchidae	9b	XP_020465034	XM_020609378
Parablennius	*parvicornis**	Actinopterygii	Blenniiformes	Blenniidae	9b		OMNM01016882/01054549

Contd...

Contd...

Salarias	*fasciatus**	Actinopterygii	Blenniiformes	Blenniidae	9b	XP_029953247	XM_030097387
Gouania	*willdenowi**	Actinopterygii	Gobiesociformes	Gobiesocidae	9b	XP_028292217	XM_028436416
Oryzias	*latipes*	Actinopterygii	Beloniformes	Adrianichthyidae	9b	XP_023809256	XM_023953488
Maylandia	*zebra**	Actinopterygii	Cichliformes	Cichlidae	9b	XP_004543135	XM_004543078
Oreochromis	*niloticus**	Actinopterygii	Cichliformes	Cichlidae	9b	XP_003442315	XM_003442267
Anableps	*anableps**	Actinopterygii	Cyprinodontiformes	Anablepidae	9b		CM026067
Cyprinodon	*variegatus**	Actinopterygii	Cyprinodontiformes	Cyprinodontidae	9b	XP_015247291	XM_015391805
Fundulus	*heteroclitus**	Actinopterygii	Cyprinodontiformes	Fundulidae	9b	XP_012710637	XM_012855183
Aphyosemion	*australe**	Actinopterygii	Cyprinodontiformes	Nothobranchiidae	9b		SSNS01000024
Nothobranchius	*furzeri**	Actinopterygii	Cyprinodontiformes	Nothobranchiidae	9b	XP_015808486	XM_015953000
Gambusia	*affinis**	Actinopterygii	Cyprinodontiformes	Poeciliidae	9b		NHOQ01002371
Poecilia	*reticulata**	Actinopterygii	Cyprinodontiformes	Poeciliidae	9b	XP_008403593	XM_008405371
Austrofundulus	*limnaeus**	Actinopterygii	Cyprinodontiformes	Rivulidae	9b	XP_013883066	XM_014027612
Kryptolebias	*marmoratus**	Actinopterygii	Cyprinodontiformes	Rivulidae	9b	XP_0172959901	XM_017440501
Gymnodraco	*acuticeps**	Actinopterygii	Perciformes	Bathydraconidae	9b	XP_034052741	XM_034196850
Cyclopterus	*lumpus**	Actinopterygii	Perciformes	Cyclopteridae	9b	XP_034415207	XM_034559316
Notothenia	*coriiceps**	Actinopterygii	Perciformes	Nototheniidae	9b	XP_010786305	XM_010788003
Perca	*flavescens**	Actinopterygii	Perciformes	Percidae	9b	XP_028429891	XM_028574090
Sparus	*aurata*	Actinopterygii	Spariformes	Sparidae	9b	T1T061	KC589387
Takifugu	*rubripes**	Actinopterygii	Tetraodontiformes	Tetraodontidae	9b	XP_003970028	XM_003969979
Anarrhichthys	*ocellatus**	Actinopterygii		Anarhichadidae	9b	XP_031702151	XM_031846291

Contd...

Contd...

Genus	Species	Class	Order	Family	AQP	GB/U AA acc.	GB/U Nu acc.
Larimichthys	*crocea**	Actinopterygii		Sciaenidae	9b	XP_027136899	XM_027281098
Cebidichthys	*violaceus**	Actinopterygii		Stichaeidae	9b		NJBE01000429
Erpetoichthys	*calabaricus**	Actinopterygii	Polypteriformes	Polypteridae	10a	XP_028649869	XM_028794036
Polypterus	*senegalus**	Actinopterygii	Polypteriformes	Polypteridae	10a	XP_039607769	XM_039751835
Acipenser	*ruthenus**	Actinopterygii	Acipenseriformes	Acipenseridae	10a	XP_034775872	XM_034919981
Lepisosteus	*oculatus**	Actinopterygii	Lepisosteiformes	Lepisosteidae	10a	XP_015191988	XM_015336502
Scleropages	*formosus**	Actinopterygii	Osteoglossiformes	Osteoglossidae	10a	XP_018590830	XM_018735314
Astyanax	*mexicanus**	Actinopterygii	Characiformes	Characidae	10a	XP_015462989	XM_015607503
Cyprinus	*carpio*	Actinopterygii	Cypriniformes	Cyprinidae	10a	A0A0K2RVT9	LC069014
Danio	*rerio*	Actinopterygii	Cypriniformes	Danionidae	10a	NP_001002349	NM_001002349
Clarias	*batrachus**	Actinopterygii	Siluriformes	Clariidae	10a		QMIH01000447
Ictalurus	*punctatus**	Actinopterygii	SIluriformes	Ictaluridae	10a	XP_017324383	XM_017468894
Pangasianodon	*hypophthalmus**	Actinopterygii	Siluriformes	Pangasiidae	10a	XP_026795883	XM_026940082
Esox	*lucius**	Actinopterygii	Esociformes	Esocidae	10a	XP_019896571	XM_020041012
Salmo	*salar*	Actinopterygii	Salmoniformes	Salmonidae	10a	XP_014007427	XM_014151952
Gadus	*morhua**	Actinopterygii	Gadiformes	Gadidae	10a	XP_030225527	XM_030369667
Gadus	*morhua**	Actinopterygii	Gadiformes	Gadidae	10ab	XP_030202662	XM_030346802
Myripristis	*murdjan**	Actinopterygii	Holocentriformes	Holocentridae	10a	XP_029927979	XM_030072119
Lesueurigobius	*sanzi**	Actinopterygii	Gobiiformes	Gobiidae	10a		OMNZ01032421
Neogobius	*melanostomus**	Actinopterygii	Gobiiformes	Gobiidae	10a		VHKM01000058

Contd...

Contd...

Boleophthalmus	*pectinirostris*	Actinopterygii	Gobiiformes	Oxudercidae	10a	XP_020783751	KN524664
Periophthalmodon	*schlosseri*	Actinopterygii	Gobiiformes	Oxudercidae	10a		KN473499
Periophthalmus	*magnuspinnatus*	Actinopterygii	Gobiiformes	Oxudercidae	10a	XP_033837260	XM_033981369
Scartelaos	*histophorus*	Actinopterygii	Gobiiformes	Oxudercidae	10a		JACN01051648
Anabas	*testudineus**	Actinopterygii	Anabantiformes	Anabantidae	10a	XP_0262272181	XM_026371433
Helostoma	*temminki**	Actinopterygii	Anabantiformes	Helostomatidae	10a		OMLM01002228
Betta	*splendens**	Actinopterygii	Anabantiformes	Osphronemidae	10a	XP_028985137	XM_029129304
Cynoglossus	*semilaevis**	Actinopterygii	Pleuronectiformes	Cynoglossidae	10a	XP_008321325	XM_008323103
Mastacembelus	*armatus**	Actinopterygii	Synbranchiformes	Mastacembelidae	10a	XP_026172915	XM_026317130
Monopterus	*albus**	Actinopterygii	Synbranchiformes	Synbranchidae	10a	XP_020462852	XM_020607196
Parablennius	*parvicornis**	Actinopterygii	Blenniiformes	Blenniidae	10a		OMNM01003748
Salarias	*fasciatus**	Actinopterygii	Blenniiformes	Blenniidae	10a	XP_029958174	XM_030102314
Gouania	*willdenowi**	Actinopterygii	Gobiesociformes	Gobiesocidae	10a	XP_028326010	XM_028470209
Oryzias	*latipes*	Actinopterygii	Beloniformes	Adrianichthyidae	10a	XP_004078186	XM_004078138
Maylandia	*zebra**	Actinopterygii	Cichliformes	Cichlidae	10a	XP_004541352	XM_004541295
Oreochromis	*niloticus**	Actinopterygii	Cichliformes	Cichlidae	10a	XP_003451010	XM_003450962
Anableps	*anableps**	Actinopterygii	Cyprinodontiformes	Anablepidae	10a1		CM026072
Cyprinodon	*variegatus**	Actinopterygii	Cyprinodontiformes	Cyprinodontidae	10a1	XP_015236041	XM_015380555
Fundulus	*heteroclitus**	Actinopterygii	Cyprinodontiformes	Fundulidae	10a1	XP_012721291	XM_012865837
Aphyosemion	*australe**	Actinopterygii	Cyprinodontiformes	Nothobranchiidae	10a1		SSNS01000005
Nothobranchius	*furzeri**	Actinopterygii	Cyprinodontiformes	Nothobranchiidae	10a1	XP_015805944	XM_015950458

Contd...

Contd...

Genus	Species	Class	Order	Family	AQP	GB/U AA acc.	GB/U Nu acc.
Gambusia	*affinis**	Actinopterygii	Cyprinodontiformes	Poeciliidae	10a1		NHOQ01001560
Poecilia	*reticulata**	Actinopterygii	Cyprinodontiformes	Poeciliidae	10a1	XP_008429913	XM_008431691
Anableps	*anableps**	Actinopterygii	Cyprinodontiformes	Anablepidae	10a2		CM026072
Cyprinodon	*variegatus**	Actinopterygii	Cyprinodontiformes	Cyprinodontidae	10a2	XP_015236029	XM_015380543
Fundulus	*heteroclitus**	Actinopterygii	Cyprinodontiformes	Fundulidae	10a2	XP_012721292	XM_012865838
Austrofundulus	*limnaeus**	Actinopterygii	Cyprinodontiformes	Rivulidae	10a	XP_013865158	XM_014009704
Kryptolebias	*marmoratus**	Actinopterygii	Cyprinodontiformes	Rivulidae	10a	XP_0172804801	XM_017424991
Gymnodraco	*acuticeps**	Actinopterygii	Perciformes	Bathydraconidae	10a	XP_034083884	XM_034227993
Cyclopterus	*lumpus**	Actinopterygii	Perciformes	Cyclopteridae	10a	XP_034409969	XM_034554078
Notothenia	*coriiceps**	Actinopterygii	Perciformes	Nototheniidae	10a	XP_010785345	XM_010787043
Perca	*flavescens**	Actinopterygii	Perciformes	Percidae	10a	XP_028436574	XM_028580773
Takifugu	*rubripes**	Actinopterygii	Tetraodontiformes	Tetraodontidae	10a	XP_011604131	XM_011605829
Anarrhichthys	*ocellatus**	Actinopterygii		Anarhichadidae	10a	XP_031708750	XM_031852890
Larimichthys	*crocea**	Actinopterygii		Sciaenidae	10a	XP_019126985	XM_019271440
Cebidichthys	*violaceus**	Actinopterygii		Stichaeidae	10a		NJBE01000427
Latimeria	*chalumnae**	Sarcopterygii			10	XP_005996556	XM_005996494
Xenopus	*laevis*	Amphibia			10	XP_041429684	XM_041573750
Alligator	*mississippiensis*	Reptilia			10	XP_006264432	XM_006264370
Gallus	*gallus*	Aves			10	XP_015154084	XM_015298598
Homo	*sapiens*	Mammalia			10	NP_536354	NM_080429

Contd...

Contd...

Erpetoichthys	*calabaricus**	Actinopterygii	Polypteriformes	Polypteridae	10b	XP_028649868	XM_028794035
Polypterus	*senegalus**	Actinopterygii	Polypteriformes	Polypteridae	10b	XP_039607747	XM_039751813
Acipenser	*ruthenus**	Actinopterygii	Acipenseriformes	Acipenseridae	10b	XP_033849960	XM_033994069
Anguilla	*anguilla**	Actinopterygii	Anguiliformes	Anguilidae	10b	XP_035284278	XM_035428387
Astyanax	*mexicanus**	Actinopterygii	Characiformes	Characidae	10b	XP_007230576	XM_007230514
Cyprinus	*carpio*	Actinopterygii	Cypriniformes	Cyprinidae	10b	A0A0K2RVX4	LC069016
Danio	*rerio*	Actinopterygii	Cypriniformes	Danionidae	10b	XP_005159449	XM_005159392
Clarias	*batrachus**	Actinopterygii	Siluriformes	Clariidae	10b		QMIH01002140
Ictalurus	*punctatus**	Actinopterygii	SIluriformes	Ictaluridae	10b	XP_017336685	XM_017481196
Pangasianodon	*hypophthalmus**	Actinopterygii	Siluriformes	Pangasiidae	10b	XP_034157370	XM_034301479
Esox	*lucius**	Actinopterygii	Esociformes	Esocidae	10b	XP_010866254	XM_010867952
Salmo	*salar*	Actinopterygii	Salmoniformes	Salmonidae	10b	XP_013997692	XM_014142217
Gadus	*morhua**	Actinopterygii	Gadiformes	Gadidae	10b	XP_030202659	XM_030346799
Myripristis	*murdjan**	Actinopterygii	Holocentriformes	Holocentridae	10b	XP_029919634	XM_030063774
Lesueurigobius	*sanzi**	Actinopterygii	Gobiiformes	Gobiidae	10b		OMNZ01032590
Neogobius	*melanostomus**	Actinopterygii	Gobiiformes	Gobiidae	10b		VHKM01000212
Boleophthalmus	*pectinirostris*	Actinopterygii	Gobiiformes	Oxudercidae	10b	XP_020783751	XM_020928092
Periophthalmodon	*schlosseri*	Actinopterygii	Gobiiformes	Oxudercidae	10b		KN474915
Periophthalmus	*magnuspinnatus*	Actinopterygii	Gobiiformes	Oxudercidae	10b	XP_033831689	XM_033975798
Scartelaos	*histophorus*	Actinopterygii	Gobiiformes	Oxudercidae	10b		JACN01031370
Sphaeramia	*orbicularis**	Actinopterygii	Kurtiformes	Apogonidae	10b1	XP_029979788	XM_030123928

Contd...

Contd...

Genus	Species	Class	Order	Family	AQP	GB/U AA acc.	GB/U Nu acc.
Sphaeramia	*orbicularis**	Actinopterygii	Kurtiformes	Apogonidae	10b2	XP_030004095	XM_030148235
Anabas	*testudineus**	Actinopterygii	Anabantiformes	Anabantidae	10b	XP_0262042871	XM_026348502
Channa	*argus**	Actinopterygii	Anabantiformes	Channidae	10b		CM015712
Helostoma	*temminki**	Actinopterygii	Anabantiformes	Helostomatidae	10b		OMLM01010766
Betta	*splendens**	Actinopterygii	Anabantiformes	Osphronemidae	10b	XP_029023866	XM_029168033
Cynoglossus	*semilaevis**	Actinopterygii	Pleuronectiformes	Cynoglossidae	10b	XP_008330120	XM_008331898
Scophthalmus	*maximus**	Actinopterygii	Pleuronectiformes	Scophthalmidae	10b	AWP21524	
Mastacembelus	*armatus**	Actinopterygii	Synbranchiformes	Mastacembelidae	10b	XP_026156039	XM_026300254
Monopterus	*albus**	Actinopterygii	Synbranchiformes	Synbranchidae	10b	XP_020479476	XM_020623821
Parablennius	*parvicornis**	Actinopterygii	Blenniiformes	Blenniidae	10b		OMNM01045385
Salarias	*fasciatus**	Actinopterygii	Blenniiformes	Blenniidae	10b	XP_029976305	XM_030120442
Gouania	*willdenowi**	Actinopterygii	Gobiesociformes	Gobiesocidae	10b	XP_028317709	XM_028461908
Oryzias	*latipes*	Actinopterygii	Beloniformes	Adrianichthyidae	10b	XP_004073758	XM_004073710
Maylandia	*zebra**	Actinopterygii	Cichliformes	Cichlidae	10b	XP_004549151	XM_004549094
Oreochromis	*niloticus**	Actinopterygii	Cichliformes	Cichlidae	10b	XP_003452499	XM_003452451
Anableps	*anableps**	Actinopterygii	Cyprinodontiformes	Anablepidae	10b		CM026079
Cyprinodon	*variegatus**	Actinopterygii	Cyprinodontiformes	Cyprinodontidae	10b	XP_015255369	XM_015399883
Fundulus	*heteroclitus**	Actinopterygii	Cyprinodontiformes	Fundulidae	10b	XP_012730801	XM_012875347
Aphyosemion	*australe**	Actinopterygii	Cyprinodontiformes	Nothobranchiidae	10b		SSNS01000031
Nothobranchius	*furzeri**	Actinopterygii	Cyprinodontiformes	Nothobranchiidae	10b	XP_015829167	XM_015973681

Contd...

Contd...

Gambusia	*affinis**	Actinopterygii	Cyprinodontiformes	Poeciliidae	10b		NHOQ01002733
Poecilia	*reticulata**	Actinopterygii	Cyprinodontiformes	Poeciliidae	10b	XP_008420750	XM_008422528
Austrofundulus	*limnaeus**	Actinopterygii	Cyprinodontiformes	Rivulidae	10b	XP_013878909	XM_014023455
Kryptolebias	*marmoratus**	Actinopterygii	Cyprinodontiformes	Rivulidae	10b	XP_0172785701	XM_017423081
Gymnodraco	*acuticeps**	Actinopterygii	Perciformes	Bathydraconidae	10b	XP_034066229	XM_034210338
Cyclopterus	*lumpus**	Actinopterygii	Perciformes	Cyclopteridae	10b	XP_034400951	XM_034545060
Notothenia	*coriiceps**	Actinopterygii	Perciformes	Nototheniidae	10b	XP_010768607	XM_010770305
Perca	*flavescens**	Actinopterygii	Perciformes	Percidae	10b	XP_028453410	XM_028597609
Sparus	*aurata*	Actinopterygii	Spariformes	Sparidae	10b	Q67EP9	AY363261
Takifugu	*rubripes**	Actinopterygii	Tetraodontiformes	Tetraodontidae	10b	XP_003969331	XM_003969282
Anarrhichthys	*ocellatus**	Actinopterygii		Anarhichadidae	10b	XP_031722270	XM_031866410
Larimichthys	*crocea**	Actinopterygii		Sciaenidae	10b	XP_019119842	XM_019264297
Cebidichthys	*violaceus**	Actinopterygii		Stichaeidae	10b		NJBE01000435
Erpetoichthys	*calabaricus**	Actinopterygii	Polypteriformes	Polypteridae	11	XP_028656204	XM_028800371
Polypterus	*senegalus**	Actinopterygii	Polypteriformes	Polypteridae	11	XP_039600162	XM_039744228
Acipenser	*ruthenus**	Actinopterygii	Acipenseriformes	Acipenseridae	11	XP_033865482	XM_034009591
Lepisosteus	*oculatus**	Actinopterygii	Lepisosteiformes	Lepisosteidae	11	XP_015196614	XM_015341128
Anguilla	*anguilla**	Actinopterygii	Anguiliformes	Anguilidae	11a	XP_035240102	XM_035384211
Scleropages	*formosus**	Actinopterygii	Osteoglossiformes	Osteoglossidae	11a	XP_029111223	XM_029255390
Astyanax	*mexicanus**	Actinopterygii	Characiformes	Characidae	11a	XP_007257850	XM_007257788
Cyprinus	*carpio*	Actinopterygii	Cypriniformes	Cyprinidae	11a	XP_018935958	XM_019080413

Contd...

Contd...

Genus	Species	Class	Order	Family	AQP	GB/U AA acc.	GB/U Nu acc.
Danio	*rerio*	Actinopterygii	Cypriniformes	Danionidae	11a	NP_001314822	NM_001327893
Clarias	*batrachus**	Actinopterygii	Siluriformes	Clariidae	11a		QMIH01000742
Ictalurus	*punctatus**	Actinopterygii	SIluriformes	Ictaluridae	11a	XP_017322285	XM_017466796
Pangasianodon	*hypophthalmus**	Actinopterygii	Siluriformes	Pangasiidae	11a	XP_026769139	XM_026913338
Esox	*lucius**	Actinopterygii	Esociformes	Esocidae	11a	XP_019905541	XM_020049982
Salmo	*salar*	Actinopterygii	Salmoniformes	Salmonidae	11a	XP_014069966	XM_014214491
Gadus	*morhua**	Actinopterygii	Gadiformes	Gadidae	11a	XP_030237689	XM_030381829
Myripristis	*murdjan**	Actinopterygii	Holocentriformes	Holocentridae	11a	XP_029922678	XM_030066818
Lesueurigobius	*sanzi**	Actinopterygii	Gobiiformes	Gobiidae	11a		OMNZ01074194
Neogobius	*melanostomus**	Actinopterygii	Gobiiformes	Gobiidae	11a		VHKM01001121
Boleophthalmus	*pectinirostris*	Actinopterygii	Gobiiformes	Oxudercidae	11a	XP_020794314	XM_020938655
Periophthalmodon	*schlosseri*	Actinopterygii	Gobiiformes	Oxudercidae	11a		KN469785
Periophthalmus	*magnuspinnatus*	Actinopterygii	Gobiiformes	Oxudercidae	11a	XP_033832428	XM_033976537
Scartelaos	*histophorus*	Actinopterygii	Gobiiformes	Oxudercidae	11a		JACN01020100
Sphaeramia	*orbicularis**	Actinopterygii	Kurtiformes	Apogonidae	11a	XP_030007154	XM_030151294
Anabas	*testudineus**	Actinopterygii	Anabantiformes	Anabantidae	11a	XP_0261947751	XM_026338990
Channa	*argus**	Actinopterygii	Anabantiformes	Channidae	11a		CM015719
Helostoma	*temminki**	Actinopterygii	Anabantiformes	Helostomatidae	11a		OMLM01018687
Betta	*splendens**	Actinopterygii	Anabantiformes	Osphronemidae	11a	XP_029027629	XM_029171796
Cynoglossus	*semilaevis**	Actinopterygii	Pleuronectiformes	Cynoglossidae	11a	XP_008307169	XM_008308947

Contd...

Contd...

Scophthalmus	*maximus**	Actinopterygii	Pleuronectiformes	Scophthalmidae	11a	AWO99237	
Mastacembelus	*armatus**	Actinopterygii	Synbranchiformes	Mastacembelidae	11a	XP_026150374	XM_026294589
Monopterus	*albus**	Actinopterygii	Synbranchiformes	Synbranchidae	11a	XP_020479216	XM_020623560
Salarias	*fasciatus**	Actinopterygii	Blenniiformes	Blenniidae	11a	XP_029964225	XM_030108365
Gouania	*willdenowi**	Actinopterygii	Gobiesociformes	Gobiesocidae	11a	XP_028320718	XM_028464917
Oryzias	*latipes*	Actinopterygii	Beloniformes	Adrianichthyidae	11a	XP_023817466	XM_023961698
Maylandia	*zebra**	Actinopterygii	Cichliformes	Cichlidae	11a	XP_004569715	XM_004569658
Oreochromis	*niloticus**	Actinopterygii	Cichliformes	Cichlidae	11a	XP_019200806	XM_019345261
Anableps	*anableps**	Actinopterygii	Cyprinodontiformes	Anablepidae	11a		CM026068
Cyprinodon	*variegatus**	Actinopterygii	Cyprinodontiformes	Cyprinodontidae	11a	XP_015241118	XM_015385632
Fundulus	*heteroclitus**	Actinopterygii	Cyprinodontiformes	Fundulidae	11a	XP_012708786	XM_012853332
Aphyosemion	*australe**	Actinopterygii	Cyprinodontiformes	Nothobranchiidae	11a		SSNS01000440
Nothobranchius	*furzeri**	Actinopterygii	Cyprinodontiformes	Nothobranchiidae	11a	XP_015806156	XM_015950670
Gambusia	*affinis**	Actinopterygii	Cyprinodontiformes	Poeciliidae	11a		NHOQ01001755
Poecilia	*reticulata**	Actinopterygii	Cyprinodontiformes	Poeciliidae	11a	XP_008424713	XM_008426491
Austrofundulus	*limnaeus**	Actinopterygii	Cyprinodontiformes	Rivulidae	11a	XP_013871377	XM_014015923
Kryptolebias	*marmoratus**	Actinopterygii	Cyprinodontiformes	Rivulidae	11a	XP_0172934911	XM_017438002
Cyclopterus	*lumpus**	Actinopterygii	Perciformes	Cyclopteridae	11a	XP_034384747	XM_034528856
Perca	*flavescens**	Actinopterygii	Perciformes	Percidae	11a	XP_028427453	XM_028571652
Takifugu	*rubripes**	Actinopterygii	Tetraodontiformes	Tetraodontidae	11a	XP_003968270	XM_003968221
Anarrhichthys	*ocellatus**	Actinopterygii		Anarhichadidae	11a	XP_031734565	XM_031878705

Contd...

Contd...

Genus	Species	Class	Order	Family	AQP	GB/U AA acc.	GB/U Nu acc.
Larimichthys	*crocea**	Actinopterygii		Sciaenidae	11a	XP_027135851	XM_027280050
Cebidichthys	*violaceus**	Actinopterygii		Stichaeidae	11a		NJBE01000193
Latimeria	*chalumnae**	Sarcopterygii			11	XP_006006366	XM_006006304
Xenopus	*laevis*	Amphibia			11	XP_018103356	XM_018247868
Alligator	*mississippiensis*	Reptilia			11	XP_006272507	XM_006272445
Chelonia	*mydas*	Reptilia			11	XP_007067464	XM_007067402
Gallus	*gallus*	Aves			11	XP_015136362	XM_015280878
Homo	*sapiens*	Mammalia			11	NP_766627	NM_173039
Mus	*musculus*	Mammalia			11	NP_780314	NM_175105
Anguilla	*anguilla**	Actinopterygii	Anguiliformes	Anguilidae	11b	XP_035287692	XM_035431801
Astyanax	*mexicanus**	Actinopterygii	Characiformes	Characidae	11b	XP_007244849	XM_007244787
Esox	*lucius**	Actinopterygii	Esociformes	Esocidae	11b	XP_010869109	XM_010870807
Salmo	*salar*	Actinopterygii	Salmoniformes	Salmonidae	11b	XP_014053194	XM_014197719
Gadus	*morhua**	Actinopterygii	Gadiformes	Gadidae	11b	XP_030218097	XM_030362237
Myripristis	*murdjan**	Actinopterygii	Holocentriformes	Holocentridae	11b	XP_029925517	XM_030069657
Lesueurigobius	*sanzi**	Actinopterygii	Gobiiformes	Gobiidae	11b		OMNZ01089478
Neogobius	*melanostomus**	Actinopterygii	Gobiiformes	Gobiidae	11b		VHKM01000583
Boleophthalmus	*pectinirostris*	Actinopterygii	Gobiiformes	Oxudercidae	11b		KN522455
Periophthalmodon	*schlosseri*	Actinopterygii	Gobiiformes	Oxudercidae	11b		KN480662
Periophthalmus	*magnuspinnatus*	Actinopterygii	Gobiiformes	Oxudercidae	11b	XP_033834468	XM_033978577

Contd...

Contd...

Scartelaos	*histophorus*	Actinopterygii	Gobiiformes	Oxudercidae	11b		JACN01076962/01183528
Sphaeramia	*orbicularis**	Actinopterygii	Kurtiformes	Apogonidae	11b	XP_030010915	XM_030155055
Anabas	*testudineus**	Actinopterygii	Anabantiformes	Anabantidae	11b	XP_0262269681	XM_026371183
Channa	*argus**	Actinopterygii	Anabantiformes	Channidae	11b		CM015733
Helostoma	*temminki**	Actinopterygii	Anabantiformes	Helostomatidae	11b		OMLM01014267
Betta	*splendens**	Actinopterygii	Anabantiformes	Osphronemidae	11b	XP_029028617	XM_029172784
Cynoglossus	*semilaevis**	Actinopterygii	Pleuronectiformes	Cynoglossidae	11b	XP_008331251	XM_008333029
Scophthalmus	*maximus**	Actinopterygii	Pleuronectiformes	Scophthalmidae	11b	AWO99707	
Mastacembelus	*armatus**	Actinopterygii	Synbranchiformes	Mastacembelidae	11b	XP_026183531	XM_026327746
Monopterus	*albus**	Actinopterygii	Synbranchiformes	Synbranchidae	11b	XP_020454278	XM_020598622
Parablennius	*parvicornis**	Actinopterygii	Blenniiformes	Blenniidae	11b		OMNM01003901
Salarias	*fasciatus**	Actinopterygii	Blenniiformes	Blenniidae	11b	XP_029957411	XM_030101551
Gouania	*willdenowi**	Actinopterygii	Gobiesociformes	Gobiesocidae	11b	XP_028322591	XM_028466790
Oryzias	*latipes*	Actinopterygii	Beloniformes	Adrianichthyidae	11b	XP_004076889	XM_004076841
Maylandia	*zebra**	Actinopterygii	Cichliformes	Cichlidae	11b	XP_004551245	XM_004551188
Oreochromis	*niloticus**	Actinopterygii	Cichliformes	Cichlidae	11b	XP_003453785	XM_003453737
Anableps	*anableps**	Actinopterygii	Cyprinodontiformes	Anablepidae	11b		CM026078
Cyprinodon	*variegatus**	Actinopterygii	Cyprinodontiformes	Cyprinodontidae	11b	XP_015243376	XM_015387890
Fundulus	*heteroclitus**	Actinopterygii	Cyprinodontiformes	Fundulidae	11b	XP_012711394	XM_012855940
Aphyosemion	*australe**	Actinopterygii	Cyprinodontiformes	Nothobranchiidae	11b		SSNS01000164
Nothobranchius	*furzeri**	Actinopterygii	Cyprinodontiformes	Nothobranchiidae	11b	XP_015817371	XM_015961885

Contd...

Contd...

Genus	Species	Class	Order	Family	AQP	GB/U AA acc.	GB/U Nu acc.
Gambusia	*affinis**	Actinopterygii	Cyprinodontiformes	Poeciliidae	11b		NHOQ01001433
Poecilia	*reticulata**	Actinopterygii	Cyprinodontiformes	Poeciliidae	11b	XP_008426765	XM_008428543
Austrofundulus	*limnaeus**	Actinopterygii	Cyprinodontiformes	Rivulidae	11b	XP_013859464	XM_014004010
Kryptolebias	*marmoratus**	Actinopterygii	Cyprinodontiformes	Rivulidae	11b	XP_0172921931	NW_016094558
Cyclopterus	*lumpus**	Actinopterygii	Perciformes	Cyclopteridae	11b	XP_034406205	XM_034550314
Notothenia	*coriiceps**	Actinopterygii	Perciformes	Nototheniidae	11b	XP_010775066	XM_010776764
Perca	*flavescens**	Actinopterygii	Perciformes	Percidae	11b	XP_028451768	XM_028595967
Takifugu	*rubripes**	Actinopterygii	Tetraodontiformes	Tetraodontidae	11b	XP_003971010	XM_003970961
Anarrhichthys	*ocellatus**	Actinopterygii		Anarhichadidae	11b	XP_031729968	XM_031874108
Larimichthys	*crocea**	Actinopterygii		Sciaenidae	11b	XP_010729426	XM_010731124
Cebidichthys	*violaceus**	Actinopterygii		Stichaeidae	11b		NJBE01000063
Erpetoichthys	*calabaricus**	Actinopterygii	Polypteriformes	Polypteridae	12	XP_028648920	XM_028793087
Polypterus	*senegalus**	Actinopterygii	Polypteriformes	Polypteridae	12	XP_039623657	XM_039767723
Acipenser	*ruthenus**	Actinopterygii	Acipenseriformes	Acipenseridae	12	XP_033894465	XM_034038574
Lepisosteus	*oculatus**	Actinopterygii	Lepisosteiformes	Lepisosteidae	12	XP_006637810	XM_006637747
Anguilla	*anguilla**	Actinopterygii	Anguiliformes	Anguilidae	12	XP_035270531	XM_035414640
Scleropages	*formosus**	Actinopterygii	Osteoglossiformes	Osteoglossidae	12	XP_018596485	XM_018740969
Astyanax	*mexicanus**	Actinopterygii	Characiformes	Characidae	12	XP_007252664	XM_007252602
Cyprinus	*carpio*	Actinopterygii	Cypriniformes	Cyprinidae	12	XP_018920906	XM_019065361
Danio	*rerio*	Actinopterygii	Cypriniformes	Danionidae	12	NP_001039327	NM_001045862

Contd...

Contd...

Clarias	*batrachus**	Actinopterygii	Siluriformes	Clariidae	12		QMIH01000089
Ictalurus	*punctatus**	Actinopterygii	SIluriformes	Ictaluridae	12	XP_017334448	XM_017478959
Pangasianodon	*hypophthalmus**	Actinopterygii	Siluriformes	Pangasiidae	12	XP_026786130	XM_026930329
Salmo	*salar*	Actinopterygii	Salmoniformes	Salmonidae	12	XP_014025041	XM_014169566
Gadus	*morhua**	Actinopterygii	Gadiformes	Gadidae	12	XP_030228996	XM_030373136
Myripristis	*murdjan**	Actinopterygii	Holocentriformes	Holocentridae	12	XP_029905906	XM_030050046
Lesueurigobius	*sanzi**	Actinopterygii	Gobiiformes	Gobiidae	12		OMNZ01109090
Neogobius	*melanostomus**	Actinopterygii	Gobiiformes	Gobiidae	12		VHKM01000494
Boleophthalmus	*pectinirostris*	Actinopterygii	Gobiiformes	Oxudercidae	12	XP_020774481	XM_020918822
Periophthalmodon	*schlosseri*	Actinopterygii	Gobiiformes	Oxudercidae	12		KN477140
Periophthalmus	*magnuspinnatus*	Actinopterygii	Gobiiformes	Oxudercidae	12	XP_033821434	XM_033965543
Scartelaos	*histophorus*	Actinopterygii	Gobiiformes	Oxudercidae	12		KN493770
Sphaeramia	*orbicularis**	Actinopterygii	Kurtiformes	Apogonidae	12	XP_029988528	XM_030132668
Anabas	*testudineus**	Actinopterygii	Anabantiformes	Anabantidae	12	XP_0262011851	XM_026345400
Channa	*argus**	Actinopterygii	Anabantiformes	Channidae	12		CM015717
Helostoma	*temminki**	Actinopterygii	Anabantiformes	Helostomatidae	12		OMLM01014674
Betta	*splendens**	Actinopterygii	Anabantiformes	Osphronemidae	12	XP_029003438	XM_029147605
Cynoglossus	*semilaevis**	Actinopterygii	Pleuronectiformes	Cynoglossidae	12	XP_024909953	XM_025054185
Scophthalmus	*maximus**	Actinopterygii	Pleuronectiformes	Scophthalmidae	12	AWP11260	
Mastacembelus	*armatus**	Actinopterygii	Synbranchiformes	Mastacembelidae	12	XP_026160559	NW_020535868
Monopterus	*albus**	Actinopterygii	Synbranchiformes	Synbranchidae	12	XP_020459066	XM_020603410

Contd...

Contd...

Genus	Species	Class	Order	Family	AQP	GB/U AA acc.	GB/U Nu acc.
Parablennius	*parvicornis**	Actinopterygii	Blenniiformes	Blenniidae	12		OMNM01037980
Salarias	*fasciatus**	Actinopterygii	Blenniiformes	Blenniidae	12	XP_029938854	XM_030082994
Gouania	*willdenowi**	Actinopterygii	Gobiesociformes	Gobiesocidae	12	XP_028300609	XM_028444808
Oryzias	*latipes*	Actinopterygii	Beloniformes	Adrianichthyidae	12	XP_004067662	XM_004067614
Maylandia	*zebra**	Actinopterygii	Cichliformes	Cichlidae	12	XP_004557241	XM_004557184
Oreochromis	*niloticus**	Actinopterygii	Cichliformes	Cichlidae	12	XP_003439959	XM_003439911
Anableps	*anableps**	Actinopterygii	Cyprinodontiformes	Anablepidae	12		CM026070
Cyprinodon	*variegatus**	Actinopterygii	Cyprinodontiformes	Cyprinodontidae	12	XP_015253060	XM_015397574
Fundulus	*heteroclitus**	Actinopterygii	Cyprinodontiformes	Fundulidae	12	XP_012721819	XM_012866365
Aphyosemion	*australe**	Actinopterygii	Cyprinodontiformes	Nothobranchiidae	12		SSNS01000082
Nothobranchius	*furzeri**	Actinopterygii	Cyprinodontiformes	Nothobranchiidae	12	XP_015827247	XM_015971761
Gambusia	*affinis**	Actinopterygii	Cyprinodontiformes	Poeciliidae	12		NHOQ01001926
Poecilia	*reticulata**	Actinopterygii	Cyprinodontiformes	Poeciliidae	12	XP_008405621	XM_008407399
Austrofundulus	*limnaeus**	Actinopterygii	Cyprinodontiformes	Rivulidae	12	XP_013877961	XM_014022507
Kryptolebias	*marmoratus**	Actinopterygii	Cyprinodontiformes	Rivulidae	12	XP_0172843241	XM_017428836
Gymnodraco	*acuticeps**	Actinopterygii	Perciformes	Bathydraconidae	12	XP_034085806	XM_034229915
Cyclopterus	*lumpus**	Actinopterygii	Perciformes	Cyclopteridae	12	XP_034387289	XM_034531398
Notothenia	*coriiceps**	Actinopterygii	Perciformes	Nototheniidae	12	XP_010789442	XM_010791140
Perca	*flavescens**	Actinopterygii	Perciformes	Percidae	12	XP_028444118	XM_028588317
Takifugu	*rubripes**	Actinopterygii	Tetraodontiformes	Tetraodontidae	12	XP_003973906	XM_003973857

Contd...

Contd...

Anarrhichthys	*ocellatus**	Actinopterygii		Anarhichadidae	12	XP_031724167	XM_031868307
Larimichthys	*crocea**	Actinopterygii		Sciaenidae	12	XP_010729348	XM_010731046
Cebidichthys	*violaceus**	Actinopterygii		Stichaeidae	12		NJBE01000129
Latimeria	*chalumnae**	Sarcopterygii			12	XP_006009788	XM_006009726
Xenopus	*laevis*	Amphibia			12	NP_001088119	NM_001094650
Alligator	*mississippiensis*	Reptilia			12	XP_006267107	XM_006267045
Chelonia	*mydas*	Reptilia			12	XP_007057031	XM_007056969
Gallus	*gallus*	Aves			12	NP_001103149	NM_001109679
Homo	*sapiens*	Mammalia			12	NP_945349	NM_198998
Mus	*musculus*	Mammalia			12	NP_808255	NM_177587
Erpetoichthys	*calabaricus**	Actinopterygii	Polypteriformes	Polypteridae	14	XP_028654863	XM_028799030
Polypterus	*senegalus**	Actinopterygii	Polypteriformes	Polypteridae	14	XP_039605586	XM_039749652
Acipenser	*ruthenus**	Actinopterygii	Acipenseriformes	Acipenseridae	14	XP_034769939	XM_034914048
Lepisosteus	*oculatus**	Actinopterygii	Lepisosteiformes	Lepisosteidae	14	XP_015199546	XM_015344060
Anguilla	*anguilla**	Actinopterygii	Anguiliformes	Anguilidae	14	XP_035243307	XM_035387416
Scleropages	*formosus**	Actinopterygii	Osteoglossiformes	Osteoglossidae	14	XP_018580791	XM_018725275
Astyanax	*mexicanus**	Actinopterygii	Characiformes	Characidae	14	XP_007233063	XM_007233001
Cyprinus	*carpio*	Actinopterygii	Cypriniformes	Cyprinidae	14	XP_018931706	XM_019076161
Danio	*rerio*	Actinopterygii	Cypriniformes	Danionidae	14	XP_005174182	XM_005174125
Clarias	*batrachus**	Actinopterygii	Siluriformes	Clariidae	14		QMIH01000020
Ictalurus	*punctatus**	Actinopterygii	SIluriformes	Ictaluridae	14	XP_017342088	XM_017486599

Contd...

Contd...

Genus	Species	Class	Order	Family	AQP	GB/U AA acc.	GB/U Nu acc.
Pangasianodon	*hypophthalmus**	Actinopterygii	Siluriformes	Pangasiidae	14	XP_034167579	XM_034311688
Esox	*lucius**	Actinopterygii	Esociformes	Esocidae	14	XP_012992002	XM_013136548
Salmo	*salar*	Actinopterygii	Salmoniformes	Salmonidae	14	XP_014001001	XM_014145526
Gadus	*morhua**	Actinopterygii	Gadiformes	Gadidae	14	XP_030206125	XM_030350265
Myripristis	*murdjan**	Actinopterygii	Holocentriformes	Holocentridae	14	XP_029912148	XM_030056288
Lesueurigobius	*sanzi**	Actinopterygii	Gobiiformes	Gobiidae	14		OMNZ01100408
Neogobius	*melanostomus**	Actinopterygii	Gobiiformes	Gobiidae	14		VHKM01000223
Boleophthalmus	*pectinirostris*	Actinopterygii	Gobiiformes	Oxudercidae	14	XP_020785193	XM_020929534
Periophthalmus	*magnuspinnatus*	Actinopterygii	Gobiiformes	Oxudercidae	14	XP_033826021	XM_033970130
Sphaeramia	*orbicularis**	Actinopterygii	Kurtiformes	Apogonidae	14	XP_029995776	XM_030139916
Anabas	*testudineus**	Actinopterygii	Anabantiformes	Anabantidae	14	XP_0262245101	XM_026368725
Helostoma	*temminki**	Actinopterygii	Anabantiformes	Helostomatidae	14		OMLM01005575
Betta	*splendens**	Actinopterygii	Anabantiformes	Osphronemidae	14	XP_029013365	XM_029157532
Cynoglossus	*semilaevis**	Actinopterygii	Pleuronectiformes	Cynoglossidae	14	XP_024914792	XR_003050763
Mastacembelus	*armatus**	Actinopterygii	Synbranchiformes	Mastacembelidae	14	XP_026189054	XM_026333269
Monopterus	*albus**	Actinopterygii	Synbranchiformes	Synbranchidae	14	XP_020466413	XM_020610757
Parablennius	*parvicornis**	Actinopterygii	Blenniiformes	Blenniidae	14		OMNM01004137
Salarias	*fasciatus**	Actinopterygii	Blenniiformes	Blenniidae	14	XP_029975644	XM_030119784
Gouania	*willdenowi**	Actinopterygii	Gobiesociformes	Gobiesocidae	14	XP_028309293	XM_028453492
Oryzias	*latipes*	Actinopterygii	Beloniformes	Adrianichthyidae	14	XP_023812913	XM_023957145

Contd...

Contd...

Maylandia	*zebra**	Actinopterygii	Cichliformes	Cichlidae	14	XP_004546273	XM_004546216
Oreochromis	*niloticus**	Actinopterygii	Cichliformes	Cichlidae	14	XP_025757563	XM_025901778
Anableps	*anableps**	Actinopterygii	Cyprinodontiformes	Anablepidae	14		CM026071
Cyprinodon	*variegatus**	Actinopterygii	Cyprinodontiformes	Cyprinodontidae	14	XP_015239032	XM_015383546
Fundulus	*heteroclitus**	Actinopterygii	Cyprinodontiformes	Fundulidae	14	XP_012721398	XM_012865944
Nothobranchius	*furzeri**	Actinopterygii	Cyprinodontiformes	Nothobranchiidae	14	XP_015823776	XM_015968290
Gambusia	*affinis**	Actinopterygii	Cyprinodontiformes	Poeciliidae	14		NHOQ01000244
Poecilia	*reticulata**	Actinopterygii	Cyprinodontiformes	Poeciliidae	14	XP_008412386	XM_008414164
Austrofundulus	*limnaeus**	Actinopterygii	Cyprinodontiformes	Rivulidae	14	XP_013862724	XM_014007270
Kryptolebias	*marmoratus**	Actinopterygii	Cyprinodontiformes	Rivulidae	14	XP_0172649441	XM_017409455
Gymnodraco	*acuticeps**	Actinopterygii	Perciformes	Bathydraconidae	14	XP_034066744	XM_034210853
Cyclopterus	*lumpus**	Actinopterygii	Perciformes	Cyclopteridae	14	XP_034393648	XM_034537757
Notothenia	*coriiceps**	Actinopterygii	Perciformes	Nototheniidae	14	XP_010776696	XM_010778394
Perca	*flavescens**	Actinopterygii	Perciformes	Percidae	14	XP_028438369	XM_028582568
Takifugu	*rubripes**	Actinopterygii	Tetraodontiformes	Tetraodontidae	14	XP_003963480	XM_003963431
Anarrhichthys	*ocellatus**	Actinopterygii		Anarhichadidae	14	XP_031700268	XM_031844408
Larimichthys	*crocea**	Actinopterygii		Sciaenidae	14	XP_019131615	XM_019276070
Cebidichthys	*violaceus**	Actinopterygii		Stichaeidae	14		NJBE01000438
Xenopus	*laevis*	Amphibia			14	XP_018101033	XM_018245544
Alligator	*mississippiensis*	Reptilia			14	XP_019350933	XM_019495388
Chelonia	*mydas*	Reptilia			14	XP_037740203	XM_037884275

Contd...

Contd...

Genus	Species	Class	Order	Family	AQP	GB/U AA acc.	GB/U Nu acc.
Gallus	*gallus*	Aves			14	XP_015155778	XM_015300292
Erpetoichthys	*calabaricus**	Actinopterygii	Polypteriformes	Polypteridae	15	XP_028674304	XM_028818471
Polypterus	*senegalus**	Actinopterygii	Polypteriformes	Polypteridae	15	XP_039596489	XM_039740555
Acipenser	*ruthenus**	Actinopterygii	Acipenseriformes	Acipenseridae	15	XP_033911334	XM_034055443
Lepisosteus	*oculatus**	Actinopterygii	Lepisosteiformes	Lepisosteidae	15	XP_006638402	XM_006638339
Anguilla	*anguilla**	Actinopterygii	Anguiliformes	Anguilidae	15	XP_035253342	XM_035397451
Scleropages	*formosus**	Actinopterygii	Osteoglossiformes	Osteoglossidae	15	XP_029110169	
Astyanax	*mexicanus**	Actinopterygii	Characiformes	Characidae	15	XP_022520808	XM_022665087
Cyprinus	*carpio*	Actinopterygii	Cypriniformes	Cyprinidae	15	XP_018947106	XM_019091561
Danio	*rerio*	Actinopterygii	Cypriniformes	Danionidae	15	XP_021327889	XM_021472214
Esox	*lucius**	Actinopterygii	Esociformes	Esocidae	15	XP_010890856	XM_010892554
Salmo	*salar*	Actinopterygii	Salmoniformes	Salmonidae	15	XP_014048256	XM_014192781
Anableps	*anableps**	Actinopterygii	Cyprinodontiformes	Anablepidae	15		CM026083
Cyprinodon	*variegatus**	Actinopterygii	Cyprinodontiformes	Cyprinodontidae	15	XP_015225942	XM_015370456
Fundulus	*heteroclitus**	Actinopterygii	Cyprinodontiformes	Fundulidae	15	XP_021176951	XM_021321276
Aphyosemion	*australe**	Actinopterygii	Cyprinodontiformes	Nothobranchiidae	15		SSNS01000439
Gambusia	*affinis**	Actinopterygii	Cyprinodontiformes	Poeciliidae	15		NHOQ01001148
Poecilia	*reticulata**	Actinopterygii	Cyprinodontiformes	Poeciliidae	15	XP_008414717	XM_017306377
Austrofundulus	*limnaeus**	Actinopterygii	Cyprinodontiformes	Rivulidae	15	XP_013878635	XM_014023181
Latimeria	*chalumnae**	Sarcopterygii			15	XP_014343048	XM_014487562

Contd...

Contd...

Chelonia	*mydas*	Reptilia			15	XP_007062049	XM_007061987
Anguilla	*anguilla**	Actinopterygii	Anguiliformes	Anguilidae	16	XP_035266901	XM_035411010
Xenopus	*laevis*	Amphibia			16	NP_001089643	NM_001096174
Alligator	*mississippiensis*	Reptilia			16	XP_019341284	XM_019485739

Table S2 Results of the branch-site test in AQP11b Gobiidae stem branch.

Aquaporin	Foreground branch	LRT	*P*-value[a]	*Q*-value[b]	ω[c]	Prop. 2a[d]	Prop. 2b[e]	Selected sites
11b	Gobiidae stem branch	17.813	1.218×10^{-5}	4.021×10^{-4}	41.228	0.110	0.025	7

[a]Uncorrected *P*-value of the LRT

[b]Multiple-test correction of the LRT *P*-value (False Discovery Rate)

[c]Omega (dN/dS) ratio of the foreground branch(es)

[d]Proportion of sites that are under positive selection (ω 2a > 1) on the foreground branch(es) and under purifying selection (ω < 1) on background branches

[e]Proportion of sites that are under positive selection (ω 2a > 1) on the foreground branch(es) and under neutral selection (ω = 1) on background branches

The tree represented in figure S1 is too large to be presented in this Ph. D. thesis document. Instead it is available for visualisation and downloading at: **https://doi.org/10.5281/zenodo.7448163**. The figure caption is listed below:

Figure S1. Maximum likelihood (IQ-TREE) cladogram of vertebrate aquaporins based on 441 aligned amino acid positions. Aquaporin classes are indicated with colour panels and class paralogues with letters (a, b, a1, a2) on the corresponding branches. Branches of amphibious fishes are highlighted in red. The names of the branches under adaptive selection are written near their terminal location in the tree.

DISCUSSION

DISCUSIÓN

Evolution and diversification of major intrinsic proteins in eukaryotes

The study of diversity and evolution of the aquaporin (MIP) superfamily began early after their discovery (Agre et al., 1993; Heymann & Engel, 1999; Pao et al., 1991; Park & Saier, 1996; Preston et al., 1992). Most of the studies regarding this superfamily have focused on humans—from a medical point of view—, animals and land plants (e.g. Agre & Kozono, 2003; Finn et al., 2014; Finn & Cerdà, 2011, 2015; Gotfryd et al., 2018; Madeira et al., 2014; Martínez-Redondo et al., 2023). However, both the expansion and rapid development of the fields of genomics and transcriptomics during the last 20 years have delivered a vast amount of data for underrepresented, non-model organisms as well. The advent of this new data—along with the development of bioinformatic protocols and pipelines, as well as software and algorithms for robust phylogenomic approaches—has yielded a great variety of publications regarding the diversity of MIPs in major groups such as prokaryotes or unicellular eukaryotes (e.g. Abascal et al., 2014; Khabudaev et al., 2014; Pettersson et al., 2005; Tesan et al., 2021; Verma et al., 2014). In this sense, the results obtained in Chapter II of this thesis update the catalogue of MIPs genes within eukaryotes with special emphasis in unicellular forms. These results highlight the great diversification of these proteins in protists and describe three clades that can be traced back to LECA (see Figs. 2 and 3 in Chapter II). This chapter is a reproduction of a manuscript submitted to *Proceedings of the Royal Society B*, and thus constitutes an unfinished study. In the following discussion paragraphs, I address some of the study limitations, and implement some of the changes that have been indicated by three anonymous referees thus far.

In general, our dataset (see Fig. 2 in Chapter II) recovered all described MIP clades of classical vertebrate aquaporins (AQP0, 1, 2, 4, 5, 6, 14, 15), aquaammoniaporins (AQP8, 16), aquaglyceroporins (AQP3, 7, 9, 10, 13) and unorthodox/superaquaporins (AQP11, AQP12) (Abascal et al., 2014; Finn et al., 2014). However, it lacks some diversity within these big clades; for instance, only one sequence—from platypus (*Ornithorhynchus anatinus*)—that could be considered an AQP13 orthologue is included (Finn et al., 2014; Yilmaz et al., 2020). Besides platypus, the AQP13 subfamily—considered an aquaglyceroporin—has been only identified in another monotreme (echidna), in a hagfish, and in a few amphibians (Finn et al., 2014; Yilmaz et al., 2020). Likewise, we could not identify any AQP16 orthologue in our dataset. These sequences were discovered in amphibians, turtles and crocodiles and they are considered a sister

clade of the AQP8 orthologues (Finn et al., 2014), although their ability for ammonia transport has not been confirmed yet. Both AQP13 and 16 classes could have emerged early during the evolution of the chordate lineage, but apparently, they were subsequently lost in most of the extant groups (Finn et al., 2014; Yilmaz et al., 2020). Previous studies (e.g. Abascal et al., 2014; Catalán-García et al., 2021; Finn et al., 2015; Jia et al., 2022; Jia & Liu, 2022; Martínez-Redondo et al., 2023; Stavang et al., 2015) situate the origin of vertebrate MIP clades at least in the common ancestor of all animals (Metazoa). Our dataset also recovers similar results for invertebrates: one clade of co-orthologues to vertebrate unorthodox/superaquaporins (see Figs. 2f and S7 in Chapter II), another clade of co-orthologues to aquaamoniaporins (see Figs. 2e and S6 in Chapter II), three clades of co-orthologues to classical aquaporins (see Figs. 2a and S2 in Chapter II), and four orthology groups to vertebrate aquaglyceroporins (see Figs. 2h and S9 in Chapter II). One of the invertebrate clades that appears as closely related to vertebrate aquaporins (named Metazoa I; see Fig. S2 in Chapter II) contains the recently described 'entomoglyceroporins' that have secondarily evolved glycerol selectivity in insects (Finn et al., 2015). Although in Chapter II we claim for the novelty of some of these clades, it is true that we may have overlooked some of the previously classified MIP clades in invertebrates (Catalán-García et al., 2021; Finn et al., 2015; Jia & Liu, 2022; Li et al., 2022; Martínez-Redondo et al., 2023). In any case, the lack of a more detailed catalogue of vertebrate MIP clades (at finer taxonomic scale) or the discovery of new co-orthologues for these in invertebrates do not limit or contradict our conclusions, which are more broadly focused on the diversification of MIPs in the major eukaryotic supergroups and the process of eukaryogenesis. In this context, we suggest that the diversity of MIPs found in metazoans—which are part of the Opisthokonta clade (Burki et al., 2020)—is sufficiently remarkable for the scope of the chapter.

As in vertebrates, our dataset allowed us to recover all the main land plant aquaporin (PIPs, TIPs, XIPs, HIPs, SIPs, NIPs) and glyceroporin (GIPs) clades (Abascal et al., 2014; Finn et al., 2014; Maurel et al., 2015; Soto et al., 2012). In a very recent study, Li et al. (2022) identified older origins for several plant MIP subfamilies than was previously accepted (Abascal et al., 2014; Soto et al., 2012). They were able to describe orthologues for PIPs, TIPs, and SIPs subfamilies in several Chlorophyta genomes suggesting that these subfamilies could be traced back to the MRCA of all green plants (Chloroplastida). This study also showed that NIPs are present within Rhodophyta genomes, tracing back their origin at least to the emergence of plants (Archaeplastida).

However, although our dataset indeed includes Chlorophyta and Rhodophyta sequences, we were not able to recover the complete diversity within these clades, except for SIPs (see Figs. 2f and S7 in Chapter II). Nevertheless, our results suggest that the duplication of NIPs into clades NIP-1 to 4 likely dates back to the MRCA of land plants and their closest algal relatives, the Zygnematophyceae (see Figs. 2g and S8 in Chapter II). These sequences likely evolved from a horizontal gene transfer event from Bacteria (Abascal et al., 2014; Danielson & Johanson, 2010; Pommerrenig et al., 2020; Zardoya et al., 2002). Early studies suggested that the evolution of NIPs in land plants could be related to the acquisition of glycerol transport given that this group lacks GIPs orthologues (Abascal et al., 2014; Danielson & Johanson, 2010; Zardoya et al., 2002). However, a recent study (Pommerrenig et al., 2020) has shown that NIPs evolved from bacterial AqpNs that are not able to transport glycerol but arsenic and boron. Therefore, these authors suggested that the evolution of NIPs as glycerol transporters could have occurred during with the expansion of NIPs in land plants. According to our results, if NIPs diverged early in the evolution of all plants maybe this function did so as well, and therefore it may not be exclusive of the NIPs of just land plants.

Beyond animal and plant MIPs, we identified all other described MIP subfamilies within unicellular eukaryotic lineages, such as: LIPs, green algal and *Phytophthora* MIPs, and AQPXs (Anderberg et al., 2011; Azad et al., 2021; Khabudaev et al., 2014; Tesan et al., 2021). For instance, LIPs were identified as part of a clade that included other ochrophyte lineages beyond diatoms but also in dinoflagellates and ciliates, and thus dating back at least to the MRCA of the Stramenopiles + Alveolata clade within SAR (see Figs. 2f and S7 in Chapter II), and much older than initially thought (Khabudaev et al., 2014). The majority of *Phytophthora* MIPs (clades PMIP-A to H) clustered together, and likely originated from ancient duplications within oomycetes (probably the MRCA of Peronosporales and Phytiales), and the entire clade probably dates back to the Stramenopiles + Alveolata ancestor within SAR, as suggested by conserved amino acid residues in oomycetes and dinoflagellates (see Figs. 2h and S9 in Chapter II). The *Phytophthora* PMIP-I clade, originally defined by a single sequence (XP_008909057 Azad et al., 2021), is corroborated by two species but remains restricted to *Phytophthora* (see Figs. 2i and S10 in Chapter II). AQPXs are recovered as two distantly-related clades, suggesting that an ancient duplication event originated them (see Figs. 2d–e and S5–6 in Chapter II), although this contradicts the results reported by Tesan et al. (2021). However, the low statistical support of basal branches and the fast evolutionary rates of AQPXs—

that make them more prone to long branch attraction artefacts—suggest that this hypothesis should be taken with caution. Besides, our dataset lacks prokaryotic AQPXs orthologues that could be key for inferring the phylogenetic position of this clade not only among Eukaryotes, but in the entire Tree of Life. In our study, we defined some clades that were not recovered within the original class assigned to them (Fig. 1b, Table S1). Most of these new clades correspond to unicellular eukaryotes, which have remained largely understudied with respect to MIPs. Quantitatively, the majority of such new MIP clades correspond to the TSAR supergroup (Strassert et al., 2019). For instance, there is one MIP clade that can be traced back to the MRCA of TSAR, and five other clades to the Stramenopiles + Alveolate clade ancestor. In addition, seven clades of Haptista—with representatives of either Prymnesiophyceae or Pavlovales—and one clade of Glaucophyta and three clades of Rhodophyta—in Archeplastida—were recovered. Interestingly, a new trypanosomid (Discoba) clade was found in the AQPXs, but also with low statistical support. Regarding the closest unicellular relatives of animals (Opisthokonta, Holozoa), three clades of choanoflagellates and one clade of ichthyosporeans were recovered. Finally, one sponge MIP clade clusters with a bacterial homologue and it might represent a HGT of an aquaglyceroporin gene in sponges (Kenny et al., 2020) dating back at least to the MRCA of Heteroscleromorpha (see Figs. 2i and S10 in Chapter II).

The above discussed results presented in Chapter II suggest that our efforts in sequence data retrieval have yielded a relatively comprehensive dataset for eukaryotic MIPs—although lacking some prokaryotic diversity (see below for further discussion). These results—along with those from the cited literature (e.g. Abascal et al., 2014; Anderberg et al., 2011; Azad et al., 2021; Finn et al., 2014; Khabudaev et al., 2014; Soto et al., 2012; Tesan et al., 2021)—point out that MIPs have diversified in all major eukaryotic supergroups. In fact, we found significant correlation between the number of available genomes and transcriptomes per eukaryotic supergroup and the number of MIPs homologues recovered (Pearson's $R=0.93$, $p=0.0021$, see Fig. S12 in Chapter II). Nevertheless, clear causality might not be so straightforward. First, because transcriptomic data do not represent well the full gene repertoire of a particular species. And second, because we have overlooked some isoforms of the same protein that might be overestimating the diversity within some groups—although many of these repeated proteins certainly belong to the prokaryotic lineage and may not affect our results. In any case, an overrepresentation of the number of orthologues, along with the fact that we are

working with relatively short-length protein sequences that diverged a long time ago, could mislead the phylogenetic reconstruction and could perhaps explain the overall low statistical support found in parts of our tree.

In spite of these limitations, we were able to identify some MIP paralogue clades that might date back to ancient gene duplications during the early evolution of eukaryotes. Our dataset recovered a strongly supported clade termed MDC1 (MIP deep clade 1) that groups several distantly related unicellular protists including slime molds (Amoebozoa, Eumycetozoa), blastocladiomycete and chytridiomycete fungi (Fungi), golden algae (Stramenopiles, Ochrophyta), and rhizarians like *Paulinella* (Rhizaria) (see Figs. 2f and S7 in Chapter II). However, looking at Froger's residues—typically used for differentiation between water and glycerol transporters (Abascal et al., 2014; Froger et al., 1998)—we cannot differentiate between AQP or GLP. These sequences present small uncharged residues in P2–P3 positions and aromatic ones at P4–5 that are typical of AQPs. On the other hand, they also display aromatic amino acids at P1, which is typical of GLPs (see Fig. 3a in Chapter II). A second clade of deep evolutionary origin—termed MDC2—is that formed by chlorophyte algal MIP-D (Archaeplastida) and the dinoflagellate (Alveolata) clade named SAR XXV (see Figs. 2c, 3b and S4 in Chapter II), which received strong statistical support and displays conserved key residues as well. This clade could represent the first deep orthology proposed for any of the enigmatic clades termed MIPA–E that are exclusively found in green algae (Archaeplastida, Chlorophyta; Anderberg et al., 2011). As in the case of MDC1, residue conservation suggests that MDC2 could be aquaporins (e.g., non-aromatic P1, small uncharged P2–P3, and aromatic P4), but they also have a non-aromatic P5 as in GLPs. Finally, there is a third deep clade that clusters XIPs—which in our analyses group plant XIPs with slime molds and other amoebas (Amoebozoa) and chlorarachnophyte algae (Rhizaria)—and one diatom (Stramenopiles) (see Figs. 2c, 3c and S4 in Chapter II). In fact, a deep evolutionary origin of XIPs has already been proposed based on phylogenetic clustering of plant XIPs with *Dictyostelium* (Amoebozoa) and fungi (Abascal et al., 2014; Danielson & Johanson, 2008; Gupta & Sankararamakrishnan, 2009). Previously identified fungal XIPs are recovered elsewhere in our analyses (see Figs 2j and S11 in Chapter II) likely due to long branch attraction artefacts, but the conservation of key residues and the synapomorphic amino acids identified by Abascal et al. (2014) are strong indicators of their deep orthology relationship.

The presence of MIP clades whose origin can be traced back to the deepest nodes in the eukaryotic tree of life is indicative of a dynamic evolutionary pattern of gene duplications, losses, and divergences in the MIP family already very early in eukaryotic evolution. In fact, some of the deep MIP clades might be traceable to LECA, suggesting a complex ancestral repertoire of MIP homologues for the first eukaryote. In contrast, in Chapter II we underestimate the repertory of prokaryotic MIPs, and despite the fact that we recovered a diverse and relatively complete dataset for eukaryotes the lack of prokaryotic orthologues limits the explanation for the inferred MIP catalogue in LECA and FECA. According to the literature, it has been proposed that prokaryotes encode five different MIP clades: AqpN, AqmM, AqpZ, GlpF and AqpX. GlpFs and AqpMs are able to transport glycerol and water, respectively, whereas AQPZs are mainly water channels and the function of AqpXs is still unknown (Finn et al., 2014; Fu, 2000; Jensen et al., 2001; Kozono et al., 2003; Tesan et al., 2021). As discussed above, Pommerrenig et al. (2020) showed that AqpNs are the closest relatives of plant NIPs and that they can transport arsenic and boron. The polyphyly of the term AQP proposed in these studies entails a far more complex evolution from FECA than the one described in this chapter (see Fig. 3d in Chapter II). Our dataset lacks prokaryotic AQPX orthologues but includes at least one sequence of all other prokaryotic clades. GlpFs are confidently clustered with other GLPs, but recovered in two big clades (see Figs. 2h, j and S9, 11 in Chapter II). This result is extremely interesting as it would suggest the polyphyly of the term GLP as well. However, due to the lack of statistical support in our phylogenetic tree, further analyses are needed to validate the result. Our dataset also includes a few AqpMs and one AqpN sequence (see Figs. S7 and S8 in Chapter II). Given that the root of MIPs is not clear, Finn et al. (2014) proposed the node that separates AqpMs from other MIPs as the putative tree root due to its central position in their prokaryotic MIPs tree. Likewise, based on this very same tree, they proposed that AqpZs could be the sister group of AqpNs. However, in our midpoint rooted tree we recovered AqpMs within a big clade with AQPZs and some unclassified metazoan clades (see Fig. S8 in Chapter II). According to Pommerrenig et al. (2020), AqpNs and plant NIPs are clustered together. However, our analysis places the AqpN sequence elsewhere in the phylogeny (see Fig. S7 in Chapter II). It is very likely that the lack of robust clades for AqpMs and AqpNs is due to the small number of representative sequences in our dataset, especially considering that the split among these prokaryotic clades should have occurred very long time ago. An updated dataset considering MIPs diversity within the prokaryotic lineage could help in

solving these ancient relationships, and hence allowing better understanding of the evolution of MIPs orthologues within eukaryotes as well.

In this context of higher diversity, the description of the three deep clades in Chapter II (MDC1,2 and XIP; see Fig. 2 and Fig. 3d in Chapter II) remains significant. Our results indicate a complex dynamic evolution of MIPs in the earliest eukaryotes, even though our interpretation of how the MIPs of LECA and FECA emerged is unclear (see Fig. 3d in Chapter II). However, all three deep MIP clades contain representatives of both Obazoa (e.g., Opisthokonta, Amoebozoa) and Diaphoretickes (e.g., SAR, Archaeplastida). According to the current understanding of the deep eukaryotic phylogeny (see Fig. 1b in Chapter II; Burki et al., 2020), such duplication could be dated back directly to LECA under the assumption of a unikont/bikont root (Derelle et al., 2015). Alternatively, it could date back to the second deepest node in the tree under the assumption of a neozoan/excavate root (He et al., 2014). According to the latest molecular clock dating estimations, these events occurred >2 billion years ago (Strassert et al., 2021). However, under the assumption of a polyphyletic AQP clade, this deep orthology in eukaryotes could be directly inherited from their prokaryotic ancestors. In fact, as we discussed above, we were not able to assign the paralogues to AQPs or GLPs based on Froger's residues. Further investigation considering the key motifs of AqpNs, AqpMs, and AqpX could be key for clarifying the origins and possible prokaryotic ancestry of these deep clades. Furthermore, the investigation of some other key motifs of MIP sequences, such as the ar/R filters, could provide some insights about their function as well. Besides from water, MIPs are known to transport small molecules including non-polar compounds like glycerol, urea, lactic acid, ROS, hydrogen peroxide, gasses (ammonia, carbon dioxide, nitric oxide), and metalloids (boron, silicon, arsenic, antimony) (Finn & Cerdà, 2015; Gupta & Sankararamakrishnan, 2009; Maurel, 2007; Mukhopadhyay et al., 2014; Pommerrenig et al., 2020). In the case of unicellular eukaryotes, MIPs have been suggested as central water transporters for dealing with low temperatures and even freezing (Tanghe et al., 2006). Besides, from a biotechnological viewpoint, MIPs have been proposed as drug targets against fungal (Verma et al., 2014) and trypanosomatid (Tesan et al., 2021) parasites. On the other hand, the diversity found at the very early evolution of eukaryotes could be explained by the initial genomic duplications of the eukaryogenesis process (Eme et al., 2017; Makarova et al., 2005; Vosseberg et al., 2021). Many physiological processes have been inferred to be present in LECA assuming a common origin for all living eukaryotes (reviewed in Eme et al.,

2017). Regarding MIPs, ROS transport, for instance, has been proposed to be likely ubiquitous across all eukaryotic MIPs (Bienert & Chaumont, 2014). Better understanding of the function of these proteins could be probably obtained with a more robust phylogenetic framework of eukaryotic MIPs in the context of the entire Tree of Life. As in the results provided by Pommerrenig et al. (2020), studying the early evolution of MIPs could enlighten the timing of emergence of some functions—such as the acquisition of glycerol transport in NIPs—as well as their association with key evolutionary events (as it appears to have happened with the evolution of land plants).

Our next steps are focused on expanding the analysed dataset to include the complete repertoire of prokaryotic MIP clades. This should provide a better picture of the evolution of eukaryotic MIPs and potential cases of HGT. Besides that, one of the main limitations of the results presented in Chapter II is the relatively low statistical support in many parts of the inferred phylogenetic tree, especially for the deep relationships among MIP paralogues. We tried to address this problem by dividing the huge dataset in 10 smaller subsets (see Fig. S1 in Chapter II). However, these subsets may not reflect the actual evolutionary process appropriately, and the comparison among subsets can be challenging. On the other hand, we are confident about the support for the clades and internal relationships recovered within each of these subsets. The expansion of the dataset, along with several new rounds of decontamination could help to improve the phylogenetic resolution in some parts of the tree—hopefully for some of the deeper relationships—and to reduce certain biases in the number of MIP orthologues within each eukaryotic supergroup. Besides, some other phylogenetic approaches such as the use of mixed models of evolution or Bayesian inference methods could lead to a more robust evolutionary hypothesis for eukaryotic MIPs. Likewise, a better phylogenetic resolution could elucidate the origin of some of the new clades discovered in unicellular eukaryotes such as TSARs, rhodophytes and fungi, and could further enable the investigation of possible cases of endosymbiotic gene transfer (EGT).

Molecular evolution of aquaporins in amphibious fishes

Amphibious behaviour in fishes has long drawn the interest of scientists. Undoubtedly, the observation of a fish out of water is an eye-catching event and there are reports as ancient as the emergence of natural sciences (reviewed in Graham, 1997). More recently, several studies have focused on the highly diversified amphibious behaviour

exhibited by fishes (e.g. Ord & Cooke, 2016; Turko & Wright, 2015; Wright & Turko, 2016). These results suggest that amphibiousness is more common than intuitively thought. The number of putative amphibious fishes is certainly high, especially within the actinopterygian lineage. However, it is worth mentioning that it is within the sarcopterygian lineage where we find the truly, fully terrestrial clade of vertebrates, the tetrapods, as well as another major clade of extant amphibious fishes, the lungfishes (Laurin, 2010). Most of the studies on amphibious fishes have tried to understand how these animals manage to breathe and locomote on terrestrial environments, and how these processes can affect their metabolism and physiology, particularly excretion or ionoregulation (e.g. Bridges, 2015; J. B. Graham, 1997; LeBlanc et al., 2010; Pace & Gibb, 2014; Randall et al., 2015; Sayer, 2005; Turko et al., 2014). However, there is less knowledge about water maintenance and osmoregulation. There are a few studies that have researched osmoregulation in emerged amphibious fishes, and have provided evidence that some species are capable of producing mucus to avoid desiccation (LeBlanc et al., 2010; Jie Zhang et al., 2000, 2003). The results proposed by Finn et al. (2014)—which suggests that the AQP2, 5 and 6 subfamilies were likely involved in the terrestrialisation of tetrapods—provided a new path for further research and pointing to aquaporins (major intrinsic proteins) and their role in the adaptation of fishes to terrestrial environments. In the following pages, I am discussing the results obtained in Chapters III and IV, which focused on the catalogue, characterisation, and molecular evolution of aquaporin orthologues in 22 amphibious fishes. As mentioned in the Introduction of this Ph.D. thesis, for the discussion of these chapters, I am going to use the term AQP as a synonym of the entire superfamily of aquaporins or major intrinsic proteins, and the term aquaglyceroporin as a synonym for GLPs, as it is the standard use in the literature when studying these proteins in animals (see e.g. Finn & Cerda, 2015; Holm et al., 2005; Ikeda et al., 2011; Saparov et al., 2007; Watanabe et al., 2016; Yakata et al., 2011).

In brief, our results suggest that, in contrast to sarcopterygians, there is no evidence of the emergence of new paralogues that can be associated with the water-to-land transition in the studied species of amphibious fishes (all actinopterygian). Instead, we found that some of the orthologues that are also present in their fully aquatic relatives appear to be under adaptive/positive selection. The dataset from Chapter IV is an expanded version of the one presented in Chapter III. These two datasets differ only in the number of genomes of amphibious fishes and close relatives and outgroups that were searched and used to construct them. Besides, there is a small difference in the number of

tetrapod genomes employed because some were removed during the construction of the dataset of Chapter IV to avoid overrepresentation. Consequently, hereafter we are going to discuss the results presented in Chapter IV as they are the most inclusive. This dataset was constructed based on the genomic data of 22 putative amphibious fishes (reviewed in Graham, 1997; Ord & Cooke, 2016; Wright & Turko, 2016; see Fig. 1 in Chapter IV). We also included comparable data for 42 additional vertebrates with special emphasis in fishes plus four sequences of lungfishes (of classes AQP0 and AQP2-like; see Table S1 in Chapter IV).

Our phylogenetic tree recovered the main AQP—or MIP—clades previously described in vertebrates: (1) the aquaglyceroporins; (2) the water-selective classical aquaporins; (3) the unorthodox or superaquaporins; and (4) the AQP8-type aquaammoniaporins (see Figs. 2 and S1 in Chapter IV) (Abascal et al., 2014; Finn et al., 2014; Soto et al., 2012; Zardoya, 2005). In agreement with previous studies, our phylogenetic tree recovered a clade with AQP11 and 12 orthologues. However, the clade of aquaglyceroporins comprises only classes 3, 7, 9 and 10 in our tree (see Figs. 2 and S1 in Chapter IV). In contrast, Finn et al. (2014) and Yilmaz et al. (2020) reported that a AQP13 was also an aquaglyceroporin. An orthologue of this subfamily was included in the dataset of Chapter III (see Fig. S1 in Chapter III) and in the first steps of the construction of the dataset of Chapter IV as well. However, given that this subfamily has not been reported in any fish genome thus far (Finn et al., 2014; Yilmaz et al., 2020)—and we do not report any hit in any of the studied fish genomes either (see Table S1 in Chapter IV)—we decided to remove it from our dataset for the sake of phylogenetic resolution in this part of the tree, which is generally low (see Fig. S1 in Chapter IV). Likewise, two agnathan aquaglyceroporins (AQP3 of *Lethenteron camtschaticum* and *Petromyzon marinus*) were excluded as well because their recovered position in the tree was likely artefactual (see Fig. S1 in Chapter III). On the other hand, we found better phylogenetic resolution within the water-selective classical aquaporins (see Fig. S1 in Chapter IV). The relationships among these subfamilies were consistent with previous phylogenetic studies Abascal et al., 2014; Chauvigné et al., 2019; Finn et al., 2014). AQP14 is the first to split from the tree, followed by AQP4, and the clade comprising AQP15—whose phylogenetic position could not be confidently resolved by Finn et al. (2014)—, AQP1, and AQP0 as the sister-group of the tetrapod-exclusive aquaporins (AQP2-like first, then AQP5 and AQP6). Their phylogenetic position suggests that an ancestral duplication of AQP0 led to the emergence of these three new classes unique to

tetrapods (Abascal et al., 2014; Finn et al., 2014; Laforenza et al., 2016), with some coelacanth and lungfish AQP2-like paralogues the precursors of them (Finn et al., 2014). Regarding aquaamoniaporins, our sequence similarity searches returned an unexpected hit for an AQP16 orthologue within the genome of the European eel, although this subfamily has been only described in a few amphibians and reptiles (Finn et al., 2014). Our dataset only contained another two AQP16 sequences (see Fig. S1 in Chapter IV) and this small number can complicate/limit the analysis. Therefore, we decided to discard this putative AQP16 sequence of *A. anguilla* for the adaptive selection analyses.

Regarding the sequences of amphibious fishes, we recovered a total of 356 putative aquaporin genes (see Fig. 1 in Chapter IV). In the case of mudskipper species, we were able to add 9 new aquaporin sequences to the dataset recovered in Chapter III (see Fig. 1 in Chapter IV). We retrieved matches across 13 different aquaporin classes—including the aforementioned putative *A. anguilla* AQP16. In general, we found one paralogue of each of the classes that had been previously described in the literature, although we spotted some possible examples of gene loss as well. For instance, we only reported a few AQP15 orthologues and several species lack this paralogue (see Figs. 1, 2 and S1 in Chapter IV). This gene loss was previously described by Finn et al. (2014) who suggested that this could be due to a genome reduction (Wolf & Koonin, 2013). The physiological impact of this loss is not well understood, but several studies have shown examples of overlapping functions among different aquaporin classes (reviewed in Cerdà & Finn, 2010; Finn & Cerda, 2015). Likewise, the patterns of expression of aquaporins are highly diverse and, even under similar conditions, there are reports of different expression patterns among species and organs (see e.g. Cerdà & Finn, 2010; Madsen et al., 2015). These studies suggest a complex evolution pattern that may not be disentangled simply using presence-absence approaches. A very recent study has found patterns of gene loss and duplications in arthropod aquaporins that can be related to the acclimation to non-marine environments and ultimately with the water-to-land transition in this invertebrate group (Martínez-Redondo et al., 2023). In agreement with previous studies (Finn et al., 2014; Finn & Cerdà, 2011), we did not find any match for putative AQP2, AQP5 and AQP6 orthologues in any of the actinopterygian fish genomes used in this research (see Figs. 1, 2 and S1 in Chapter IV). Likewise, we were not able to catalogue any novel aquaporin class that could have appeared in amphibious fishes related to their out-of-water transition. Nevertheless, we found a high diversity in the number of certain aquaporin paralogues in the genomes of the actinopterygian fishes analysed (see Figs. 2

and S1 in Chapter IV). Most of these duplications are due to a whole genome duplication (WGD) event that occurred on the teleost stem branch. However, there are reports of some tandem and inter-chromosomal duplications in fishes as well (Cerdà and Finn 2010; Crow et al. 2006; Finn et al. 2014; Finn and Kristoffersen 2007; Cerdà & Finn, 2010; Crow et al., 2006; R. N. Finn et al., 2014; R. N. Finn & Kristoffersen, 2007; J. S. Taylor et al., 2001; Tingaud-Sequeira et al., 2010; Yilmaz et al., 2020). For instance, there is evidence of a tandem duplication of an AQP10 sequence that predates the teleost stem branch since *E. calabaricus*, part of the Polypteriformes, encodes two AQP10 paralogues (Yilmaz et al., 2020).

Our results indicate that, unlike in the sarcopterygian lineage, the development of an amphibious behaviour in the studied actinopterygian fishes is not related to the emergence of new aquaporin classes nor with the tetrapod-exclusive aquaporins. Finn et al. (2014) proposed that this clade of tetrapod-exclusive aquaporins could have arisen before the split between Actinopterygii and Sarcopterygii, thus being present in both lineages but subsequently lost early in the actinopterygians branch. Our results do not provide any new insight to either support or reject this hypothesis. In contrast, our results do suggest that, if aquaporins were relevant for the adaptation to an amphibious lifestyle in some actinopterygian fishes, this could have been achieved by changes at the molecular sequence level—which could entail conformational changes affecting function— in aquaporins subfamilies that are also present in their fully-aquatic relatives. In order to explore this possibility, we conducted adaptive selection analyses on the amphibious fish aquaporin catalogue (except for the *A. anguilla* AQP16, see above for discussion).

Adaptive or positive selection relies on the ratio between non-synonymous and synonymous nucleotide substitutions (dN/dS), usually known as the omega value (ω) (Yang, 2008). When adaptive selection occurs, the fixation of advantageous mutations can potentially lead to the development of evolutionary innovations due to the emergence of novel molecular functions (Nielsen & Yang, 1998). In this regard, two recent studies have reported signatures of adaptive selection in aquaporins, one focused on the cetacean land-to-water transition, and the other one focused on the terrestrial adaptation of reptiles (São Pedro et al., 2015; Zang et al., 2019). Our results on the dataset of Chapter IV suggest that adaptive selection may have occurred in 19 different branches of the studied amphibious fishes, spanning seven different orders (see Table 1 in Chapter IV). However, specific positions under adaptive selection could only be identified in 13 out of these 19 branches (see Fig. 3 in Chapter IV). This discrepancy could indicate that in some branches

the signal of adaptive selection is cumulative and not strong enough at any particular site. Moreover, the branch-site test used for detecting sequence positions that could have undergone adaptive selection is highly conservative and sometimes may lack enough statistical power (see Methods in Chapter IV and Yang et al., 2005; Zhang et al., 2005). As an example of this, the signal found on the *Anableps anableps* AQP15 branch (see Table 1 in Chapter IV) suggest a significant event of adaptive selection. However, its omega value indicates neutral selection (ω=1) and no position was highlighted under adaptive selection (see Table 1 and Fig. 3 in Chapter IV). This could certainly be due to the small number of AQP15 orthologues in our dataset, and it has been reported that the employed tests do not perform well with small-size alignments (Zhang et al., 2005).

Among the positively selected sites detected in our study (see Figs.3 and 4 in Chapter IV), some are found on or around the key motifs of AQPs. The NPA motifs are responsible for pore formation, and they are directly related with the aquaporin function (Ikeda et al., 2011; Ishibashi, 2006; Morishita et al., 2005; Takahashi et al., 2014). These residues are in the cytoplasmatic half of the aquaporin protein along with many of the conserved amino acids across the entire aquaporin or MIP superfamily (Abascal et al., 2014). In Chapter IV, we were able to detect adaptive selection at one NPA motif of the Gobiidae AQP11b stem branch, which suggests a conformational change of the pore that may have permitted the transport of a different molecule in this aquaporin class. This particular change was initially ascribed to the stem branch of mudskippers in Chapter III (and related to the terrestrialisation of this group), but the denser representation of gobies in Chapter IV allowed a more precise allocation of the branch with signals of adaptive selection (see *Evolution of aquaporins in the Gobiidae clade* below). Furthermore, and looking at the 3D molecular structure of AQP11 (see Fig. 4C in Chapter IV), we can see that there are two more putative positively selected residues in the vicinity of the modified NPA. These three regions (including the NPA itself) are relatively close to each other suggesting a possible modification of the pore constriction. Also, on the Siamese fighting fish (*Betta splendens)* AQP11b branch there are two positively selected positions located around this region that could be related with pore formation as well. AQP11 orthologues seem to be promising candidates for being involved in the amphibious development of the studied fishes with up to 15 putatively selected positions (one shared among three lineages; see below) corresponding to six different AQP11 branches. Nevertheless, this aquaporin class is still relatively unknown, and its function and expression patterns remain poorly understood. It is clustered with the AQP12 class within the

superaquaporins or unorthodox clade (see Figs.2 and S1 in Chapter IV and Abascal et al., 2014; Finn et al., 2014). Some studies have reported that it is fully capable of water transport (Yakata et al., 2011) and that it can be related with seawater acclimation (e.g. Kim et al., 2014; Ma et al., 2020). Furthermore, there are also reports of transport of H_2O_2 through it, and it has been associated with the reduction of cellular stress in the endoplasmic reticulum as well (Bestetti et al., 2020; Ishibashi & Tanaka, 2021; Yakata et al., 2011). However, whether AQP11 can deal with the rise of ROS, and therefore oxidative stress, when fishes adapt to land and air-breathing conditions (Pelster & Wood, 2018) remains unknown. A more accurate resolution of the 3D structure of AQP11, as well as experimental functional evidence, is needed to better understand the actual implications of the adaptive changes detected, especially considering that superaquaporins constitute one of the most highly diverse clades within the aquaporin superfamily (Abascal et al., 2014; Ishibashi et al., 2014).

Interestingly, another positively-selected residue on an AQP11 branch (*B. splendens* AQP11b) is one site away from the first ar/R selectivity filter (see Fig. 3 in Chapter IV). These four filters induce a constriction that defines the permeant transported (Fu, 2000; Sui et al., 2001), and the selection found could be related with changes in solute specificity. Abascal et al. (2014) found that the sites related with solute filtering are more prone to be found in the external half of the aquaporin protein. This includes the ar/R selectivity filters and the differentially conserved amino acids in aquaglyceroporins (P1-P5, Fig. 3) ((Froger et al., 1998; Fu, 2000; Sánchez et al., 2011; Sui et al., 2001). Around the first ar/R filter residue, we found some positively-selected sites on the *M. albus* AQP8a1 branch as well, although these sites are mainly in the internal half of the molecule (see Fig. 4 in Chapter IV) suggesting a lesser impact on solute specificity. It is on the cytoplasmic half of the aquaporin molecule where most of the regulation processes occur (reviewed in Törnroth-Horsefield et al., 2010). Aquaporin function can be regulated by gating and trafficking, especially on the cytoplasmatic c-terminal region (reviewed in Kreida & Törnroth-Horsefield, 2015) in which we found some positively selected sites on the AQP3 and AQP12 branches (see Fig. 3 in Chapter IV). Finally, it is worth mentioning that there are two sites that may have undergone adaptive selection in more than one branch in the same paralogue—*M. albus* and *Parablennius parvicornis* AQP8a1 branches, and *B. splendens*, *Kryptolebias marmoratus* and the Gobiidae clade AQP11b branches—, and across two different paralogues in two different clades—*Clarias batrachus* AQP3b and the Gobiidae AQP11b stem branches, and *M. albus* AQP8a1 and

the Gobiidae AQP11b stem branch (see Figs. 3 and 4 in Chapter IV). These results are quite interesting because they might be suggesting a possible case of convergent evolution. In general, and although little is known about how these positively selected sites could exactly affect solute specificity, it is possible that these single mutations can affect aquaporin solute recognition—and therefore change their function—as well as aquaporin regulation. Consequently, it is reasonably plausible that the detected sites under adaptive selection could have, somehow, facilitated the development of an amphibious lifestyle in these fishes, or at least perhaps some intermediate transitional steps that may involve osmoregulation adaptation (for example, from seawater to brackish and freshwater).

Evolution of aquaporins in the Gobiidae clade

The results in Chapter III (see Table 2 and Fig. 2 in Chapter III) showed positively selected amino acid residues in the AQP10a and AQP11b branches of a clade of four mudskipper species (Gobiidae: Oxudercinae): *Boleophthalmus pectinirostris*, *Periophthalmodon schlosseri*, *Periophthalmus magnuspinnatus*, and *Scartelaos histophorus*. In that Chapter, we hypothesise that this adaptive selection could suggest that these sites may have played a role in the terrestrialisation of these amphibious fishes. However, in Chapter IV we expanded our original taxon sampling to include two additional species of gobies (Gobiidae), *Neogobius melanostomus* and *Lesueurigobius sanzi* (Agorreta et al., 2013; Simonovic et al., 2001), as well as a representative of Apogonidae (*Sphaeramia orbicularis*) as their sister group (see Fig. S1 in Chapter IV). The results in Chapter IV indicate that there is no significant adaptive selection on the mudskippers AQP10a stem branch (the signal is lost or, at least, become too weak with the expanded taxon sampling), but we do find signal of adaptive selection on the mudskipper AQP11b, and on the AQP8a and AQP7 as well. In these three cases, however, the tests conducted do not highlight specific sites as positively selected.

Taken together, these results are slightly contradictory, especially taking into account that the most promising positively-selected position discussed in Chapter III was located at the first NPA motif of the AQP11b mudskippers clade stem branch: the canonical proline (N) was replaced by a serine (S), generating a SFI motif instead (see Fig.2 in Chapter III). The more comprehensive results of Chapter IV suggest that adaptive selection has likely occurred, but not on the SFI motif, or at least, not with enough

magnitude when comparing the mudskippers with other gobies. As we discussed above, the test for the identification of specific sites under adaptive selection —Bayes Empirical Bayes (BEB; see Methods in Chapter IV)— is fairly conservative and sometimes lacks statistical power (Yang et al., 2005; Zhang et al., 2005). We explored the results offered by a deprecated test, the Naïve Empirical Bayes (NEB), which is more liberal and somewhat less reliable (Nielsen & Yang, 1998; Yang et al., 2005). This test showed a few positions that could have experimented adaptive selection, even though, none of them were located at the SFI motif. Given this situation, we tested adaptive selection for the AQP11b stem branch of all the Gobiidae family (as represented in our dataset by the four mudskippers plus *Neogobius* and *Lesueurigobius*) in order to more precisely trace the selection over the SFI motif. The results indicate that it is the Gobiidae AQP11b stem branch the one that has experienced adaptive selection indeed, and the BEB test highlights that the serine located in the SFI motif is positively selected, along with six other sequence positions (see Fig. 3 in Chapter IV). This suggests that the modification of the NPA motif occurred before the origin of the mudskippers clade, and constitutes a synapomorphy of the entire Gobiidae (or even Gobioidei) that is not present in the sister group Apogonidae. Even though further research is still needed, this could be a case of exaptation (Gould & Vrba, 1982). A recent study has suggested that mudskippers underwent two convergent events of terrestrialisation and maybe the conditions (i.e. some molecular and genetic basis) for the development of an amphibious lifestyle evolved before the conquest of mudflats by this group (Steppan et al., 2022). In a similar way, in Chapter III we also indicated that the entire clade of sarcopterygian aquaporins—lungfish and coelacanth AQP2-like and tetrapods AQP2, 5 and 6 (see Fig. S1 in Chapter III and Fig. S1 in Chapter IV)—experienced adaptive selection as well (see Table 2 in Chapter III). This result points out that the adaptive selection could predate the split between the Actinistia and the Dipnoi + Tetrapoda clade and suggest that it may have occurred even before the development of the amphibious behaviour in the sarcopterygian lineage. Besides, if the hypothesis postulated by Finn et al. (2014) is assumed correct, and the first AQP2-like appeared even before the split between actinopterygian and sarcopterygian fishes, these results could be a sign of the evolutionary process that ended up with the conservation and division of this aquaporin clade within the sarcopterygian lineage.

In Chapter IV, we also tested adaptive selection for the AQP10a stem branch of all the Gobiidae family, but, unlike AQP11b, our results did not indicate any signature of adaptive selection on this branch. Therefore, the result in Chapter III appears to be

spurious or artefactual, likely influenced by the lower density of the taxon sampling. In the dataset of Chapter III, apart from the mudskippers themselves, only two sequences of fish AQP10a could be included (and both of the same family Cyprinidae), and it has been reported that adaptive selection tests do not perform well in situations with such small-size alignments (Zhang et al., 2005). The same explanation could be applied to the AQP7 class (with adaptive selection detected in Chapter IV, but not Chapter III) because the sampling is much denser in the latter study. In the case of the AQP8 class, adaptive selection could not be detected in Chapter III (but it was in Chapter IV) because the AQP8a orthologues were not found in this chapter (see Table 1 in Chapter III and Fig. 1 in Chapter IV). This emphasises the importance of having a solid protocol for sequence identification. For the construction of the dataset in Chapter IV, we took advantage of the workflow and pipeline described in Chapter I. The advent of a huge amount of genomic data—thanks to the appearance of high-throughput sequencing technologies—have prompted the development of robust bioinformatic pipelines for sequence retrieval and alignment that can be subsequently used in studies focused on phylogenetic diversity and molecular evolution. In our case, the application of the protocol described in Chapter I allowed us to construct the comprehensive dataset of Chapter IV, being able to extract, for example nine additional aquaporin sequences from mudskippers that had been overlooked in Chapter III (see Fig. 1 in Chapter IV).

Future directions and perspectives

The bioinformatic workflow of Chapter I and the results described in Chapter III are both already published in international peer-reviewed journals. Chapter II constitutes a manuscript that is currently under revision in an international peer-reviewed journal as well. In contrast, Chapter IV constitutes a manuscript that is still in preparation, and thus some aspects and analyses are still ongoing, and not yet reflected in the manuscript. For instance, we are currently investigating the evolutionary history of aquaporins genes in our phylogenetic tree using the software Bayesian Estimation for Gene Family Evolution (BEGFE) v. 2.0 (Liu et al., 2011). This method estimates the probability of gene family numbers expanding, contracting, or remaining constant on each node of the species tree. We want to test if aquaporins in vertebrates showed a higher or lower birth/death rate (lambda, λ) than other gene families. Besides, although they are out of the scope of this Ph.D. thesis, this project opens up some directions for further analysis. Perhaps, one of

the most straightforward next steps would be to conduct a protein modelling analysis to enable a more accurate reconstruction of the 3D structure of the different aquaporins that appear to be under adaptive selection. In particular, it would be interesting to model how the point mutations, especially those observed at the first NPA motif of the Gobiidae clade, can affect the protein structure. Also, it would be interesting to perform docking analysis as well in order to predict the function of mutated aquaporins, as this could shed some light in the context of amphibiousness in the studied fishes. Obviously, proper in vitro/vivo functional analyses, experimentally expressing these proteins into the oocytes of a vertebrate model species (*Danio*, *Xenopus*) will be even better for depicting their function. Another interesting approach could be the investigation of differences in expressions in aquaporins using transcriptomic analysis, collecting samples from different tissues that could be related, for instance, with the osmoregulatory process in amphibious fishes, such as the gut, the gills, and the kidney, under both aquatic and terrestrial conditions. Finally, more sophisticated approaches such as the use of CRISPR for gene edition could be used for investigating the effect of these mutated aquaporins in aquatic fishes and then study their amphibious behaviour on land.

CONCLUSIONS

CONCLUSIONES

Conclusions

From the studies presented in this Ph.D. thesis concerning the molecular evolution and role of aquaporins in the water-to-land transitions of amphibious fishes, the following conclusions can be drawn:

1. Aquaporins or major intrinsic proteins represent a highly diverse family; they are present in all living organisms from bacteria and archaea to land plants and mammals. The evolution of these proteins within Eukaryotes is still unresolved but MIPs homologues are virtually present in all eukaryotic supergroups. This diversity suggests an early burst in function that can be traced back at least to LECA.

2. The discovery of three new possible deep clades is in agreement with the idea of a highly diverse MIP repertory in LECA. However, further analysis—in progress—are needed to disentangle whether the origin of these clades occurred in the transition from FECA to LECA or if they evolved from prokaryotic MIPs, thus suggesting that they were already present in FECA.

3. In vertebrates, aquaporins appear to have facilitated the water-to-land transition of tetrapods. Sarcopterygians retain an AQP2-like that may have played a key role in the amphibious development of lungfishes and—after two rounds of duplication—the colonization of land by tetrapods. In contrast, actinopterygian amphibious fishes appear to have developed an amphibious lifestyle without any branch-specific duplication giving rise to novel AQP classes.

4. The study of adaptive evolution at the molecular level has revealed multiple sequence positions under positive selection in several aquaporins of the studied amphibious fishes. In this context, AQP11 orthologues arise as promising candidates for more exhaustive investigation. Besides, the adaptive selection found on the AQP11b stem branch of the Gobiidae clade points out that the modification of this sequence predates the emergence of mudskippers and suggest a possible case of exaptation.

5. Most aquaporin sequence positions found under adaptive/positive selection are in relatively close proximity to protein motifs involved in pore formation and substrate selectivity—as well as in regulation. Therefore, these results suggest that the adaptive changes detected (all or some) in aquaporins could have led to protein modifications in structure and/or function somehow related to the water-to-land transition of these amphibious fishes (especially considering the role of AQPs or MIPs in osmoregulation).

6. High-throughput sequencing technologies can yield an enormous amount of genome-scale data that often remains underexploited. In general, publicly available databases provide access to most of this data, but certainly state-of-the-art bioinformatic workflows and pipelines for gene mining and retrieval in a phylogenetic context are key for comprehensive molecular evolution analysis.

REFERENCES

BIBLIOGRAFÍA

Abascal, F., Irisarri, I., & Zardoya, R. (2014). Diversity and evolution of membrane intrinsic proteins. *Biochimica et Biophysica Acta (BBA) - General Subjects*, *1840*(5), 1468–1481.

Abascal, F., Zardoya, R., & Posada, D. (2005). ProtTest: selection of best-fit models of protein evolution. *Bioinformatics*, *21*(9), 2104–2105.

Abascal, F., Zardoya, R., & Telford, M. J. (2010). TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Research*, *38*(SUPPL. 2), 7–13.

Agorreta, A., San Mauro, D., Schliewen, U., van Tassell, J. L., Kovačić, M., Zardoya, R., & Rüber, L. (2013). Molecular phylogenetics of Gobioidei and phylogenetic placement of European gobies. *Molecular Phylogenetics and Evolution*, *69*(3), 619–633.

Agre, P., & Kozono, D. (2003). Aquaporin water channels: molecular mechanisms for human diseases. *FEBS Letters*, *555*(1), 72–78.

Agre, P., Preston, G. M., Smith, B. L., Jung, J. S., Raina, S., Moon, C., Guggino, W. B., & Nielsen, S. (1993). Aquaporin CHIP: the archetypal molecular water channel. *The American Journal of Physiology*, *265*, F463–F476.

Agre, P., Sasaki, S., & Chrispeels, M. J. (1993). Aquaporins: a family of water channel proteins. *American Journal of Physiology-Renal Physiology*, *265*(3), F461–F461.

Alioto, T., Blanco, E., Parra, G., & Guigó, R. (2007). Using geneid to identify genes. *Current Protocols in Bioinformatics*, *18*(1), 4.3.1-4.3.28.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410.

Anderberg, H. I., Danielson, J. Å. H., & Johanson, U. (2011). Algal MIPs, high diversity and conserved motifs. *BMC Evolutionary Biology*, *11*(1), 110.

Anisimova, M., & Yang, Z. (2007). Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. *Molecular Biology and Evolution*, *24*(5), 1219–1228.

Ashkenazy, H., Abadi, S., Martz, E., Chay, O., Mayrose, I., Pupko, T., & Ben-Tal, N. (2016). ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Research*, *44*(W1), W344–W350.

Azad, A. K., Ahmed, J., Hakim, A., Hasan, M., Alum, M., Hasan, M., Raihan, T., Ishikawa, T., & Sawa, Y. (2021). Genome-wide characterization deciphers distinct properties of aquaporins in six *Phytophthora* species. *Current Bioinformatics*, *16*(6), 880–898.

Bateman, A. (2019). UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Research*, *47*(D1), D506–D515.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2013). GenBank. *Nucleic Acids Research*, *41*(D1), D36–D42.

Bestetti, S., Galli, M., Sorrentino, I., Pinton, P., Rimessi, A., Sitia, R., & Medraño-Fernandez, I. (2020). Human aquaporin-11 guarantees efficient transport of H_2O_2 across the endoplasmic reticulum membrane. *Redox Biology*, *28*, 101326.

Bienert, G. P., Cavez, D., Besserer, A., Berny, M. C., Gilis, D., Rooman, M., & Chaumont, F. (2012). A conserved cysteine residue is involved in disulfide bond formation between plant plasma membrane aquaporin monomers. *Biochemical Journal*, *445*(1), 101–111.

Bienert, G. P., & Chaumont, F. (2014). Aquaporin-facilitated transmembrane diffusion of hydrogen peroxide. *Biochimica et Biophysica Acta (BBA) - General Subjects*, *1840*(5), 1596–1604.

Boone, M., & Deen, P. M. T. (2008). Physiology and pathophysiology of the vasopressin-regulated renal water reabsorption. *Pflügers Archiv - European Journal of Physiology*, *456*(6), 1005–1024.

Boumis, R. J., Ferry, L. A., Pace, C. M., & Gibb, A. C. (2014). heads or tails: do stranded fish (mosquitofish, *Gambusia affinis*) know where they are on a slope and how to return to the water? *PLOS ONE*, *9*(8), e104569-.

Breves, J. P. (2020). Chapter Twelve - Hormonal regulation of aquaporins in fishes. In G. B. T.-V. and H. Litwack (Ed.), *Aquaporin Regulation* (Vol. 112, pp. 265–287). Academic Press.

Bridges, C. R. (2015). Respiratory adaptations in intertidal fish. *American Zoologist*, *28*(1), 79–96.

Broekhuyse, R. M., Kuhlmann, E. D., & Stols, A. L. H. (1976). Lens membranes II. Isolation and characterization of the main intrinsic polypeptide (MIP) of bovine lens fiber membranes. *Experimental Eye Research*, *23*(3), 365–371.

Brůna, T., Lomsadze, A., & Borodovsky, M. (2020). GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genomics and Bioinformatics*, *2*(2), lqaa026.

Bruno, W. J., & Halpern, A. L. (1999). Topological bias and inconsistency of maximum likelihood using wrong models. *Molecular Biology and Evolution*, *16*(4), 564–566.

Burki, F., Roger, A. J., Brown, M. W., & Simpson, A. G. B. (2020). The new tree of Eukaryotes. *Trends in Ecology & Evolution*, *35*(1), 43–55.

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, *25*(15), 1972–1973.

Carrol, R. L. (2001). The origin and early adaptation of terrestrial vertebrates. *Journal of Paleontology*, *75*(6), 1202–1213.

Catalán-García, M., Chauvigné, F., Stavang, J. A., Nilsen, F., Cerdà, J., & Finn, R. N. (2021). Lineage-level divergence of copepod glycerol transporters and the emergence of isoform-specific trafficking regulation. *Communications Biology*, *4*(1), 643.

Cavalier-Smith, T. (2002). The phagotrophic origin of Eukaryotes and phylogenetic classification of Protozoa. *International Journal of Systematic and Evolutionary Microbiology*, *52*(2), 297–354.

Cerdà, J., & Finn, R. N. (2010). Piscine aquaporins: an overview of recent advances. *Journal of Experimental Zoology Part A: Ecological Genetics and Physiology*, *313 A*(10), 623–650.

Chauvigné, F., Yilmaz, O., Ferré, A., Fjelldal, P. G., Finn, R. N., & Cerdà, J. (2019). The vertebrate Aqp14 water channel is a neuropeptide-regulated polytransporter. *Communications Biology*.

Cheng, a, van Hoek, a N., Yeager, M., Verkman, a S., & Mitra, a K. (1997). Three-dimensional organization of a human water channel. *Nature*, *387*(6633), 627–630.

Chew, S. F., Gan, J., & Ip, Y. K. (2005). Nitrogen metabolism and excretion in the swamp eel, *Monopterus albus*, during 6 or 40 days of estivation in mud. *Physiological and Biochemical Zoology*, *78*(4), 620–629.

Chew, S. F., Hong, L. N., Wilson, J. M., Randall, D. J., & Ip, Y. K. (2003). Alkaline environmental pH has no effect on ammonia excretion in the mudskipper *Periophthalmodon schlosseri* but inhibits ammonia excretion in the related species *Boleophthalmus boddaerti*. *Physiological and Biochemical Zoology*, *76*(2), 204–214.

Chew, S. F., Sim, M. Y., Phua, Z. C., Wong, W. P., & Ip, Y. K. (2007). Active ammonia excretion in the giant mudskipper, *Periophthalmodon schlosseri* (Pallas), during emersion. *Journal of Experimental Zoology Part A: Ecological Genetics and Physiology*, *307*(6), 357–369.

Chew, S. F., Wilson, J. M., Ip, Y. K., & Randall, D. J. B. T.-F. P. (2005). Nitrogen excretion and defense against ammonia toxicity. In *The Physiology of Tropical Fishes* (Vol. 21, pp. 307–395). Academic Press.

Chrispeels, M. J., & Agre, P. (1994). Aquaporins: water channel proteins of plant and animal cells. *Trends in Biochemical Sciences*, *19*(10), 421–425.

Church, R. L., & Wang, J. (1992). Assignment of the lens intrinsic membrane protein MP19 structural gene to human chromosome 19. *Curr Eye Res*, *11*(5), 421–424.

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & De Hoon, M. J. L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423.

Connolly, D. L., Shanahan, C. M., & Weissberg, P. L. (1998). The aquaporins. A family of water channel proteins. *International Journal of Biochemistry and Cell Biology*, *30*(2), 169–172.

Consortium, T. U. (2008). The Universal Protein Resource (UniProt). *Nucleic Acids Research*, *36*(suppl_1), D190–D195.

Crow, K. D., Stadler, P. F., Lynch, V. J., Amemiya, C., & Wagner, G. P. (2006). The fish-specific Hox cluster duplication is coincident with the origin of teleosts. *Molecular Biology and Evolution*, *23*(1), 121–136.

Cunningham, C. W., Zhu, H., & Hillis, D. M. (1998). Best-fit maximum-likelihood models for phylogenetic inference: empirical tests with known phylogenies. *Evolution*, *52*(4), 978–987.

Dabruzzi, T. F., Wygoda, M. L., Wright, J. E., Eme, J., & Bennett, W. A. (2011). Direct evidence of cutaneous resistance to evaporative water loss in amphibious mudskipper (family Gobiidae) and rockskipper (family Blenniidae) fishes from Pulau Hoga, southeast Sulawesi, Indonesia. *Journal of Experimental Marine Biology and Ecology*, *406*(1–2), 125–129.

Daniels, C. B., Orgeig, S., Sullivan, L. C., Ling, N., Bennett, M. B., Schürch, S., Val, A. L., & Brauner, C. J. (2004). The origin and evolution of the surfactant system in fish: insights into the evolution of lungs and swim bladders. *Physiological and Biochemical Zoology*, *77*(5), 732–749.

Danielson, J. Å. H., & Johanson, U. (2008). Unexpected complexity of the aquaporin gene family in the moss *Physcomitrella patens*. *BMC Plant Biology*, *8*(1), 45.

Danielson, J. Å. H., & Johanson, U. (2010). Phylogeny of major intrinsic proteins. In T. P. Jahn & G. P. Bienert (Eds.), *MIPs and Their Role in the Exchange of Metalloids* (pp. 19–31). Springer New York.

Darriba, D., Taboada, G. L., Doallo, R., & Posada, D. (2011). ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics*, *27*(8), 1164–1165.

Darriba, D., Taboada, G. L., Doallo, R., & Posada, D. (2012). JModelTest 2: more models, new heuristics and parallel computing. *Nature Methods*, *9*(8), 1164–1165.

Davenport, J., & Matin, A. K. M. A. (1990). Terrestrial locomotion in the climbing perch, *Anabas testudineus* (Bloch) (Anabantidea, Pisces). *Journal of Fish Biology*, *37*(1), 175–184.

Deen, P. M. T., Verdijk, M. A. J., Knoers, N. V. A. M., Wieringa, B., Monnens, L. A. H., van Os, C. H., & van Oost, B. A. (1994). Requirement of human renal water channel aquaporin-2 for vasopressin-dependent concentration of urine. *Science*, *264*(5155), 92–95.

Dehal, P., & Boore, J. L. (2005). Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biology*, *3*(10), e314.

Derelle, R., Torruella, G., Klimeš, V., Brinkmann, H., Kim, E., Vlček, Č., Lang, B. F., & Eliáš, M. (2015). Bacterial proteins pinpoint a single eukaryotic root. *Proceedings of the National Academy of Sciences*, *112*(7), E693–E699.

Di Tommaso, P., Moretti, S., Xenarios, I., Orobitg, M., Montanyola, A., Chang, J. M., Taly, J. F., & Notredame, C. (2011). T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucleic Acids Research*, *39*(SUPPL. 2), W13–W17.

Ekblom, R., & Galindo, J. (2011). Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, *107*(1), 1–15.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E., & Finn, R. D. (2019). The Pfam protein families database in 2019. *Nucleic Acids Research*, *47*(D1), D427–D432.

Ellerby, D. J., Spierts, I. L. Y., & Altringham, J. D. (2001). Fast muscle function in the European eel (*Anguilla Anguilla* L.) during aquatic and terrestrial locomotion. *Journal of Experimental Biology*, *204*(13), 2231 LP – 2238.

Eme, L., Spang, A., Lombard, J., Stairs, C. W., & Ettema, T. J. G. (2017). Archaea and the origin of eukaryotes. *Nature Reviews Microbiology*, *15*(12), 711–723.

Escriva, H., Bertrand, S., Germain, P., Robinson-Rechavi, M., Umbhauer, M., Cartry, J., Duffraisse, M., Holland, L., Gronemeyer, H., & Laudet, V. (2006). Neofunctionalization in vertebrates: the example of retinoic acid receptors. *PLoS Genetics*, *2*(7), 0955–0965.

Evans, D. H. (2008). Teleost fish osmoregulation: what have we learned since August Krogh, Homer Smith, and Ancel Keys. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, *295*(2), R704–R713.

Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Biology*, *27*(4), 401–410.

Finkelstein, A. (1987). Water movement through lipid bilayers, pores, and plasma membranes. *Theory and Reality*, *10*.

Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, *39*(SUPPL. 2), W29–W37.

Finn, R. N., & Cerdà, J. (2011). Aquaporin evolution in fishes. *Frontiers in Physiology*, *JUL*, 44.

Finn, R. N., & Cerdà, J. (2015). Evolution and functional diversity of aquaporins. *Biological Bulletin*, *229*(1), 6–23.

Finn, R. N., Chauvigné, F., Hlidberg, J. B., Cutler, C. P., & Cerdà, J. (2014). The lineage-specific evolution of aquaporin gene clusters facilitated tetrapod terrestrial adaptation. *PLoS ONE*, *9*(11), e113686.

Finn, R. N., Chauvigné, F., Stavang, J. A., Belles, X., & Cerdà, J. (2015). Insect glycerol transporters evolved by functional co-option and gene replacement. *Nature Communications*, *6*(1), 7814.

Finn, R. N., & Kristoffersen, B. A. (2007). Vertebrate vitellogenin gene duplication in relation to the 3R hypothesis: correlation to the pelagic egg and the oceanic radiation of teleosts. *PLoS ONE*, *2*(1), e169.

Flammang, B. E., Suvarnaraksha, A., Markiewicz, J., & Soares, D. (2016). Tetrapod-like pelvic girdle in a walking cavefish. *Scientific Reports*, *6*(1), 23711.

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L., & Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, *151*(4), 1531–1545.

Freitas, R., Gómez-Marín, C., Wilson, J. M., Casares, F., & Gómez-Skarmeta, J. L. (2012). Hoxd13 contribution to the evolution of vertebrate appendages. *Developmental Cell*, *23*(6), 1219–1229.

Frick, N. T., & Wright, P. A. (2002a). Nitrogen metabolism and excretion in the mangrove killifish *Rivulus marmoratus* I. The influence of environmental salinity and external ammonia. *Journal of Experimental Biology*, *205*(1), 79 LP – 89.

Frick, N. T., & Wright, P. A. (2002b). Nitrogen metabolism and excretion in the mangrove killifish *Rivulus marmoratus* II. Significant ammonia volatilization in a teleost during air-exposure. *Journal of Experimental Biology*, *205*(1), 91 LP – 100.

Froger, A., Thomas, D., Delamarche, C., & Tallur, B. (1998). Prediction of functional residues in water channels and related proteins. *Protein Science*, *7*(6), 1458–1468.

Fu, D. (2000). Structure of a glycerol-conducting channel and the basis for its selectivity. *Science*, *290*(5491), 481–486.

Giffard-Mena, I., Boulo, V., Aujoulat, F., Fowden, H., Castille, R., Charmantier, G., & Cramb, G. (2007). Aquaporin molecular characterization in the sea-bass (*Dicentrarchus labrax*): The effect of salinity on AQP1 and AQP3 expression. *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology*, *148*(2), 430–444.

Gillis, G. B. (1998). Environmental effects on undulatory locomotion in the American eel *Anguilla rostrata*: kinematics in water and on land. *The Journal of Experimental Biology*, *201*(7), 949 LP – 961.

Goldman, N. (1998). Effects of sequence alignment procedures on estimates of phylogeny. *BioEssays*, *20*(4), 287–290.

Goldman, N., Thorne, J. L., & Jones, D. T. (1998). Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics*, *149*(1), 445–458.

Gorin, M. B., Yancey, S. B., Cline, J., Revel, J.-P., & Horwitz, J. (1984). The major intrinsic protein (MIP) of the bovine lens fiber membrane: characterization and structure based on cDNA cloning. *Cell*, *39*(1), 49–59.

Gotfryd, K., Mósca, A. F., Missel, J. W., Truelsen, S. F., Wang, K., Spulber, M., Krabbe, S., Hélix-Nielsen, C., Laforenza, U., Soveral, G., Pedersen, P. A., & Gourdon, P. (2018). Human adipose glycerol flux is regulated by a pH gate in AQP10. *Nature Communications*, *9*(1), 4749.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., Di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–652.

Grachev, M. A., Annenkov, V. V, & Likhoshway, Y. V. (2008). Silicon nanotechnologies of pigmented heterokonts. *BioEssays*, *30*(4), 328–337.

Graham, J. B. (1997). *Air-breathing fishes: evolution, diversity and adaptation.* Academic Press, San Diego.

Graham, J. B., & Lee, H. J. (2004). Breathing air in air: in what ways might extant amphibious fish biology relate to prevailing concepts about early tetrapods, the evolution of vertebrate air breathing, and the vertebrate land transition? *Physiological and Biochemical Zoology*, *77*(5), 720–731.

Graham, J., L. Lee, H., & Wegner, N. (2007). Transition from water to land in an extant group of fishes: air breathing and the acquisition sequence of adaptations for amphibious life in oxudercine gobies. In MN. Fernandes, FT. Rantin, ML. Glass, & B. Kapoor (Eds.), *Fish Respiration and Environment* (pp. 255–288). Science Publishers, Enfield.

Grosberg, R. K., Vermeij, G. J., & Wainwright, P. C. (2012). Biodiversity in water and on land. *Current Biology*, *22*(21), R900–R903.

Grosell, M. (2010). The role of the gastrointestinal tract in salt and water balance. In *Fish Physiology*.

Gould, S. J., & Vrba, E. S. (1982). Exaptation—a Missing Term in the Science of Form. *Paleobiology*, *8*(1), 4–15.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology*, *59*(3), 307–321.

Guindon, S., Lethiec, F., Duroux, P., & Gascuel, O. (2005). PHYML Online - A web server for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Research*, *33*(SUPPL. 2), W557–W559.

Gupta, A. B., & Sankararamakrishnan, R. (2009). Genome-wide analysis of major intrinsic proteins in the tree plant *Populus trichocarpa*: characterization of XIP subfamily of aquaporins from evolutionary perspective. *BMC Plant Biology*, *9*(1), 134.

Gustavsson, S., Lebrun, A.-S., Nordén, K., Chaumont, F., & Johanson, U. (2005). A novel plant major intrinsic protein in *Physcomitrella patens* most similar to bacterial glycerol channels. *Plant Physiology*, *139*(1), 287–295.

Han, M. V., Demuth, J. P., McGrath, C. L., Casola, C., & Hahn, M. W. (2009). Adaptive evolution of young gene duplicates in mammals. *Genome Research*, *19*(5), 859–867.

Hardison, R. C. (2003). Comparative genomics. *PLOS Biology*, *1*(2), e58.

Harris, V. A. (1960). On the locomotion of the mud-skipper *Periophthalmus koelreuteri* (Pallas): (gobiidae). *Proceedings of the Zoological Society of London*, *134*(1), 107–135.

He, D., Fiz-Palacios, O., Fu, C.-J., Fehling, J., Tsai, C.-C., & Baldauf, S. L. (2014). An alternative root for the eukaryote tree of life. *Current Biology*, *24*(4), 465–470.

Hedfalk, K., Törnroth-Horsefield, S., Nyblom, M., Johanson, U., Kjellbom, P., & Neutze, R. (2006). Aquaporin gating. *Current Opinion in Structural Biology*, *16*(4), 447–456.

Heymann, J. B., & Engel, A. (1999). Aquaporins: phylogeny, structure, and physiology of water channels. *News in Physiological Sciences*, *14*, 187–193.

Heymann, J. B., & Engel, A. (2000). Structural clues in the sequences of the aquaporins. *Journal of Molecular Biology*, *295*(4), 1039–1053.

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2018). UFBoot2: improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution*, *35*(2), 518–522.

Holm, L. M., Jahn, T. P., Møller, A. L. B., Schjoerring, J. K., Ferri, D., Klaerke, D. A., & Zeuthen, T. (2005). NH3 and NH4+ permeability in aquaporin-expressing *Xenopus oocytes*. *Pflügers Archiv*, *450*(6), 415–428.

Holm, L. M., Klaerke, D. A., & Zeuthen, T. (2004). Aquaporin 6 is permeable to glycerol and urea. *Pflügers Archiv*, *448*(2), 181–186.

Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Ridwan Amode, M., Armean, I. M., Azov, A. G., Bennett, R., Bhai, J., Billis, K., Boddu, S., Charkhchi, M., Cummins, C., da Rin Fioretto, L., Davidson, C., Dodiya, K., El Houdaigui, B., Fatima, R., … Flicek, P. (2021). Ensembl 2021. *Nucleic Acids Research*, *49*(D1), D884–D891.

Hub, J. S., & de Groot, B. L. (2008). Mechanism of selectivity in aquaporins and aquaglyceroporins. *Proceedings of the National Academy of Sciences*, *105*(4), 1198–1203.

Huelsenbeck, J. P., & Hillis, D. M. (1993). Success of phylogenetic methods in the four taxon case. *Systematic Biology*, *42*(3), 247–264.

Huelsenbeck, J. P., & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, *17*(8), 754–755.

Huerta-Cepas, J., Serra, F., & Bork, P. (2016). ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Molecular Biology and Evolution*, *33*(6), 1635–1638.

Hughes, L. C., Ortí, G., Huang, Y., Sun, Y., Baldwin, C. C., Thompson, A. W., Arcila, D., Betancur-R., R., Li, C., Becker, L., Bellora, N., Zhao, X., Li, X., Wang, M., Fang, C., Xie, B., Zhou, Z., Huang, H., Chen, S., … Shi, Q. (2018). Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. *Proceedings of the National Academy of Sciences*, *115*(24), 6249 LP – 6254.

Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., Finn, R. D., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Laugraud, A., Letunic, I., Lonsdale, D., … Yeats, C. (2009). InterPro: The integrative protein signature database. *Nucleic Acids Research*, *37*(SUPPL. 1), D211–D215.

Huson, D. H., & Scornavacca, C. (2012). Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Systematic Biology*, *61*(6), 1061–1067.

Ikeda, M., Andoo, A., Shimono, M., Takamatsu, N., Taki, A., Muta, K., Matsushita, W., Uechi, T., Matsuzaki, T., Kenmochi, N., Takata, K., Sasaki, S., Ito, K., & Ishibashi, K. (2011). The NPC motif of aquaporin-11, unlike the NPA motif of known aquaporins, is essential for full expression of molecular function. *Journal of Biological Chemistry*.

Innan, H. (2009). Population genetic models of duplicated genes. *Genetica*, *137*(1), 19–37.

Inoue, J., & Satoh, N. (2019). ORTHOSCOPE: an automatic web tool for phylogenetically inferring bilaterian orthogroups with user-selected taxa. *Molecular Biology and Evolution*, *36*(3), 621–631.

Ip, Y. K., Chew, S. F., & Randall, D. J. (2004). Five tropical air-breathing fishes, six different strategies to defend against ammonia toxicity on land. *Physiological and Biochemical Zoology*, *7*(5), 768–782.

Ip, Y. K., Chew, S. F., & Randall, D. J. B. T.-F. P. (2001). Ammonia toxicity, tolerance, and excretion. In *Nitrogen Excretion* (Vol. 20, pp. 109–148). Academic Press.

Ip, Y. K., Randall, D. J., Kok, T. K. T., Barzaghi, C., Wright, P. A., Ballantyne, J. S., Wilson, J. M., & Chew, S. F. (2004). The giant mudskipper *Periophthalmodon schlosse*ri facilitates active NH 4+ excretion by increasing acid excretion and decreasing NH3 permeability in the skin. *Journal of Experimental Biology*, *207*(5), 787–801.

Ip, Y. K., Soh, M. M. L., Chen, X. L., Ong, J. L. Y., Chng, Y. R., Ching, B., Wong, W. P., Lam, S. H., & Chew, S. F. (2013). Molecular characterization of branchial aquaporin 1aa and effects of seawater acclimation, emersion or ammonia exposure on its mrna expression in the gills, gut, kidney and skin of the freshwater climbing perch, *Anabas testudineus*. *PLoS One*, *8*(4), 1–14.

Ip, Y. K., Tay, A. S. L., Lee, K. H., & Chew, S. F. (2004). Strategies for surviving high concentrations of environmental ammonia in the swamp eel *Monopterus albus*. *Physiological and Biochemical Zoology*, *77*(3), 390–405.

Ishibashi, K. (2006). Aquaporin subfamily with unusual NPA boxes. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, *1758*(8), 989–993.

Ishibashi, K., Morishita, Y., & Tanaka, Y. (2017). The evolutionary aspects of aquaporin family. *Aquaporins Advances in Experimental Medicine and Biology*, *212*(14), 2183–2193.

Ishibashi, K., Tanaka, Y., & Morishita, Y. (2014). The role of mammalian superaquaporins inside the cell. *Biochimica et Biophysica Acta (BBA) - General Subjects*, *1840*(5), 1507–1512.

Ishibashi, K., Tanaka, Y., & Morishita, Y. (2021). The role of mammalian superaquaporins inside the cell: an update. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, *1863*(7), 183617.

Ishimatsu, A., & Gonzales, T. (2011). Mudskippers: front runners in the modern invasion of land. In R. A. Patzner, J. L. Van Tassell, M. Kovačić, & B. G. Kapoor (Eds.), *The Biology of Gobies* (pp. 609–638). CRC Press and Science Publishers.

Ishimatsu, A., Hishida, Y., Takita, T., Kanda, T., Oikawa, S., Takeda, T., & Huat, K. K. (1998). Mudskippers store air in their burrows. *Nature*, *391*(6664), 237–238.

Jaafar, Zeehan., & Murdy, E. O. (2017). *Fishes out of water: biology and ecology of mudskippers*. CRC Press, Boca Raton.

Jahn, T. P., Møller, A. L. B., Zeuthen, T., Holm, L. M., Klærke, D. A., Mohsin, B., Kühlbrandt, W., & Schjoerring, J. K. (2004). Aquaporin homologues in plants and mammals transport ammonia. *FEBS Letters*, *574*(1–3), 31–36.

Janssens, P. A. (1964). The metabolism of the aestivating African lungfish. *Comparative Biochemistry and Physiology*, *11*(1), 105–117.

Jensen, M. Ø., Tajkhorshid, E., & Schulten, K. (2001). The mechanism of glycerol conduction in aquaglyceroporins. *Structure*, *9*(11), 1083–1093.

Jia, Y., & Liu, X. (2022). Diversification of the aquaporin family in geographical isolated oyster species promote the adaptability to dynamic environments. *BMC Genomics*, *23*(1), 211.

Jia, Y., Xu, F., & Liu, X. (2022). Duplication and subsequent functional diversification of aquaporin family in Pacific abalone *Haliotis discus hannai*. *Molecular Phylogenetics and Evolution*, *168*, 107392.

Johanson, Z. (2011). How vertebrates left the water. *Acta Zoologica*, *92*(3), 10–12.

Jones, D. T., Taylor, W. R., & Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Bioinformatics*, *8*(3), 275–282.

Katoh, K., Misawa, K., Kuma, K. I., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066.

Katoh, K., Rozewicki, J., & Yamada, K. D. (2019). MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics*, *20*(4), 1160–1166.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., & Drummond, A. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, *28*(12), 1647–1649.

Kenny, N. J., Francis, W. R., Rivera-Vicéns, R. E., Juravel, K., de Mendoza, A., Díez-Vives, C., Lister, R., Bezares-Calderón, L. A., Grombacher, L., Roller, M., Barlow, L. D., Camilli, S., Ryan, J. F., Wörheide, G., Hill, A. L., Riesgo, A., & Leys, S. P. (2020). Tracing animal genomic evolution with the chromosomal-level assembly of the freshwater sponge *Ephydatia muelleri*. *Nature Communications*, *11*(1), 3676.

Khabudaev, K. V., Petrova, D. P., Grachev, M. A., & Likhoshway, Y. V. (2014). A new subfamily LIP of the major intrinsic proteins. *BMC Genomics*, *15*(1), 173.

Kim, Y. K., Lee, S. Y., Kim, B. S., Kim, D. S., & Nam, Y. K. (2014). Isolation and mRNA expression analysis of aquaporin isoforms in marine medaka *Oryzias dancena*, a euryhaline teleost. *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology*, *171*, 1–8.

Kok, W. K., Lim, C. B., Lam, T. J., & Ip, Y. K. (1998). The mudskipper *Periophthalmodon schlosseri* respires more efficiently on land than in water and vice versa for *Boleophthalmus boddaerti*. *Journal of Experimental Zoology*, *280*(1), 86–90.

Konno, N., Hyodo, S., Yamaguchi, Y., Kaiya, H., Miyazato, M., Matsuda, K., & Uchiyama, M. (2009). African lungfish, Protopterus annectens, possess an arginine vasotocin receptor homologous to the tetrapod V2-type receptor. *Journal of Experimental Biology*.

Konno, N., Hyodo, S., Yamaguchi, Y., Matsuda, K., & Uchiyama, M. (2010). Vasotocin/V2-type receptor/aquaporin axis exists in African lungfish kidney but is functional only in terrestrial condition. *Endocrinology*, *151*(3), 1089–1096.

Kornobis, E., Cabellos, L., Aguilar, F., Frías-López, C., Rozas, J., Marco, J., & Zardoya, R. (2015). TRUFA: a user-friendly web server for de novo RNA-seq analysis using cluster computing. *Evolutionary Bioinformatics*, *11*, EBO-S23873.

Koshi, J. M., & Goldstein, R. A. (1995). Context-dependent optimal substitution matrices. *Protein Engineering, Design and Selection*, *8*(7), 641–645.

Kozono, D., Ding, X., Iwasaki, I., Meng, X., Kamagata, Y., Agre, P., & Kitagawa, Y. (2003). Functional expression and characterization of an archaeal aquaporin. AqpM from *Methanothermobacter marburgensis*. *Journal of Biological Chemistry*, *278*(12), 10649–10656.

Kreida, S., & Törnroth-Horsefield, S. (2015). Structural insights into aquaporin selectivity and regulation. *Current Opinion in Structural Biology*, *33*, 126–134.

Kruse, E., Uehlein, N., & Kaldenhoff, R. (2006). The aquaporins. *Genome Biology*, *7*(2), 206.

Ku, C., Nelson-Sathi, S., Roettger, M., Sousa, F. L., Lockhart, P. J., Bryant, D., Hazkani-Covo, E., McInerney, J. O., Landan, G., & Martin, W. F. (2015). Endosymbiotic origin and differential loss of eukaryotic genes. *Nature*, *524*(7566), 427–432.

Laforenza, U., Bottino, C., & Gastaldi, G. (2016). Mammalian aquaglyceroporin function in metabolism. In *Biochimica et Biophysica Acta - Biomembranes* (Vol. 1858, Issue 1, pp. 1–11).

Lanfear, R., Calcott, B., Ho, S. Y. W., & Guindon, S. (2012). PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Molecular Biology and Evolution*, *29*(6), 1695–1701.

Larsen, E. H., Deaton, L. E., Onken, H., O'Donnell, M., Grosell, M., Dantzler, W. H., & Weihrauch, D. (2014). Osmoregulation and excretion. *Comprehensive Physiology*, *4*(2), 405–573.

Laurin, M. (2010). How vertebrates left the water. In *University of California Press, Berkeley, CA.*

Le, S. Q., Lartillot, N., & Gascuel, O. (2008). Phylogenetic mixture models for proteins. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1512), 3965–3976.

LeBlanc, D. M., Wood, C. M., Fudge, D. S., & Wright, P. A. (2010). A fish out of water: gill and skin remodeling promotes osmo- and ionoregulation in the mangrove killifish *Kryptolebias marmoratus*. *Physiological and Biochemical Zoology*, *83*(6), 932–949.

Lee, C. Y., Chiu, Y. C., Wang, L. B., Kuo, Y. L., Chuang, E. Y., Lai, L. C., & Tsai, M. H. (2013). Common applications of next-generation sequencing technologies in genomic research. *Translational Cancer Research*, *2*(1), 33–45.

Lee, Y.-J., Choi, Y., & Ryu, B.-S. (1995). A taxonomic revision of the genus *Periophthalmus* (Pisces: Gobiidae) from Korea with description of a new species. *Korean Journal of Ichthyology*, *7*(2), 120–127.

Lemoine, F., Domelevo Entfellner, J.-B., Wilkinson, E., Correia, D., Dávila Felipe, M., De Oliveira, T., & Gascuel, O. (2018). Renewing Felsenstein's phylogenetic bootstrap in the era of big data. *Nature*, *556*(7702), 452–456.

Li, G., Chen, T., Zhang, Z., Li, B., & Tian, S. (2020). Roles of aquaporins in plant-pathogen interaction. In *Plants* (Vol. 9, Issue 9).

Li, Q., Tong, T., Jiang, W., Cheng, J., Deng, F., Wu, X., Chen, Z. H., Ouyang, Y., & Zeng, F. (2022). Highly conserved evolution of aquaporin pips and tips confers their crucial contribution to flowering process in plants. *Frontiers in Plant Science*, *12*.

Lignot, J. H., Cutler, C. P., Hazon, N., & Cramb, G. (2002). Immunolocalisation of aquaporin 3 in the gill and the gastrointestinal tract of the European eel *Anguilla anguilla* (L.). *Journal of Experimental Biology*, *205*(17), 2653–2663.

Linnaeus, C. v. (1758). Systema Naturae, edition X, vol. 1 (Systema naturae per regna tria naturae, secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis. Tomus I. Editio decima, reformata). *Holmiae Salvii*, *824*.

Liu, L., Yu, L., Kalavacharla, V., & Liu, Z. (2011). A Bayesian model for gene family evolution. *BMC Bioinformatics*, *12*(1), 426.

Long, J. A., & Gordon, M. S. (2004). The greatest step in vertebrate history: a paleobiological review of the fish-tetrapod transition. *Physiological and Biochemical Zoology*, *77*(5), 700–719.

Lorente-Martínez, H., Agorreta, A., & San Mauro, D. (2022). Genomic fishing and data processing for molecular evolution research. *Methods and Protocols*, *5*(2), 26.

Lorente-Martínez, H., Agorreta, A., Torres-Sánchez, M., & San Mauro, D. (2018). Evidence of positive selection suggests possible role of aquaporins in the water-to-land transition of mudskippers. *Organisms Diversity & Evolution*, *18*(4), 499–514.

Löytynoja, A. (2014). Phylogeny-aware alignment with PRANK. In D. J. Russell (Ed.), *Multiple Sequence Alignment Methods* (1st ed., pp. 155–170). Humana Press.

Lutteropp, S., Kozlov, A. M., & Stamatakis, A. (2020). A fast and memory-efficient implementation of the transfer bootstrap. *Bioinformatics*, *36*(7), 2280–2281.

Lynch, M., & Force, A. (2000). The probability of duplicate gene preservation by subfunctionalization. *Genetics*, *154*(1), 459–473.

Ma, Q., Liu, X., Li, A., Liu, S., & Zhuang, Z. (2020). Effects of osmotic stress on the expression profiling of aquaporin genes in the roughskin sculpin (Trachidermus fasciatus). *Acta Oceanologica Sinica*, *39*(6), 19–25.

Macey, R. I. (1984). Transport of water and urea in red blood cells. *American Journal of Physiology-Cell Physiology*, *246*(3), C195–C203.

Maclver, B., Cutler, C. P., Yin, J., Hill, M. G., Zeidel, M. L., & Hill, W. G. (2009). Expression and functional characterization of four aquaporin water channels from the European eel (*Anguilla anguilla*). *Journal of Experimental Biology*, *212*(17), 2856–2863.

Madeira, A., Fernández-Veledo, S., Camps, M., Zorzano, A., Moura, T. F., Ceperuelo-Mallafré, V., Vendrell, J., & Soveral, G. (2014). Human aquaporin-11 is a water and glycerol channel and localizes in the vicinity of lipid droplets in human adipocytes. *Obesity*, *22*(9), 2010–2017.

Madsen, S. S., Bujak, J., & Tipsmark, C. K. (2014). Aquaporin expression in the Japanese medaka (*Oryzias latipes*) in freshwater and seawater: challenging the paradigm of intestinal water transport? *The Journal of Experimental Biology*, *217*(17), 3108 LP – 3121.

Madsen, S. S., Engelund, M. B., & Cutler, C. P. (2015). Water transport and functional dynamics of aquaporins in osmoregulatory organs of fishes. *Biological Bulletin*, *229*(1), 70–92.

Makarova, K. S., Wolf, Y. I., Mekhedov, S. L., Mirkin, B. G., & Koonin, E. v. (2005). Ancestral paralogs and pseudoparalogs and their role in the emergence of the eukaryotic cell. *Nucleic Acids Research*, *33*(14), 4626–4638.

Marchissio, M. J., Francés, D. E. A., Carnovale, C. E., & Marinelli, R. A. (2012). Mitochondrial aquaporin-8 knockdown in human hepatoma HepG2 cells causes ROS-induced mitochondrial depolarization and loss of viability. *Toxicology and Applied Pharmacology*, *264*(2), 246–254.

Margulis, L., Chapman, M., Guerrero, R., & Hall, J. (2006). The last eukaryotic common ancestor (LECA): acquisition of cytoskeletal motility from aerotolerant spirochetes in the Proterozoic Eon. *Proceedings of the National Academy of Sciences*, *103*(35), 13080–13085.

Martin, K. L. M. (1995). Time and tide wait for no fish: intertidal fishes out of water. *Environmental Biology of Fishes*, *44*, 165–181.

Martinez, A., Cutler, C. P., Wilson, G. D., Phillips, C., Hazon, N., Cramb, G., Wil-, G. D., & Cramb, G. (2005). Regulation of expression of two aquaporin homologs in the intestine of the European eel: effects of seawater acclimation and cortisol treatment. *American Journal of Physiology. Regulatory, Integrative and Comparative Physiology*, *288*(6), 1733–1743.

Martinez, A.-S., Cutler, C. P., Wilson, G. D., Phillips, C., Hazon, N., & Cramb, G. (2005). Cloning and expression of three aquaporin homologues from the European eel (*Anguilla anguilla*): effects of seawater acclimation and cortisol treatment on renal expression. *Biology of the Cell / under the Auspices of the European Cell Biology Organization*, *97*, 615–627.

Martínez-Redondo, G. I., Guerrero, C. S., Aristide, L., Balart-García, P., Tonzo, V., & Fernández, R. (2023). Parallel duplication and loss of aquaporin-coding genes during the "out of the sea" transition as potential key drivers of animal terrestrialization. *Molecular Ecology*, *00*, 1–19.

Maurel, C. (2007). Plant aquaporins: Novel functions and regulation properties. *FEBS Letters*, *581*(12), 2227–2236.

Maurel, C., Boursiac, Y., Luu, D.-T., Santoni, V., Shahzad, Z., & Verdoucq, L. (2015). Aquaporins in plants. *Physiological Reviews*, *95*(4), 1321–1358.

Meyer, A., Schloissnig, S., Franchini, P., Du, K., Woltering, J. M., Irisarri, I., Wong, W. Y., Nowoshilow, S., Kneitz, S., Kawaguchi, A., Fabrizius, A., Xiong, P., Dechaud, C., Spaink, H. P., Volff, J.-N., Simakov, O., Burmester, T., Tanaka, E. M., & Schartl, M. (2021). Giant lungfish genome elucidates the conquest of land by vertebrates. *Nature*, *590*(7845), 284–289.

Miller, M. A., Pfeiffer, W., & Schwartz, T. (2010). Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In *Gateway Computing Environments Workshop, GCE* (pp. 1–8).

Minh, B. Q., Nguyen, M. A. T., & Von Haeseler, A. (2013). Ultrafast approximation for phylogenetic bootstrap. *Molecular Biology and Evolution*, *30*(5), 1188–1195.

Morishita, Y., Matsuzaki, T., Hara-chikuma, M., Andoo, A., Shimono, M., Matsuki, A., Kobayashi, K., Ikeda, M., Yamamoto, T., Verkman, A., Kusano, E., Ookawara, S., Takata, K., Sasaki, S., & Ishibashi, K. (2005). Disruption of aquaporin-11 produces polycystic kidneys following vacuolization of the proximal tubule. *Molecular and Cellular Biology*, *25*(17), 7770–7779.

Moriyama, Y., Ito, F., Takeda, H., Yano, T., Okabe, M., Kuraku, S., Keeley, F. W., & Koshiba-Takeuchi, K. (2016). Evolution of the fish heart by sub/neofunctionalization of an elastin gene. *Nature Communications*, *7*, 10397.

Morris, C. (1892). The origin of lungs, a chapter in evolution. *The American Naturalist*, *26*(312), 975–986.

Mukhopadhyay, R., Bhattacharjee, H., & Rosen, B. P. (2014). Aquaglyceroporins: generalized metalloid channels. *Biochimica et Biophysica Acta (BBA) - General Subjects*, *1840*(5), 1583–1591.

Müller, C., Sendler, M., & Hildebrandt, J.-P. (2006). Downregulation of aquaporins 1 and 5 in nasal gland by osmotic stress in ducklings, *Anas platyrhynchos*: implications for the production of hypertonic fluid. *The Journal of Experimental Biology*, *209*(Pt 20), 4067–4076.

Munday, J. C., Settimo, L., & de Koning, H. P. (2015). Transport proteins determine drug sensitivity and resistance in a protozoan parasite, *Trypanosoma brucei*. *Frontiers in Pharmacology*, *6*. https://www.frontiersin.org/articles/10.3389/fphar.2015.00032

Murata, K., Mitsuoka, K., Hirai, T., Walz, T., Agre, P., Heymann, J. B., Engel, A., & Fujiyoshi, Y. (2000). Structural determinants of water permeation through aquaporin-1. *Nature*, *407*(6804), 599–605.

Murdy, E. O. (1989). A taxonomic revision and cladistic analysis of the oxudercine gobies (Gobiidae: Oxudercinae). *Records of the Australian Museum*, *Suppl. 11*, 1–93.

Murdy, E. O. (2011). Systematics of Oxudercinae. In Patzner, R.A., J. L. van Tassell, M. Kovačić, & B. G. Kapoor (Eds.), *The Biology of Gobies* (pp. 99–106). CRC Press and Science Publishers.

Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274.

Nielsen, R., & Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*, *148*(3), 929–936.

Nosil, P. (2012). *Ecological speciation*. Oxford University Press.

Notredame, C. (2007). Recent evolutions of multiple sequence alignment algorithms. *PLoS Computational Biology*, *3*(8), e123.

Öberg, F., Sjöhamn, J., Fischer, G., Moberg, A., Pedersen, A., Neutze, R., & Hedfalk, K. (2011). Glycosylation increases the thermostability of human aquaporin 10 protein. *The Journal of Biological Chemistry*, *286*(36), 31915–31923.

Ogden, T. H., & Rosenberg, M. S. (2006). Multiple sequence alignment accuracy and phylogenetic inference. *Systematic Biology*, *55*(2), 314–328.

Ogino, Y., Kuraku, S., Ishibashi, H., Miyakawa, H., Sumiya, E., Miyagawa, S., Matsubara, H., Yamada, G., Baker, M. E., & Iguchi, T. (2016). Neofunctionalization of androgen receptor by gain-of-function mutations in teleost fish lineage. *Molecular Biology and Evolution*, *33*(1), 228–244.

Ogushi, Y., Mochida, H., Nakakura, T., Suzuki, M., & Tanaka, S. (2007). Immunocytochemical and phylogenetic analyses of an arginine vasotocin-dependent aquaporin, AQP-h2K, specifically expressed in the kidney of the tree frog, *Hyla japonica*. *Endocrinology*, *148*(12), 5891–5901.

Ord, T. J., & Cooke, G. M. (2016). Repeated evolution of amphibious behavior in fish and its implications for the colonization of novel environments. *Evolution*, *70*(8), 1747–1759.

Ozu, M., Galizia, L., Acuña, C., & Amodeo, G. (2018). Aquaporins: more than functional monomers in a tetrameric arrangement. *Cells*, *7*(11), 209.

Pace, C. M., & Gibb, a C. (2009). Mudskipper pectoral fin kinematics in aquatic and terrestrial environments. *The Journal of Experimental Biology*, *212*(Pt 14), 2279–2286.

Pace, C. M., & Gibb, A. C. (2011). Locomotor behavior across an environmental transition in the ropefish, *Erpetoichthys calabaricus*. *The Journal of Experimental Biology*, *214*(4), 530 LP – 537.

Pace, C. M., & Gibb, A. C. (2014). Sustained periodic terrestrial locomotion in air-breathing fishes. *Journal of Fish Biology*, *84*(3), 639–660.

Pallas, P. S. (1780). *Spicilegia zoologica: quibus novae imprimis et obscurae animalium species iconibus, descriptionibus atque commentariis illustrantur*. Berolini,Gottl, August, Lange.

Pao, G. M., Wu, L.-F., Johnson, K. D., Höfte, H., Chrispeels, M. J., Sweet, G., Sandal, N. N., & Saier Jr, M. H. (1991). Evolution of the MIP family of integral membrane transport proteins. *Molecular Microbiology*, *5*(1), 33–37.

Park, J. H., & Saier, M. H. (1996). Phylogenetic characterization of the MIP family of transmembrane channel proteins. *Journal of Membrane Biology*, *153*(3), 171–180.

Pelster, B., & Wood, C. M. (2018). Ionoregulatory and oxidative stress issues associated with the evolution of air-breathing. *Acta Histochemica*, *120*(7), 667–679.

Pettersson, N., Filipsson, C., Becit, E., Brive, L., & Hohmann, S. (2005). Aquaporins in yeasts and filamentous fungi. *Biology of the Cell*, *97*(7), 487–500.

Phillips, A., Janies, D., & Wheeler, W. (2000). Multiple sequence alignment in phylogenetic analysis. *Molecular Phylogenetics and Evolution*, *16*(3), 317–330.

Pittis, A. A., & Gabaldón, T. (2016). Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry. *Nature*, *531*(7592), 101–104.

Polgar, G., Ghanbarifardi, M., Milli, S., Agorreta, A., Aliabadian, M., Esmaeili, H. R., & Khang, T. F. (2017). Ecomorphological adaptation in three mudskippers (Teleostei: Gobioidei: Gobiidae) from the Persian Gulf and the Gulf of Oman. *Hydrobiologia*, *795*(1), 91–111.

Pommerrenig, B., Diehn, T. A., Bernhardt, N., Bienert, M. D., Mitani-Ueno, N., Fuge, J., Bieber, A., Spitzer, C., Bräutigam, A., Ma, J. F., Chaumont, F., & Bienert, G. P. (2020). Functional evolution of nodulin 26-like intrinsic proteins: from bacterial arsenic detoxification to plant nutrient transport. *New Phytologist*, *225*(3), 1383–1396.

Preston, G. M., & Agre, P. (1991). Isolation of the cDNA for erythrocyte integral membrane protein of 28 kilodaltons: member of an ancient channel family. *Proceedings of the National Academy of Sciences*, *88*(24), 11110–11114.

Preston, G. M., Carroll, T. P., Guggino, W. B., & Agre, P. (1992). Appearance of water channels in *Xenopus oocytes* expressing red cell CHIP28 protein. *Science*, *256*(5055), 385–387.

Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2007). NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, *35*(SUPPL. 1), D61–D65.

R Development Core Team. (2016). R: a language and environment for statistical computing. In *R Found. Stat. Comput. Vienna Austria* (Vol. 0, Issue 2.11.1, p. 409).

Rambaut, A. *FigTree v1.4.4*. Retrieved January 27, 2022, from http://tree.bio.ed.ac.uk/software/figtree/

Randall, D. J., Burggren, W. W., Farrell, A. P., & Haswell, M. S. (1981). *The evolution of air breathing in vertebrates*. Cambridge University Press.

Randall, D. J., Ip, Y. K., Chew, S. F., & Wilson, J. M. (2015). Air breathing and ammonia excretion in the giant mudskipper, *Periophthalmodon schlosseri*. *Physiological and Biochemical Zoology : PBZ*, *77*(5), 783–788.

Reeves, J. H. (1992). Heterogeneity in the substitution process of amino acid sites of proteins coded for by mitochondrial DNA. *Journal of Molecular Evolution*, *35*(1), 17–31.

Richter, D. J., Berney, C., Strassert, J. F. H., Burki, F., & de Vargas, C. (2020). EukProt: a database of genome-scale predicted proteins across the diversity of eukaryotic life. *BioRxiv*.

Roger, A. J., Muñoz-Gómez, S. A., & Kamikawa, R. (2017). The origin and diversification of mitochondria. *Current Biology*, *27*(21), R1177–R1192.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., & Huelsenbeck, J. P. (2012). MrBayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, *61*(3), 539–542.

Ruhr, I. M., Wood, C. M., Schauer, K. L., Wang, Y., Mager, E. M., Stanton, B., & Grosell, M. (2020). Is aquaporin-3 involved in water-permeability changes in the killifish during hypoxia and normoxic recovery, in freshwater or seawater? *Journal of Experimental Zoology Part A: Ecological and Integrative Physiology*.

Ruvinsky, I., & Gibson-Brown, J. J. (2000). Genetic and developmental bases of serial homology in vertebrate limb evolution. *Development*, *127*(24), 5233–5244.

Saad, R., Cohanim, A. B., Kosloff, M., & Privman, E. (2018). Neofunctionalization in ligand binding sites of ant olfactory receptors. *Genome Biology and Evolution*, *10*(9), 2490–2500.

Saitoh, Y., Ogushi, Y., Shibata, Y., Okada, R., Tanaka, S., & Suzuki, M. (2014). Novel vasotocin-regulated aquaporins expressed in the ventral skin of semiaquatic anuran amphibians: evolution of cutaneous water-absorbing mechanisms. *Endocrinology*, *155*(6), 2166–2177.

Sakamoto, T., Mori, C., Minami, S., Takahashi, H., Abe, T., Ojima, D., Ogoshi, M., & Sakamoto, H. (2011). Corticosteroids stimulate the amphibious behavior in mudskipper: potential role of mineralocorticoid receptors in teleost fish. *Physiology & Behavior*, *104*(5), 923–928.

Sakamoto, T., Nishiyama, Y., Ikeda, A., Takahashi, H., Hyodo, S., Kagawa, N., & Sakamoto, H. (2015). Neurohypophysial hormones regulate amphibious behaviour in the mudskipper goby. *PLOS ONE*, *10*(7), e0134605.

San Mauro, D., & Agorreta, A. (2010). Molecular systematics: a synthesis of the common methods and the state of knowledge. *Cellular and Molecular Biology Letters*, *15*(2), 311–341.

Sánchez, R., Serra, F., Tárraga, J., Medina, I., Carbonell, J., Pulido, L., de María, A., Capella-Gutíerrez, S., Huerta-Cepas, J., Gabaldón, T., Dopazo, J., & Dopazo, H. (2011). Phylemon 2.0: a suite of web-tools for molecular evolution, phylogenetics, phylogenomics and hypotheses testing. *Nucleic Acids Research*, *39*(SUPPL. 2), W470–W474.

São Pedro, S. L., Alves, J. M. P., Barreto, A. S., & de Souza Lima, A. O. (2015). Evidence of positive selection of aquaporins genes from *Pontoporia blainvillei* during the evolutionary process of cetaceans. *PLoS ONE*, *10*(7), e0134516.

Saparov, S. M., Liu, K., Agre, P., & Pohl, P. (2007). Fast and selective ammonia transport by aquaporin-8. *Journal of Biological Chemistry*, *282*(8), 5296–5301.

Sayer, M. D. J. (2005). Adaptations of amphibious fish for surviving life out of water. *Fish and Fisheries*, *6*(3), 186–211.

Sayer, M. D. J., & Davenport, J. (1991). Amphibious fish: why do they leave water? *Reviews in Fish Biology and Fisheries*, *1*, 159–181.

Sayers, E. W., Beck, J., Bolton, E. E., Bourexis, D., Brister, J. R., Canese, K., Comeau, D. C., Funk, K., Kim, S., Klimke, W., Marchler-Bauer, A., Landrum, M., Lathrop, S., Lu, Z., Madden, T. L., O'Leary, N., Phan, L., Rangwala, S. H., Schneider, V. A., … Sherry, S. T. (2021). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *49*(D1), D10–D17.

Schadt, E. E., Turner, S., & Kasarskis, A. (2010). A window into third-generation sequencing. *Human Molecular Genetics*, *19*(R2), R227–R240.

Schluter, D. (2000). *The ecology of adaptive radiation*. OUP Oxford.

Schneider, I., & Shubin, N. H. (2012). Making limbs from fins. *Developmental Cell*, *23*(6), 1121–1122.

Schoch, C. L., Ciufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'Neill, K., Robbertse, B., Sharma, S., Soussov, V., Sullivan, J. P., Sun, L., Turner, S., & Karsch-Mizrachi, I. (2020). NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database*, *2020*, baaa062.

Schrödinger LLC. (2015). *The PyMOL Molecular Graphics System, Version~1.8*.

Schuler, G. D., Epstein, J. A., Ohkawa, H., & Kans, J. A. (1996). Entrez: molecular biology database and retrieval system. *Methods in Enzymology*, *266*, 141–161.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068–2069.

Self, S. G., & Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, *82*(398), 605–610.

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, *26*(10), 1135–1145.

Shubin, N., Tabin, C., & Carroll, S. (1997). Fossils, genes and the evolution of animal limbs. *Nature*, *388*(6643), 639–648.

Simonovic, P., Paunović, M., & Popović, S. (2001). Morphology, feeding, and reproduction of the round goby, *Neogobius melanostomus* (Pallas), in the Danube river basin, Yugoslavia. *Journal of Great Lakes Research*, *27*(3), 281–289.

Smith, H. W. (1930). Metabolism of the lung-fish *Protopterus aethiopicus*. *Journal of Biological Chemistry*, *88*(1), 97–130.

Soto, G., Alleva, K., Amodeo, G., Muschietti, J., & Ayub, N. D. (2012). New insight into the evolution of aquaporins from flowering plants and vertebrates: orthologous identification and functional transfer is possible. *Gene*, *503*(1), 165–176.

Soveral, G., & Casini, A. (2017). Aquaporin modulators: a patent review (2010–2015). *Expert Opinion on Therapeutic Patents*, *27*(1), 49–62.

Spring, K. R. (2011). Mechanism of fluid transport by epithelia. In *Handbook of physiology. The gastrointestinal system.* (Vol. 5, pp. 195–207).

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313.

Stavang, J. A., Chauvigné, F., Kongshaug, H., Cerdà, J., Nilsen, F., & Finn, R. N. (2015). Phylogenomic and functional analyses of salmon lice aquaporins uncover the molecular diversity of the superfamily in Arthropoda. *BMC Genomics*, *16*(1), 618.

Stechmann, A., & Cavalier-Smith, T. (2002). Rooting the eukaryote tree by using a derived gene fusion. *Science*, *297*(5578), 89–91.

Steppan, S. J., Meyer, A. A., Barrow, L. N., Alhajeri, B. H., Al-Zaidan, A. S. Y., Gignac, P. M., & Erickson, G. M. (2022). Phylogenetics and the evolution of terrestriality in mudskippers (Gobiidae: Oxudercinae). *Molecular Phylogenetics and Evolution*, *169*, 107416.

Strassert, J. F. H., Irisarri, I., Williams, T. A., & Burki, F. (2021). A molecular timescale for eukaryote evolution with implications for the origin of red algal-derived plastids. *Nature Communications*, *12*(1), 1879.

Strassert, J. F. H., Jamy, M., Mylnikov, A. P., Tikhonenkov, D. v, & Burki, F. (2019). New phylogenomic analysis of the enigmatic phylum telonemia further resolves the eukaryote tree of life. *Molecular Biology and Evolution*, *36*(4), 757–765.

Sui, H., Han, B. G., Lee, J. K., Walian, P., & Jap, B. K. (2001). Structural basis of water-specific transport through the AQP1 water channel. *Nature*, *414*(6866), 872–878.

Suzuki, M., Hasegawa, T., Ogushi, Y., & Tanaka, S. (2007). Amphibian aquaporins and adaptation to terrestrial environments: A review. *Comparative Biochemistry and Physiology - A Molecular and Integrative Physiology*, *148*(1 SPEC. ISS.), 72–81.

Suzuki, M., & Tanaka, S. (2009). Molecular and cellular regulation of water homeostasis in anuran amphibians by aquaporins. *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology*, *153*(3), 231–241.

Swanson, B. O., & Gibb, A. C. (2004). Kinematics of aquatic and terrestrial escape responses in mudskippers. *Journal of Experimental Biology*, *207*(23), 4037 LP – 4044.

Takahashi, S., Muta, K., Sonoda, H., Kato, A., Abdeen, A., & Ikeda, M. (2014). The role of Cysteine 227 in subcellular localization, water permeability, and multimerization of aquaporin-11. *FEBS Open Bio*, *4*, 315–320.

Takata, K., Matsuzaki, T., & Tajika, Y. (2004). Aquaporins: water channel proteins of the cell membrane. *Progress in Histochemistry and Cytochemistry*, *39*(1), 1–83.

Talavera, G., & Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology*, *56*(4), 564–577.

Tamura, K., Stecher, G., & Kumar, S. (2021). MEGA11: Molecular evolutionary genetics analysis version 11. *Molecular Biology and Evolution*, *38*(7), 3022–3027.

Tanghe, A., van Dijck, P., & Thevelein, J. M. (2006). Why do microorganisms have aquaporins? *Trends in Microbiology*, *14*(2), 78–85.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *American Mathematical Society: Lectures on Mathematics in the Life Sciences*, *17*, 57–86.

Tay, A. S. L., Chew, S. F., & Ip, Y. K. (2003). The swamp eel *Monopterus albus* reduces endogenous ammonia production and detoxifies ammonia to glutamine during 144 h of aerial exposure. *Journal of Experimental Biology*, *206*(14), 2473 LP – 2486.

Taylor, D. S. (2000). Biology and ecology of *Rivulus Marmoratus*: new insights and a review. *Florida Scientist*, *63*(4), 242–255.

Taylor, J. S., van de Peer, Y., Braasch, I., & Meyer, A. (2001). Comparative genomics provides evidence for an ancient genome duplication event in fish. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *356*(1414), 1661–1679.

Tesan, F. C., Lorenzo, R., Alleva, K., & Fox, A. R. (2021). AQPX-cluster aquaporins and aquaglyceroporins are asymmetrically distributed in trypanosomes. *Communications Biology*, *4*(1), 953.

Thompson, A. W., Hayes, A., Podrabsky, J. E., & Ortí, G. (2017). Gene expression during delayed hatching in fish-out-of-water. *Ecological Genetics and Genomics*, *3–5*, 52–59.

Thompson, J. D., Gibson, Toby. J., & Higgins, D. G. (2003). Multiple sequence alignment using ClustalW and ClustalX. *Current Protocols in Bioinformatics*, *1*, 2.3.1-2.3.22.

Thorne, J. L., Goldman, N., & Jones, D. T. (1996). Combining protein evolution and secondary structure. *Molecular Biology and Evolution*, *13*(5), 666–673.

Tingaud-Sequeira, A., Calusinska, M., Finn, R. N., Chauvigné, F., Lozano, J., & Cerdà, J. (2010). The zebrafish genome encodes the largest vertebrate repertoire of functional aquaporins with dual paralogy and substrate specificities similar to mammals. *BioMed Central Evolutionary Biology*, *10*(1), 38.

Tingaud-Sequeira, A., Chauvigné, F., Fabra, M., Lozano, J., Raldúa, D., & Cerdà, J. (2008). Structural and functional divergence of two fish aquaporin-1 water channels following teleost-specific gene duplication. *BMC Evolutionary Biology*, *8*(1), 259.

Tingaud-Sequeira, A., Zapater, C., Chauvigné, F., Otero, D., & Cerdà, J. (2009). Adaptive plasticity of killifish (*Fundulus heteroclitus*) embryos: dehydration-stimulated

development and differential aquaporin-3 expression. *American Journal of Physiology - Regulatory Integrative and Comparative Physiology*, *296*(4), 1041–1052.

Tipsmark, C. K., Sorensen, K. J., & Madsen, S. S. (2010). Aquaporin expression dynamics in osmoregulatory tissues of Atlantic salmon during smoltification and seawater acclimation. *Journal of Experimental Biology*, *213*(3), 368–379.

Törnroth-Horsefield, S., Hedfalk, K., Fischer, G., Lindkvist-Petersson, K., & Neutze, R. (2010). Structural insights into eukaryotic aquaporin regulation. *FEBS Letters*, *584*(12), 2580–2588.

Törnroth-Horsefield, S., Wang, Y., Hedfalk, K., Johanson, U., Karlsson, M., Tajkhorshid, E., Neutze, R., & Kjellbom, P. (2006). Structural mechanism of plant aquaporin gating. *Nature*, *439*, 688–694.

Tsuhako, Y., Ishimatsu, A., Takeda, T., Huat, K. K., & Tachihara, K. (2003). The eggs and larvae of the giant mudskipper, *Periophthalmodon schlosseri*, collected from a mudflat in Penang, Malaysia. *Ichthyological Research*, *50*(2), 178–181.

Turko, A. J., Cooper, C. A., & Wright, P. A. (2012). Gill remodelling during terrestrial acclimation reduces aquatic respiratory function of the amphibious fish *Kryptolebias marmoratus*. *Journal of Experimental Biology*, *215*(22), 3973–3980.

Turko, A. J., Robertson, C. E., Bianchini, K., Freeman, M., & Wright, P. A. (2014). The amphibious fish *Kryptolebias marmoratus* uses different strategies to maintain oxygen delivery during aquatic hypoxia and air exposure. *Journal of Experimental Biology*, *217*(22), 3988–3995.

Turko, A. J., & Wright, P. A. (2015). Evolution, ecology and physiology of amphibious killifishes (Cyprinodontiformes). *Journal of Fish Biology*, *87*(4), 815–835.

Valenciennes, A. (1837). Histoire naturelle des Poissons. Tome douzième. Suite du livre quatorzième. Gobioïdes. Livre quinzième. Acanthoptérygiens à pectorales pédiculées. *FG Levrault, Paris*, *12*(1–24), 1–507.

van Balkom, B. W. M., Savelkoul, P. J. M., Markovich, D., Hofman, E., Nielsen, S., van der Sluijs, P., & Deen, P. M. T. (2002). The role of putative phosphorylation sites in the targeting and shuttling of the aquaporin-2 water channel. *Journal of Biological Chemistry*, *277*(44), 41473–41479.

Verma, R. K., Prabh, N. D., & Sankararamakrishnan, R. (2014). New subfamilies of major intrinsic proteins in fungi suggest novel transport properties in fungal channels: implications for the host-fungal interactions. *BMC Evolutionary Biology*, *14*(1), 173.

Virkki, L. V, Franke, C., Somieski, P., & Boron, W. F. (2002). Cloning and functional characterization of a novel aquaporin from *Xenopus laevis* oocytes. *The Journal of Biological Chemistry*, *277*(43), 40610–40616.

Vosseberg, J., van Hooff, J. J. E., Marcet-Houben, M., van Vlimmeren, A., van Wijk, L. M., Gabaldón, T., & Snel, B. (2021). Timing the origin of eukaryotic cellular complexity with ancient duplications. *Nature Ecology & Evolution*, *5*(1), 92–100.

Wallace, I. S., Choi, W.-G., & Roberts, D. M. (2006). The structure, function and regulation of the nodulin 26-like intrinsic protein family of plant aquaglyceroporins. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, *1758*(8), 1165–1175.

Walz, T., Hirai, T., Murata, K., Heymann, J. B., Mitsuoka, K., Fujiyoshi, Y., Smith, B. L., Agre, P., & Engel, A. (1997). The three-dimensional structure of aquaporin-1. *Nature*, *387*, 624–627.

Watanabe, S., Kaneko, T., & Aida, K. (2005). Aquaporin-3 expressed in the basolateral membrane of gill chloride cells in Mozambique tilapia *Oreochromis mossambicus* adapted to freshwater and seawater. *Journal of Experimental Biology*, *208*(14), 2673–2682.

Watanabe, S., Moniaga, C. S., Nielsen, S., & Hara-Chikuma, M. (2016). Aquaporin-9 facilitates membrane transport of hydrogen peroxide in mammalian cells. *Biochemical and Biophysical Research Communications*, *471*(1), 191–197.

Wheeler, D., & Bhagwat, M. (2007). BLAST QuickStart: example-driven web-based BLAST tutorial. *Methods in Molecular Biology*, *395*, 149–176.

Wolf, Y. I., & Koonin, E. v. (2013). Genome reduction as the dominant mode of evolution. *BioEssays*, *35*(9), 829–837.

Wright, P. A. (1995). Nitrogen excretion: three end products, many physiological roles. *The Journal of Experimental Biology*, *198*(2), 273 LP – 281.

Wright, P. A., & Turko, A. J. (2016). Amphibious fishes: evolution and phenotypic plasticity. *The Journal of Experimental Biology*, *219*(15), 2245–2259.

Wspalz, T., Fujiyoshi, Y., & Engel, A. (2009). The AQP structure and functional implications. In E. Beitz (Ed.), *Aquaporins. Handbook of Experimental Pharmacology* (pp. 31–56). Springer Berlin Heidelberg.

Wu, B., & Beitz, E. (2007). Aquaporins with selectivity for unconventional permeants. *Cellular and Molecular Life Sciences*, *64*(18), 2413–2421.

Yakata, K., Hiroaki, Y., Ishibashi, K., Sohara, E., Sasaki, S., Mitsuoka, K., & Fujiyoshi, Y. (2007). Aquaporin-11 containing a divergent NPA motif has normal water channel activity. *Biochimica et Biophysica Acta - Biomembranes*, *1768*(3), 688–693.

Yakata, K., Tani, K., & Fujiyoshi, Y. (2011). Water permeability and characterization of aquaporin-11. *Journal of Structural Biology*, *174*(2), 315–320.

Yang, Z. (1993). Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution*, *10*(6), 1188–1195.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution*, *39*(3), 306–314.

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, *24*(8), 1586–1591.

Yang, Z. (2008). Adaptive molecular evolution. In *Handbook of Statistical Genetics: Third Edition* (Vol. 1, pp. 375–406).

Yang, Z., & Dos Reis, M. (2011). Statistical properties of the branch-site test of positive selection. *Molecular Biology and Evolution*, *28*(3), 1217–1228.

Yang, Z., Wong, W. S. W., & Nielsen, R. (2005). Bayes empirical bayes inference of amino acid sites under positive selection. *Molecular Biology and Evolution*, *22*(4), 1107–1118.

Yilmaz, O., Chauvigné, F., Ferré, A., Nilsen, F., Fjelldal, P. G., Cerdà, J., & Finn, R. N. (2020). Unravelling the complex duplication history of deuterostome glycerol transporters. *Cells*, *9*(7), 1663.

You, X., Bian, C., Zan, Q., Xu, X., Liu, X., Chen, J., Wang, J., Qiu, Y., Li, W., Zhang, X., Sun, Y., Chen, S., Hong, W., Li, Y., Cheng, S., Fan, G., Shi, C., Liang, J., Tom Tang, Y., … Shi, Q. (2014). Mudskipper genomes provide insights into the terrestrial adaptation of amphibious fishes. *Nature Communications*, *5*, 5594.

You, X., Sun, M., Li, J., Bian, C., Chen, J., Yi, Y., Yu, H., & Shi, Q. (2018). Mudskippers and their genetic adaptations to an amphibious lifestyle. *Animals*, *8*(2), 24.

Zander, C. D. (2011). Morphological adaptations to special environments of gobies. In R. A. Patzner, J. L. van Tassell, M. Kovačić, & B. G. Kapoor (Eds.), *The Biology of Gobies* (pp. 345–366). CRC Press and Science Publishers.

Zang, Y., Chen, J., Zhong, H., Ren, J., Zhao, W., Man, Q., Shang, S., & Tang, X. (2019). Genome-wide analysis of the aquaporin gene family in reptiles. *International Journal of Biological Macromolecules*, *126*, 1093–1098.

Zardoya, R. (2005). Phylogeny and evolution of the major intrinsic protein family. *Biology of the Cell*, *97*(6), 397–414.

Zardoya, R., Ding, X., Kitagawa, Y., & Chrispeels, M. J. (2002). Origin of plant glycerol transporters by horizontal gene transfer and functional recruitment. *Proceedings of the National Academy of Sciences*, *99*(23), 14893–14896.

Zardoya, R., Irisarri, I., & Abascal, F. (2016). Aquaporin discovery in the genomic era. In G. Soveral, S. Nielsen, & A. Casini (Eds.), *Aquaporins in health and disease: New molecular targets for drug discovery* (1st ed., pp. 19–32). CRC Press Boca Raton, FL, USA.

Zardoya, R., & Villalba, S. (2001). A phylogenetic framework for the aquaporin family in Eukaryotes. *Journal of Molecular Evolution*, *52*(5), 391–404.

Zaremba-Niedzwiedzka, K., Caceres, E. F., Saw, J. H., Bäckström, D., Juzokaite, L., Vancaester, E., Seitz, K. W., Anantharaman, K., Starnawski, P., Kjeldsen, K. U., Stott, M. B., Nunoura, T., Banfield, J. F., Schramm, A., Baker, B. J., Spang, A., & Ettema, T. J. G. (2017). Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature*, *541*(7637), 353–358.

Zhang, J., Nielsen, R., & Yang, Z. (2005). Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Molecular Biology and Evolution*, *22*(12), 2472–2479.

Zhang, J., Taniguchi, T., Takita, T., & Ali, A. B. (2000). On the epidermal structure of *Boleophthalmus* and *Scartelaos* mudskippers with reference to their adaptation to terrestrial life. *Ichthyological Research*, *47*(3), 359–366.

Zhang, J., Taniguchi, T., Takita, T., & Ali, A. B. (2003). A study on the epidermal structure of *Periophthalmodon* and *Periophthalmus* mudskippers with reference to their terrestrial adaptation. *Ichthyological Research*, *50*(4), 310–317.