Biochimica et Biophysica Acta 1840 (2014) 1468–1481

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

journal homepage: www.elsevier.com/locate/bbagen

Diversity and evolution of membrane intrinsic proteins ☆

Federico Abascal [a,1], Iker Irisarri [b,1,2], Rafael Zardoya [b,*]

[a] Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), Melchor Fernández Almagro, 3, 28029 Madrid, Spain
[b] Department of Biodiversity and Evolutionary Biology, Museo Nacional de Ciencias Naturales-CSIC (MNCN-CSIC), José Gutiérrez Abascal 2, 28006 Madrid, Spain

ARTICLE INFO

Article history:
Received 18 September 2013
Received in revised form 5 December 2013
Accepted 9 December 2013
Available online 16 December 2013

Keywords:
Membrane intrinsic protein
Aquaporin
Evolutionary relationship
Molecular phylogeny

ABSTRACT

Background: Membrane intrinsic proteins (MIPs) are the proteins in charge of regulating water transport into cells. Because of this essential function, the MIP family is ancient, widespread, and highly diverse.
Scope of review: The rapidly accumulating genomic and transcriptomic data from previously poorly known groups such as unicellular eukaryotes, fungi, green algae, mosses, and non-vertebrate animals are contributing to expand our view of MIP evolution throughout the diversity of life. Here, by analyzing more than 1700 sequences, we provide an updated and comprehensive phylogeny of MIPs
Major conclusions: The reconstructed phylogeny supports (i) deep orthology of X intrinsic proteins (XIPs; present from unicellular eukaryotes to plants); (ii) that the origin of small intrinsic proteins (SIPs) traces back to the common ancestor of all plants; and (iii) the expansion of aquaglyceroporins (GLPs) in Oomycetes, as well as their loss in vascular plants and in the ancestor of endopterygote insects. Additionally, conserved positions in the protein, and residues involved in glycerol selectivity are reviewed within a phylogenetic framework. Furthermore, functional diversification of human and *Arabidopsis* paralogs are analyzed in an evolutionary genomic context.
General significance: Our results show that while bacteria and archaea generally function with one copy of each a water channel (aquaporin or AQP) and a GLP, recurrent independent expansions have greatly diversified the structures and functions of the different members of both MIP paralog subfamilies throughout eukaryote evolution (and not only in flowering plants and vertebrates, as previously thought). This article is part of a Special Issue entitled Aquaporins.

1. Introduction

Membrane intrinsic proteins (MIPs), also termed aquaporins, are ubiquitous channel proteins of molecular mass 26–35 kDa that facilitate the transport of water and small solutes (particularly glycerol but also urea, ammonia, metalloids, or even carbon dioxide; [1,2] across cell membranes in all living organisms [3–6]). These proteins are essential for life and constitute an ancient, abundant, and highly diversified protein family whose members are recognized by several key conserved structural features [7,8]. MIPs form tetramers in cell membranes [9], and each monomer is made of six transmembrane helices that delimit a pore with two selectivity filters. One is formed by two opposite NPA (Asn–Pro–Ala) motifs that establish hydrogen bonds with the water molecule and create an electrostatic repulsion of protons [10,11]. The other selectivity filter is named ar/R, and it is formed by two aromatic amino acids and one Arg, which create the narrowest section of the channel, and are thought to be determinant for substrate specificity [12].

The great diversity of forms and functions displayed by the different members of the MIP family can only be fully understood within an evolutionary framework. Several phylogenies of the whole family and of different members have been reported [13–20]. The earliest node in the MIP phylogenetic tree most likely corresponds to the ancient split of MIPs into two distinct subfamilies: water channels or aquaporins (AQPs) and glycerol transporters or aquaglyceroporins (GLPs). Up to five relatively conserved amino acid residues designed P1–P5 [21] discriminate AQPs from GLPs. Of these, P2 (Asp), located after the second NPA motif, is crucial in increasing the size of the pore to permeate larger molecules such as glycerol [22]. Bacteria and archaea, generally retain the ancestral condition of the family and have one AQP and one GLP that function as water and glycerol transporters, respectively. However, in eukaryotes many gene duplications have occurred greatly diversifying protein structures and functions [19]. Best-known examples are found in vertebrates and flowering plants, where expansions of the family are linked in particular to subfunctionalization of the different paralogs in different tissues. For instance, up to 13 and 35 genes have been described in human [17] and *Arabidopsis* [23,24], respectively. In other cases, diversification is achieved through neofunctionalization, either after gene duplication, as different analyses have suggested for intracellular aquaporins [17,25], or through co-option as has been proposed for the glycerol transport in the NOD26-like intrinsic proteins (NIPs) of plants [26].

☆ This article is part of a Special Issue entitled Aquaporins.
* Corresponding author.
E-mail address: rafaz@mncn.csic.es (R. Zardoya).
[1] These authors contributed equally to the paper.
[2] Present address: Laboratory for Zoology and Evolutionary Biology, Department of Biology, University of Konstanz, Universitätsstraße 10, 78457 Konstanz, Germany.

0304-4165/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.bbagen.2013.12.001

Current sequencing technologies have dramatically expanded the number of genome projects of non-model organisms [27]. Moreover, available transcriptomes allow investigating expression profiles of protein family members in an evolutionary context [28,29]. For MIPs, a wealth of new sequence, structural, and functional data is accumulating rapidly, providing an unprecedented rich outlook on the evolution and diversification of the family (e.g. [14]). As the sequence data of new complete genomes have been released, genetic information on previously poorly known MIPs from e.g. non-vertebrate animals, unicellular eukaryotes, or early-branched plants has become available, being crucial in deciphering the origins of the different MIP subfamilies, and determining to what extent known sequence motifs associated to functional specificity are conserved across living organisms. For instance, the great diversity of MIP paralogs found in angiosperms such as *Arabidopsis*, maize, and rice [30] (with up to five subfamilies: TIPs or tonoplast intrinsic proteins; PIPs or plasma membrane intrinsic proteins; NIPs or NOD26-like intrinsic proteins; SIPs or small basic intrinsic proteins; and XIPs or X intrinsic proteins), has indeed its roots in the most basal lineages of land plants including spike mosses (*Selaginella*; [14]) and mosses (*Physcomitrella*; [15]). These early-branched land plants possess additional MIP subfamilies adding up to a total of six (including also HIPs or hybrid intrinsic proteins) and seven (including additionally GIPs or GlpF-like intrinsic proteins) subfamilies for *Selaginella* and *Physcomitrella*, respectively [14,15]. Moreover, paralog diversity could be extended further back in time in the plant lineage since green algal MIPs can also be grouped into seven subfamilies, of which PIPs and GIPs are common to land plants and the rest (named MIP A–E) are specific of the green algae lineage [13]. Besides gene duplications, events of horizontal gene transfer (HGT) have also been suggested to be in part responsible for the diversity of MIPs in plants [13,26,31]. Thus far, the analyses of animal MIPs indicate that maximum diversity in terms of total number of subfamilies (AQP0 to 12) is achieved in fishes [32] and land vertebrates [19] as a result of several rounds of whole genome duplication.

Here, we reconstructed phylogenetic relationships within the MIP family using proteins from genome projects available in public databases. Our goal was to set an updated phylogenetic framework onto which interpret structural and functional patterns observed in MIPs across living organisms in order to understand the evolution of this protein family and how its extraordinary diversity was generated.

2. Diversity and evolutionary relationships of MIPs

A careful and balanced selection of publicly available genomes was conducted at UniProt [33] and Ensembl [34]. A total of 1714 MIPs were retrieved using hits with MIPs of Pfam [35] and InterPro [36] from 175 complete proteomes and the incomplete proteomes of *Equisetales*, conifers, and annelids (added to improve lineage representation). In addition, certain proteins were included a posteriori to improve the representation of XIPs, GIPs and unicellular eukaryote MIPs. We filtered out proteins that aligned less than 150 amino acids with the Pfam MIP Hidden Markov Model (HMM), rendering an alignment of 1613 MIPs. When sequences from the same species were more than 99% identical, only one was retained, reducing the alignment to 1489 sequences. A preliminary phylogenetic analysis using this alignment was performed to delimit major groups of paralogy within the MIP family (the alignment and tree are available at http://pc16141.mncn.csic.es/aqps.html). Further selection of sequences was done based on the preliminary phylogeny aiming to reconstruct a global phylogeny of MIPs that (1) maximized lineage diversity; (2) avoided overrepresentation of species from lineages that currently concentrate sequencing efforts such as e.g. bacteria, flowering plants and mammals; and (3) reduced when possible the number of long branches. In addition, phylogenies of the different subfamilies were reconstructed separately. Sequences were aligned with Mafft v. 7.055 (using the E-INS-i strategy optimized for alignments with multiple conserved domains and long gaps; [37]) and trimmed using trimAL v.1.3 [38] with a gap threshold of 0.8. Phylogenetic reconstruction was performed under maximum likelihood using the rapid hill-climbing algorithm as implemented in RAxML v.7.2.8 [39]. Prottest v.3.0 [40–42] was used to determine best-fit models for each alignment (see Appendix A). Support for internal branches was evaluated by non-parametric bootstrapping [43] with 1000 pseudo-replicates.

The global phylogeny of MIPs included 162 selected sequences (a graphical summary is shown in Fig. 1 and the full phylogeny is shown in the Appendix A). Since there is no known outgroup to the MIP family, the root of the phylogenetic tree had to be placed arbitrarily within the ingroup. We looked for a branch in the tree to place the root that needs to be highly supported (100% bootstrap), taxonomy congruent, and long. The only branch that fulfilled the above criteria is the one that separates GLPs from the rest of the MIPs (hereafter referred as AQPs), which automatically forms two reciprocally monophyletic groups. This rooting is further supported by previous phylogenetic analyses [19], and the fact that the ancestral state in bacteria is the possession of one GLP and one AQP. Previous phylogenies that were mostly restricted to vertebrates and flowering plants showed good resolution (e.g., [18,19]). However, reconstructing the present phylogeny proved more challenging given the heterogeneous rates of substitution among paralogs and among the numerous lineages, as well as the lack of enough shared derived alignment positions to unite different homologs. Hence, the general lack of statistical support for many internal nodes and the misplacement (according to taxonomy) of some lineages are mostly due to long-branch attraction (LBA) phenomena (Fig. 1). Nevertheless, several interesting phylogenetic patterns were evident, confirming and expanding previous knowledge on the evolution of this gene family. The first and most apparent phylogenetic pattern is the significant difference in terms of diversification between GLPs and AQPs. The former are a rather compact group in which paralog subfamily diversification is most obvious in vertebrates, whereas in the latter, plants and animals experienced successive events of gene duplication accompanied by extraordinary sequence and functional divergences that generated the exceptional diversity of subfamilies found in AQPs (Fig. 1; [19]), and caused some AQPs (e.g., SIPs and AQP11/12) to be hardly recognizable as members of the MIP family. By alternatively rooting the phylogeny in any of the poorly resolved, short nodes at the base of the AQP clade (e.g., at SIPs and AQP11/12 or at NIPs), an alternative, less plausible, view of the evolution of the family could be hypothesized, in which non-GLPs would be paraphyletic with respect to GLPs. While the evolutionary scenario preferred here proposes an increase in complexity of the family during its evolutionary history through successive gene duplication events and functional divergence in the two subfamilies (AQPs and GLPs), the latter view would imply in contrast ancient diversity of MIP families and multiple independent losses in important groups such as bacteria, archaea, unicellular eukaryotes, fungi and non-vertebrate animals.

Another interesting pattern is the presence of GLPs in green algae (*Chlorophyta*) and mosses (*Bryophyta*). This is noteworthy, as they are absent in vascular plants (*Tracheophyta*) (Fig. 1; [14]). Plant GIPs show a long branch that is recovered within bacterial GLPs, supporting their origin from a HGT event from bacteria [31]. The secondary loss of GLPs in vascular plants might be related to redundancy with NIP function as glycerol transporter [26] (see below).

Bacteria and archaea generally possess a single AQP copy, and unicellular eukaryotes and fungi show heterogeneous number of genes, whereas the diversification of AQPs is most outstanding in plants and animals. The phylogenetic tree shows three major AQP groups that encompass plants and animals: (1) plant SIPs plus animal AQPs 11 and 12; (2) plant XIPs, HIPs and TIPs plus animal AQP8; and (3) plant PIPs plus animal AQPs 4, 1, 0, 2, 5 and 6. It is tempting to propose that these three groupings represent instances of deep orthology, and that diversity within AQPs could be concentrated in few early gene duplication events in the ancestor of eukaryotes [18]. However, this possibility is difficult to prove because (1) support for the corresponding internal

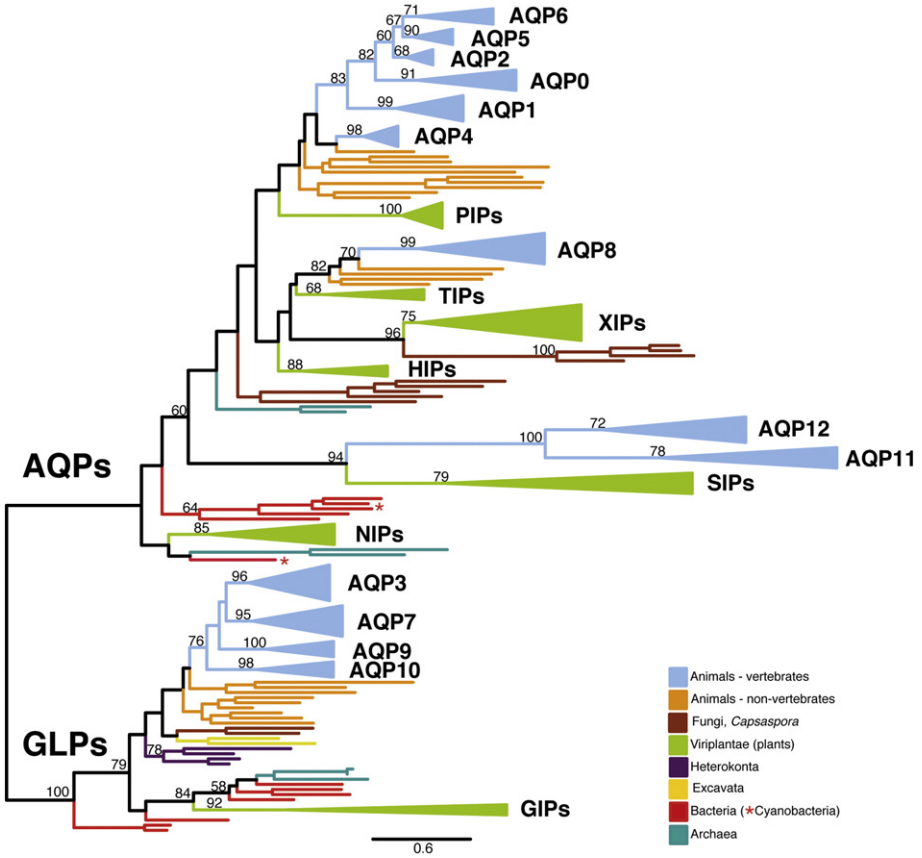

Fig. 1. General phylogeny of MIP proteins. Major subfamilies of MIP proteins are shown as collapsed sets of nodes. In this and following tree figures, branches are colored according to the taxonomy in the accompanying legends. Numbers above nodes indicate bootstrap support from 1000 pseudoreplicates (in percentage) for selected nodes of interest.

nodes in the phylogenetic tree is minimal; (2) the only one of these internal nodes grouping animal (only vertebrates) and plant AQPs that has relatively strong support (SIPs + AQPs 11 and 12) leads to extremely long branches, and thus the possibility of LBA is very high; and (3), both the lack of fungi orthologs as sister groups of animals (Opisthokonta) in two of the three groupings, and the lack of non-vertebrate orthologs at the base of AQP11 and AQP12 would require postulating many independent gene losses, which is unlikely. An alternative hypothesis to deep orthology for the three above-mentioned groupings would be that AQPs in plants and animals have independent origins associated to the need of functional diversification of AQPs in these complex multicellular organisms, and that analogous functions in plants and animals may have led through natural selection to the minimal sequence convergence required to spuriously unite the three groups in the phylogenetic tree. Because we are dealing with a phylogenetic pattern of ancient and short internal nodes leading to relatively long branches at the tips, discerning between both scenarios is challenging. Exceptionally, XIPs form a monophyletic group that includes orthologs from unicellular eukaryotes, fungi, and plants, and this could be in support of deep orthology of XIPs, HIPs, TIPs and AQP8s.

A final, most intriguing evolutionary pattern derived from the reconstructed phylogeny is the relative position of NIPs. These proteins are found from mosses to flowering plants (Fig. 1; [44]). NIPs are multifunctional proteins found in plant nodules with a high glycerol transport rate, low intrinsic water permeability, and capable of transporting formamide, urea, ammonia, and metalloids [16]. Moreover, they show their own distinct ar/R filters [45]. In the reconstructed phylogeny, they are grouped with low statistical support together with cyanobacterial and archaeal (NIP-like) proteins as the most basal lineage within AQPs and separated from other bacterial AQPs (Fig. 1; [16]). This phylogenetic pattern further supports that plant NIPs could have been acquired through HGT [26], and may point to the bacterial groups involved in the event (as per the ar/R filter of the bacterial NIP-like, glycerol transporting could be already a property of these bacterial proteins [16] or a function co-opted in plants after HGT). Alternatively, it could indicate an ancient origin of NIPs that would trace back to some lineages of bacteria [16], but this hypothesis would require recurrent losses in all eukaryote lineages but land plants. Finally, convergent evolution in several residues of some NIP-like bacterial proteins and NIPs cannot be discarded. The ability of NIPs to transport glycerol could explain the loss of GIPs in flowering plants due to functional redundancy. In this regard, it would be interesting to experimentally characterize glycerol transport in mosses, which are the only organisms in which GIPs and NIPs coexist [15], as well as to determine the function of NIP-like proteins in those bacteria where they coexist with GLPs [16].

Additional phylogenies were separately reconstructed for bacteria and archaea (Fig. 2), unicellular eukaryotes (Fig. 3A), fungi (Fig. 3B), and animals (Fig. 4). Given the high diversity of MIPs in plants, independent phylogenies for PIPs, TIPs, SIPs (Fig. 5) and NIPs (Fig. 6) were also reconstructed. These phylogenies were based on larger taxon samplings compared to the general phylogeny (Fig. 1), allowing a zooming in into each of the analyzed groups. Moreover, since they were focused on subsets of the global phylogeny of MIPs, they were based on more phylogenetically informative alignment positions thus providing further support to the different internal nodes, and allowing more robust conclusions regarding internal phylogenetic relationships.

2.1. Bacterial and archaeal MIPs

In the phylogeny of bacterial and archaeal MIPs, there was no resolution further than clearly separating GLPs from AQPs (Fig. 2). Within each paralog group, the phylogeny of bacteria and archaea at the phylum level was not recovered, which could be due to either lack of enough shared derived positions to reconstruct the corresponding

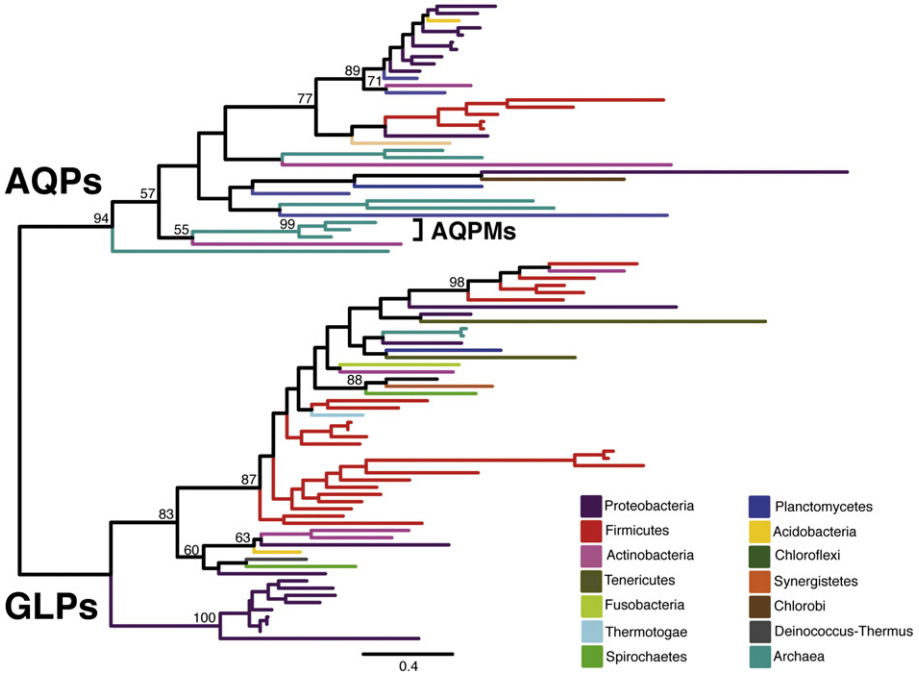

Fig. 2. Phylogeny of bacterial and archaeal MIP proteins.

nodes or the presence of HGT phenomena. The distribution in number of copies of GLPs and AQPs in bacteria is rather heterogeneous (see Appendix A). This distribution does not correlate with the bacterial phylogeny (Fig. 2) but it is likely associated to different lifestyles. Some genera have one copy of each paralog such as *Escherichia*, *Shigella*, *Pseudomonas*, *Vibrio*, *Burkholderia*, or *Bacillus*. Many seem to have only GLPs such as *Staphylococcus*, *Listeria*, *Clostridium*, *Salmonella*, *Yersinia* or *Borrelia*. On the opposite side are nitrogen-fixing bacteria (Alphaproteobacteria: Rhizobiales) such as *Rhizobium*, *Agrobacterium*, *Nitrobacter*, *Sinorhizobium* or *Methylobacterium* that have only AQPs. It is noteworthy that many intracellular bacteria such as *Rickettsia*, *Coxiella*, *Bartonella*, *Wolbachia*, *Ehrlichia* or *Chlamydia* have no MIPs, likely lost during the host-pathogen co-evolution process. In addition, *Thermotoga*, *Aquifex* or *Thermus* do not have MIPs, which might be related to their thermophilic lifestyle. Thus far, we have not found any Archaea that has both GLPs and AQPs concurrently. Methane-producing archaea such as *Methanococcus*, *Methanothermobacter*, *Methanosarcina* or *Methanobrevibacter* have only AQPs. However, these AQPs are reported to have special properties being able to also transport glycerol [46]. Few GLPs were found in Archaea, all of them within Halobacteria. Most Archaea do not have any MIP including *Thermococcus*, *Pyrococcus*, *Thermoplasma* or *Pyrobaculum*, which are thermophilic.

2.2. Unicellular eukaryote and fungal MIPs

Unicellular eukaryotes (traditionally classified as "protists", but currently known to be paraphyletic; Fig. 3A) and fungi (Fig. 3B) follow a similar pattern with a clear division between GLPs and AQPs (that include XIP orthologs), and a heterogeneous distribution in the number of copies of each paralog in the different genera. Until recently, the evolutionary history of unicellular eukaryote MIPs was the least known due to the paucity of available data [19]. Although, our genomic data set is still far from complete in terms of lineage representation, a clear pattern of MIP expansions is evident within some of the major groups analyzed (Amoebozoa, Heterokonta, Euglenozoa, Choanoflagellida, Alveolata) (Fig. 3A). While some genera such as *Naegleria* (Heterolobosea) and *Capsaspora* (Filasterea) have few paralogs, others experienced spectacular bursts of gene duplications, being the best examples *Paramecium* (Alveolata), *Phytophthora* (Heterokonta; Oomycetes), and *Leishmania* and *Trypanosoma* (Euglenozoa). Interestingly, in *Paramecium* and *Leishmania* the bulk of the expansion occurs in the AQPs whereas in *Phytophthora* and other Oomycetes (Heterokonta) it occurs in the GLPs. An interesting distribution pattern is found within *Trypanosoma* (Euglenozoa) with the different species having either AQPs or GLPs but not both (see Appendix A). This pattern is not observed in the closely related *Leishmania*, indicating that relatively recent gene losses occurred in the different *Trypanosoma* species. Neither Heterolobosea nor Amoebozoa have GLPs, but the number of species sequenced for Heterolobosea (thus far only one) is insufficient to draw a firm conclusion in this respect. On the other hand, analyzed Oomycetes (*Phytophthora*, *Hyaloperonospora* and *Pythium*) have several GLPs but no AQP, contrasting to other heterokont phyla including Bacillariophyceae (diatoms), Phaeophyceae (brown algae) and Pelagophyceae, which experienced less duplications but have both AQPs and GLPs. This pattern found within Heterokonta may reveal an evolutionary relationship between the loss of AQPs and consequent GLPs expansion or vice versa. Interestingly, the species *Tetrahymena thermophila* (Alveolata), which lives at high temperatures, has neither AQP nor GLP, as happens in termophilic bacteria (see Appendix A). MIPs are also absent from *Giardia intestinalis* (Fornicata; not included in the phylogenetic tree).

In fungi, classical AQPs and GLPs are evenly distributed across phyla. The analyzed species posses between none and six copies of each AQPs and GLPs, and always have at least one of the two (see Appendix A). Fungal XIP orthologs [47] are recovered as sister group to the remaining AQPs (Fig. 3B; [48]) whereas classic AQPs fail to group together in unicellular eukaryotes and the relative phylogenetic position of XIPs remains unresolved in this group (Fig. 3A). All analyzed fungal phyla (Basidiomycota, Ascomycota, Microsporidia and Chytridiomycota) have representatives of XIPs, but only few species within each phylum actually have XIP orthologs. In contrast, unicellular eukaryote XIPs are concentrated in the order Dictyosteliida (Amoebozoa).

2.3. Animal MIPs

Within animals, GLP and AQP phylogenies were reconstructed separately (Fig. 4). Expansion of GLPs into paralog groups is widespread in all major lineages but most apparent in vertebrates due to the larger

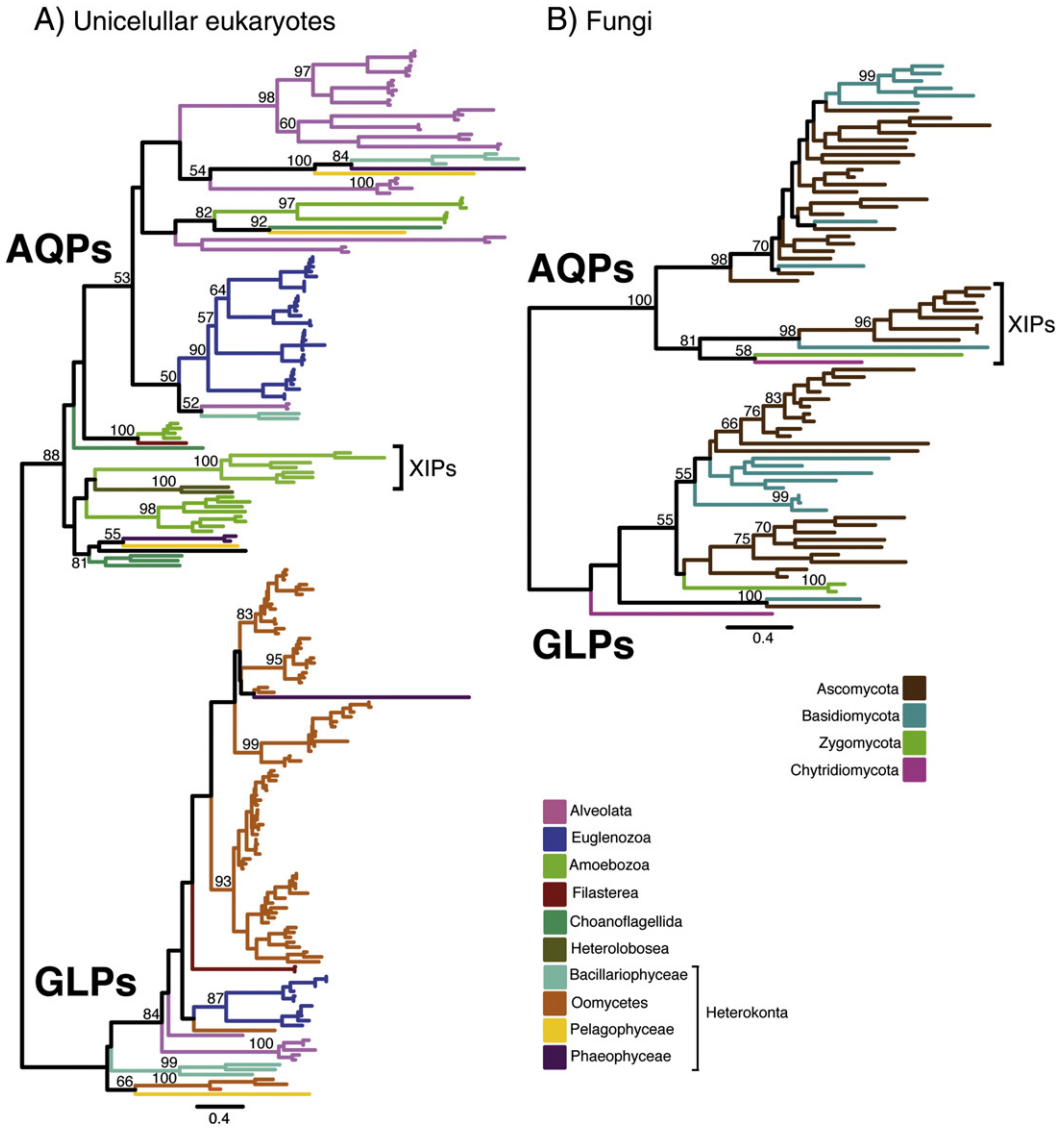

Fig. 3. Phylogeny of MIP proteins in unicellular eukaryotes (A) and Fungi (B).

taxon sampling [18,19]. For instance, in nematodes, up to five different paralog groups could be recognized but due to the lower sampling effort centered exclusively on Rhabditida, it is not possible to trace precisely their origins within the phylum at the moment. An interesting pattern is found within arthropods: GLPs are found in Quelicerata (*Ixodes*), Branchiopoda (*Daphnia*), and in Anoplura (*Pediculus*) within Insecta. However, they are missing in other Insecta such as Hymenoptera (*Apis*, *Atta*), Lepidoptera (*Bombyx*), Diptera (*Drosophila*, *Aedes*, *Anopheles*), and Coleoptera (*Tribolium*), indicating the loss of GLPs in the ancestor of endopterygote insects. The diversity of GLP paralogs (AQP3, 7, 9 and 10) within vertebrates could be associated mostly to the rounds of whole genome duplication experienced by this group early on its evolutionary history.

Animal AQPs can be divided into three major groups (Fig. 4). The first group includes classical AQP4, 1, 0, 5, 6 and 2, which are found in vertebrates, and are likely the result of whole genome duplications at early stages of their evolutionary history, together with more recent tandem gene duplication events involving AQP2, 5 and 6 as deduced from their close localization in the chromosome. Non-vertebrate AQPs are recovered at the base of a clade including the above-mentioned vertebrate AQP paralogs. A second group includes AQP8 orthologs, which are found from nematodes to mammals. Yet, it cannot be discarded that both the groups of arthropod AQPs branching off after and the cnidarian AQP branching off before the AQP8s in the phylogeny could actually belong to the same group of orthology. The most basal group in the phylogeny of animal AQPs corresponds to AQP11 and AQP12, which are only found in ray-finned fishes and sarcopterygians (Fig. 4). Therefore, it is likely that these paralogs arose as a result of one of the whole genome duplications experienced by vertebrates. Both AQP11 and 12 are intracellular AQPs that have diverged largely in sequence and function [17,25], what is reflected in their extremely long branches and explains their basal position in the phylogeny (i.e., a LBA artifact).

2.4. Plant MIPs

MIPs are most diverse in plants (Figs. 5 and 6; [4,14]). There are five paralog groups in seed plants (PIPs, TIPs, NIPs, SIPs, and XIPs), one more in spike mosses (HIP) [14] and two more in mosses (HIP and GIP) [15]. The actual number of paralogs in ferns is unknown, although here we show that *Equisetum* has at least PIPs, TIPs, and NIPs (see Appendix A). Green algae have five subfamilies (MIP A–E) not found in any other plant lineage, and two subfamilies (PIPs and GIPs) that might have been acquired through HGT [13]. PIPs can be subdivided into two highly conserved paralog groups, PIP1 and PIP2 (Fig. 5A), which can be traced back to mosses, suggesting that they were already present in the

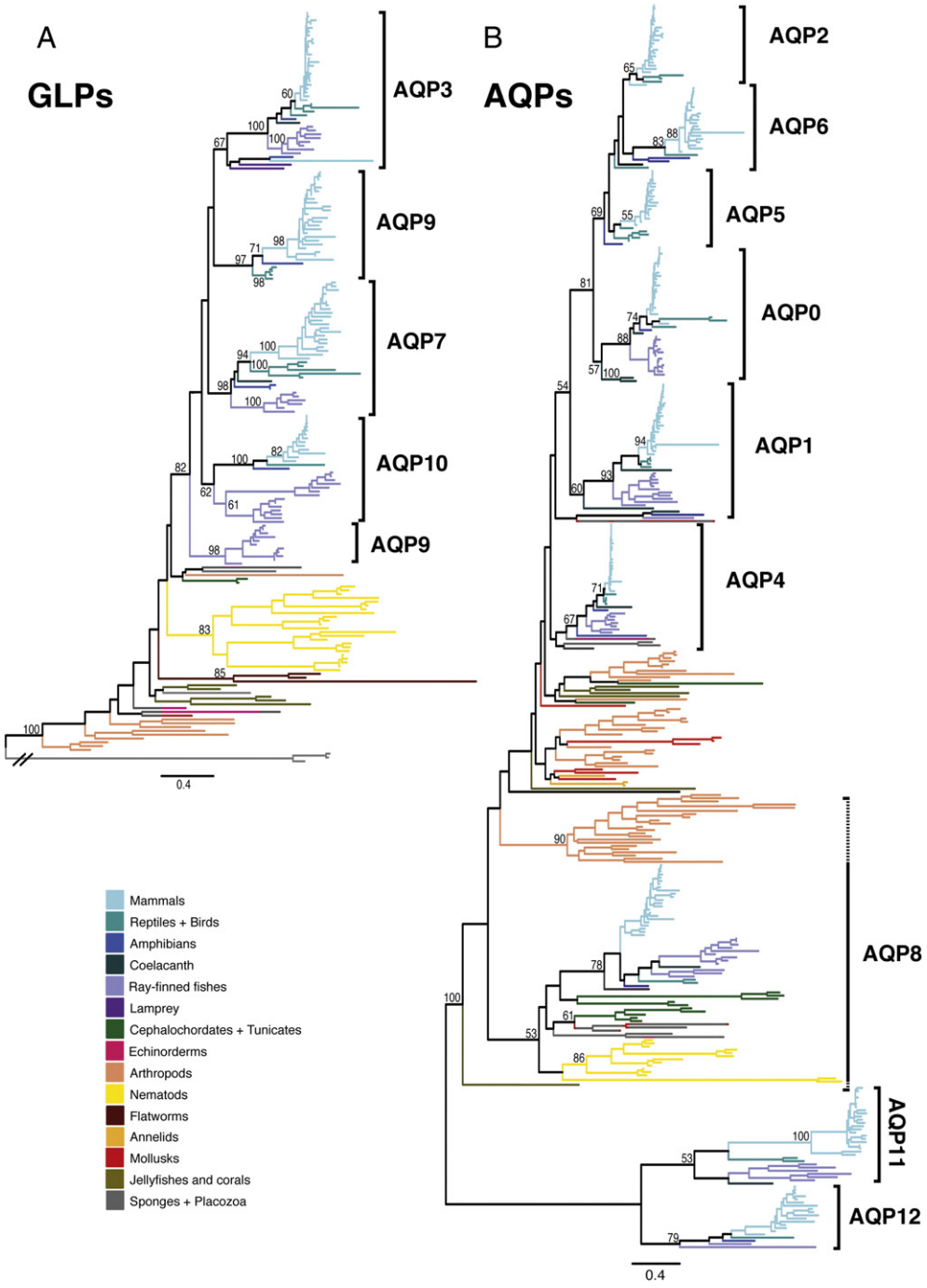

Fig. 4. Phylogeny of animal GLPs (A) and AQPs (B).

ancestor of Embryophyta [15]. Within each of these two main paralog groups, there have been family-specific expansions, as exemplified by grasses (Poaceae), crucifers (Brassicaceae) or legumes (Fabaceae).

In TIPs, up to five groups of paralogy can be distinguished in seed plants (Spermatophyta) based on the phylogenetic tree (Fig. 5B): TIP1 to 5. TIP1 and 3 and TIP2 and 5 are sister groups, respectively. Each of these groups shows internal family-specific gene duplications. Independent expansions of TIPs occurred both in *Selaginella* (Lycopodiophyta) and *Physcomitrella* (Bryophyta). XIPs and HIPs are recovered as close relatives of TIPs in the global MIP phylogeny (Fig. 1). The former were first identified in *Physcomitrella* and dicots [15], and later also found in *Selaginella* [14]. There are also XIPs reported out of plants in Dictyosteliida and Fungi [15,47], and the sparse and odd phylogenetic distribution of XIPs among lineages suggests that either multiple independent losses or HGT events need to be invoked to explain their evolutionary history and current taxonomic distribution. HIPs are only present in mosses and spike mosses, and therefore they were lost between the ancestor of vascular and seed plants (depending on whether they are found or not in ferns).

SIPs are intracellular MIPs [49] characterized by extremely long branches, and were originally found in seed plants [50]. Afterwards, they were also described in spike mosses [14] and mosses [15]. Moreover, a subfamily of algal MIPs (MIPC) present in *Ostreococcus* and *Micromonas* was recovered as putative sister group of SIPs [13]. However, MIPC was discarded as potential ortholog of SIPs due to its very different sequence, and its relative phylogenetic position was explained as a LBA artifact [13]. Here, we report the presence of true SIP orthologs in several green algae including *Volvox* (gi

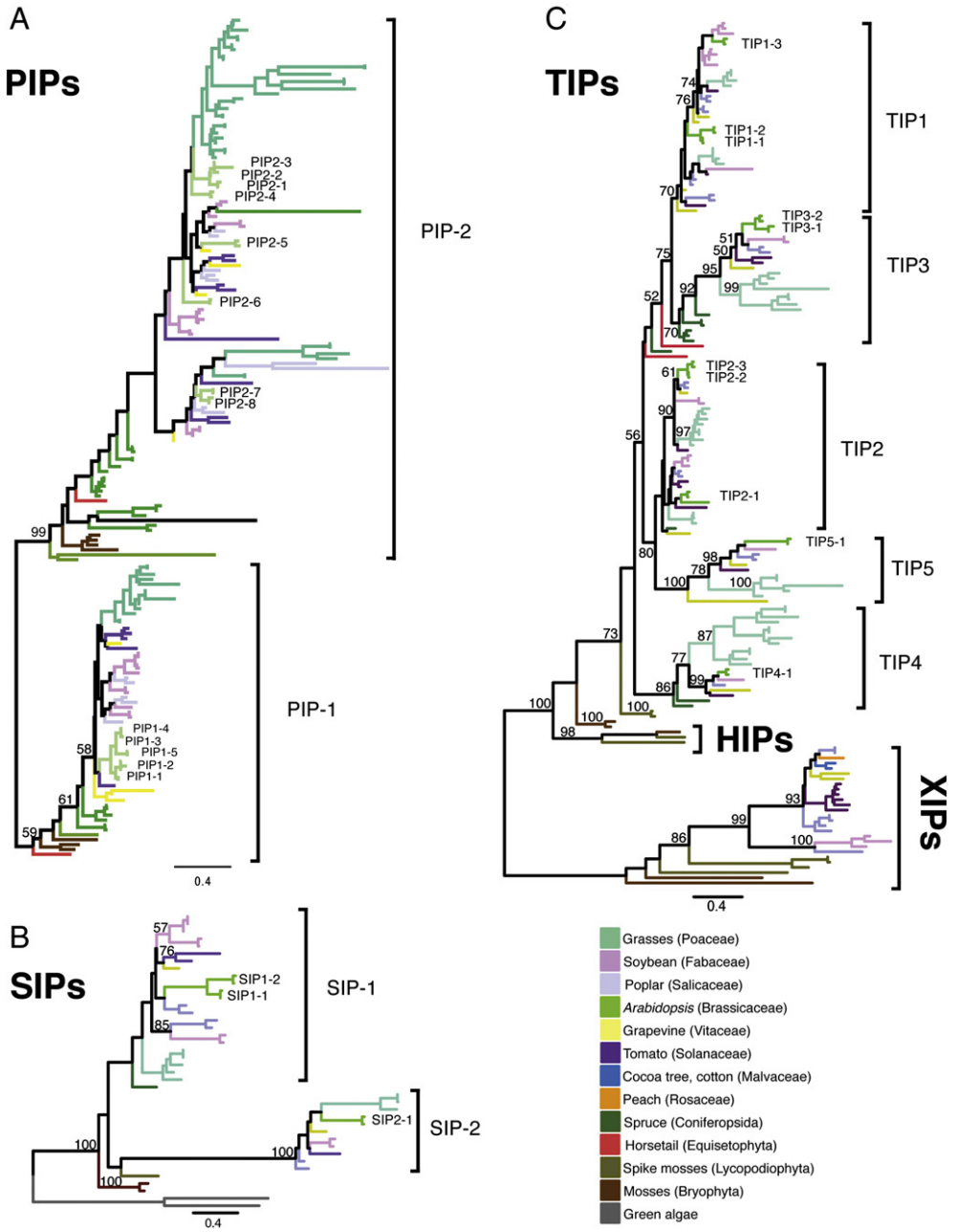

Fig. 5. Phylogeny of plant PIPs (A); TIPs, HIPs and XIPs (B); and SIPs (C).

302844971), *Chlorella* (gi 307102614), *Coccomyxa* (gi 384247825), and *Chlamydomonas* (gi 159463790), which share most of the conserved residues that define SIPs [50]. The reconstructed phylogeny shows that algal SIPs are basal to land plant SIPs (Fig. 5B). Given that Chlorophyta are assumed to be monophyletic [51], SIPs are either a true ancient family in plants with members independently lost in some Chlorophyta lineages or the result of HGT. Indeed, the presence of a GIP in a virus infecting *Chlorella* has been suggested as evidence of a possible vector for HGT [13]. Seed plant SIPs could be divided into two paralogs, SIP1 and SIP2. The latter showed a relatively long branch, and had been reported to not function as a water channel [25]. The long branch of this paralog introduced important biases in the phylogenetic analyses. Therefore, for reconstructing this particular tree, we had to retain all positions in the alignment in the phylogenetic analyses in order to maximize the number of phylogenetically informative characters between algal and seed plant SIPs (this was not needed when the phylogenetic analyses included only seed plant SIPs; not shown). According to the reconstructed phylogeny, the two SIP orthologs of mosses result from a more recent moss-specific duplication and outgroup seed plant SIP1 and SIP2 (Fig. 5B). Although spike mosses are recovered as sister group of SIP2, this relationship could be spurious due to its low support and high divergence. Hence, it is likely that *Selaginella* may also outgroup SIP1 and SIP2, and that the duplication that led to both paralogs occurred in the ancestor of seed plants. It would be important to search for SIPs in ferns and conifers, as genomes from these groups become complete, to fully understand the evolutionary history of this family.

The NIP phylogeny was rooted according to the global phylogeny. The recovered tree showed four main paralog groups (Fig. 6). Three of these (NIP1-3) were already well known, but only NIP3 was reported to have orthologs in *Selaginella* [14] and *Physcomitrella* [15]. The recovered tree shows that also NIP2 has orthologs in ferns, spike mosses and mosses (these sequences were previously assigned to another paralog group, NIP5, which according to our results is not valid [14,15]). A fourth

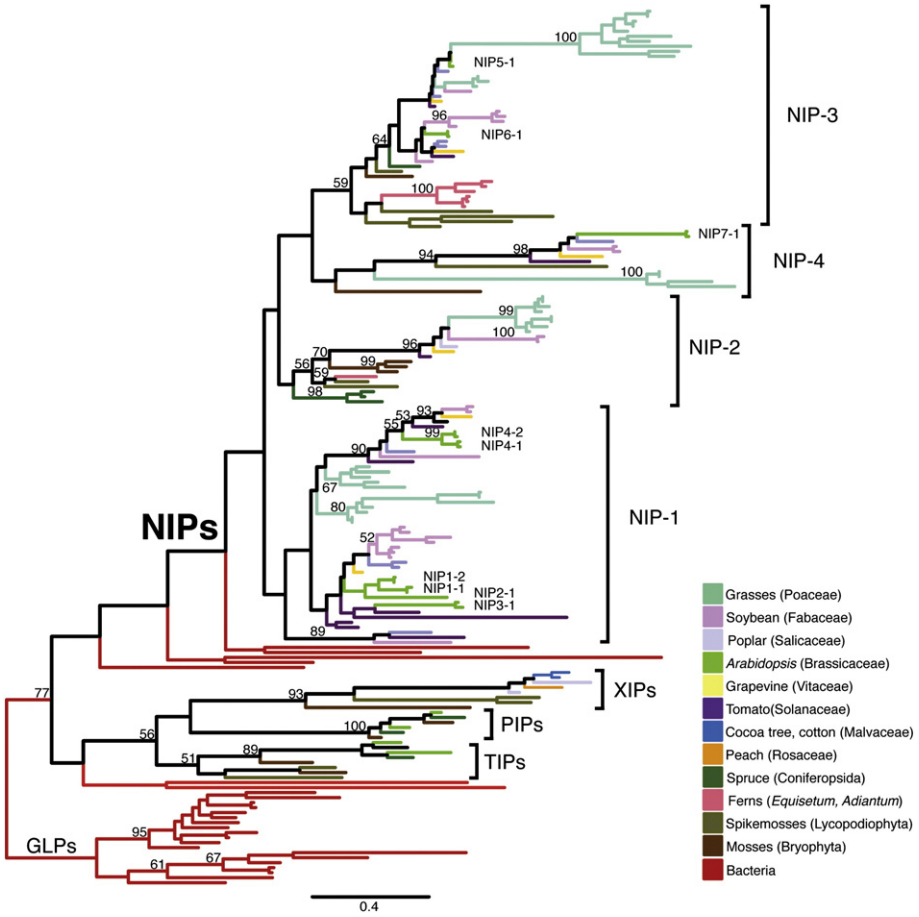

Fig. 6. Phylogeny of plant NIPs, together with bacterial AQPs and GLPs and a representation of other plant MIPs.

paralog group (NIP4) with relatively long branches is first described here. This paralog is present from mosses to flowering plants (including the gene named NIP7 of *Arabidopsis*; [23]). Therefore, only NIP1 is exclusive of seed plants, and its relative position in the phylogeny could be explained invoking secondary losses in *Selaginella* and *Physcomitrella* (it is not possible at present to conclude whether this paralog is also absent in *Equisetum*). In addition, the phylogeny shows groups of paralogy within NIP1 and NIP3 that likely arose through gene duplications in the ancestor of seed plants (Fig. 6). Interestingly, NIP2 is missing in some of the analyzed flowering plant species (e.g., *Arabidopsis*, cocoa tree, and cotton).

3. Structure of MIPs

Despite the ancient origin, widespread taxonomic distribution, and high sequence diversity of MIPs, comparative sequence analyses have shown that several amino acid residues key for their function as membrane channels are highly conserved [7,8,19,21]. To define selectively conserved and potentially important positions for the function in the MIP protein, we built a multiple sequence alignment of MIPs (670 sequences) in which highly divergent AQPs (e.g., SIPs or AQP12) and sequences with long insertions or deletions were removed. Next, we looked for residues conserved in at least 90% of the aligned proteins (Fig. 7). This filtering rendered the following key positions: (1) the two NPA boxes (we included the Ala of the first and second NPA motifs although they were conserved in 85.7% and 87.3%, respectively); (2) Asp8 (position numbers correspond to *Escherichia coli* AqpZ), which is in transmembrane helix 1 (H1); (3) Ser58 and Gly59 in loop B (i.e., close to the first NPA motif); (4) Gln88 and Gly91 in H3; (5) Asn182 (Gly in most MIPs) and Arg189 in loop E (before and after the second NPA, respectively); and (6) Pro212 and Gly215 in H6. Similar results were obtained with Consurf [52], which measures conservation in a phylogenetic framework (see Appendix A). Interestingly, conserved residues concentrated in the MIPs half region close to the cytoplasm. We further investigated conservation of these key positions in each of the main MIP subfamilies. In most subfamilies, these positions are conserved and exceptions are restricted exclusively to certain species without any particular phylogenetic pattern (see Appendix A). However, systematic deviations from the canon were evident in SIPs, AQP11s and AQP12s. In SIPs, Asp8 (*E. coli* AqpZ coordinates) is replaced by Glu; Ser57 is most frequently replaced by Gly or a basic residue (Arg or Lys); the Ala of the first NPA motif becomes highly variable; and most remarkably, Arg189 (one of the most conserved residues in MIPs) is replaced by different amino acids (Ser, Ala, or Asn, among others). In addition, the loop C has been reported to be considerably shorter in these proteins [49]. Moreover, our comparative analyses identified several residues that are different in other MIPs but highly conserved in SIPs (including the new algal SIPs reported here), and thus that could be considered shared derived in evolutionary terms i.e. synapomorphies: a Trp in H1, two Pro in H3 and H4, another Trp after loop B, and some other residues at the last helix and the C-terminus. Both AQP11 and 12 deviate from the canon at Ser57 (Asp in AQP12, variable in AQP11), at the first NPA (NPT in most AQP12, NPC in AQP11; [53]), and at Arg189 (replaced by Ala in both). The alignment between AQP11 and 12 was particularly problematic given the high sequence divergence, and proper site homology could not be reliably established for some regions. Nevertheless, some conserved residues between AQP11 and 12 could be distinguished, including two Cys residues, one at H5 and the other at the end of the loop E. Although AQP11/12 and SIPs might have somewhat similar functions (they are

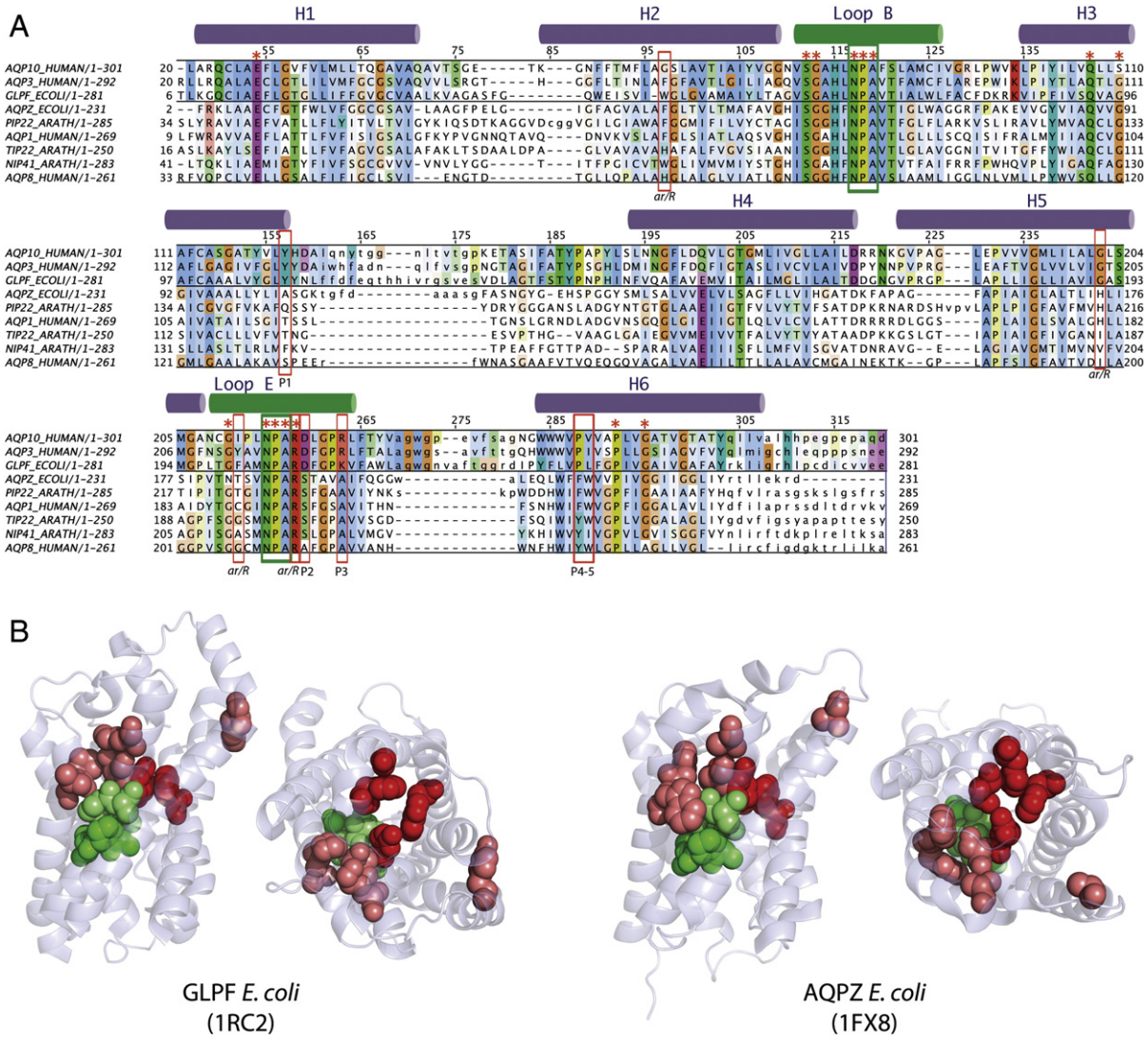

Fig. 7. Sequence and structural view of positions determining specificity in MIP proteins. (A) Multiple sequence alignment of selected MIPs. Residues forming the ar/R selectivity filter, reported to confer glycerol selectivity (P1–P5), or defined here as mostly conserved (red asterisks; >90% identity) are mapped onto the alignment. (B) NPA boxes (green), ar/R filter sites (red), and P1–P5 sites (pink) are mapped onto the three dimensional structures of *E. coli* GlpF (PDB ID: 1RC2) and AqpZ (PDB ID: 1FX8).

basic proteins and are located in the endoplasmic reticulum [49]), none of the specifically conserved residues in AQP11 and 12 is shared with SIPs, suggesting that the recovered node in the phylogenetic tree connecting these vertebrate and plant subfamilies probably reflects a LBA artifact rather than a true relationship of orthology, as previously proposed [18]. Solving the 3D structure of these subfamilies would probably help to a better understanding of their evolutionary history.

Structural analyses showed that the ar/R filter (residues Phe43, His174, Thr183, and Arg189 in *E. coli* AqpZ) mainly dictates substrate selectivity of MIPs by delimiting the narrowest part of the pore [12,44,45]. Furthermore, certain sites were described as differentially conserved in AQPs and GLPs, and thus potentially related to glycerol selectivity. These include the sites commonly known as P1–P5 ([21]; residues Ala103, Ser190, Ala194, Phe208, and Trp209 in *E. coli* AqpZ). However, mutagenesis of these sites has shown restricted alteration of MIPs selectivity [54,55], indicating that more complex evolutionary patterns have to be considered. In this regard, Statistical Coupling Analysis revealed more than ten differentially coevolving pairs of residues in AQPs and GLPs [56], some of which had experimental support. Interestingly, while most conserved MIP residues are located in the cytoplasmic half of MIP proteins, sites responsible for selectivity (including the ar/R filter) mostly concentrate on the extracellular half. Since MIP proteins are made of two structurally identical hemipores (hourglass model; [57]) arranged in opposite directions, such evolutionary pattern most probably reflects functional differences between the two hemipores rather than evolutionary constraints associated to the common (inverted) structure.

Plant NIPs and some archeal AQPs (AqpMs) represent independent cases of natural mutagenesis towards glycerol selectivity. In NIPs, the first residue of the ar/R filter mutated to Trp, and the P1 and P5 sites were replaced by aromatic and small hydrophobic amino acids, respectively, as observed in typical GLPs (Fig. 7; [16,26,44]). The shared derived residues of NIPs include a Pro (usually followed by a Trp) at the end of loop B, a Thr and a Pro before H4, and a Tyr and an Arg at H6. Regarding AqpMs, the archaeal aquaglyceroporins, mutations concentrate in P2 (Ser in AQPs, Asp in GLPs, Thr in AqpM), P3 (Ala in AQPs, basic in GLPs, Tyr in AqpM), and P5 (Trp in AQPs, small hydrophobic in GLPs, and Tyr in AqpM), although they show also conserved residues at the end of loop E (charged amino acid) or close to H6 (Pro), among others. The evolutionary patterns observed in NIPs and archaeal aquaglyceroporins show us that different solutions to confer glycerol selectivity are possible [16,44,45], but at the same time that some key positions, like Trp in P5, are recurrently mutated (see Appendix A).

XIPs have been found in land plants (although lost in monocots), fungi and *Dictyostelium* [15,47], and it was controversial whether the XIPs of *Dictyostelium* were true orthologs or their grouping with plant

and fungal XIPs was a LBA artifact [15]. Indeed, plant XIPs have a larger loop after helix 5, whereas *Dictyostelium* XIPs have a shorter helix 5, and alignment in this region is problematic. Our phylogenetic analyses including new potential XIPs from unicellular eukaryotes, all from the order Dictyosteliida, further supported the true orthology of these genes (Fig. 3). Remarkably, shared derived residues conserved among XIPs of plants, fungi and Dictyosteliida could be distinguished in the alignments. These included an aromatic residue (usually Trp) in TM1, a conserved Cys in loop C, a small polar (Ser/Thr) residue before loop B, and most notably, a Cys just after the second NPA motif (see Appendix A). Overall, these residues might be responsible of functional specificities of XIP proteins and further support the true orthology of the group as recovered in the phylogenetic tree.

To further explore whether deep orthologies between plant TIPs and animal AQP8, as well as between plant PIPs and classic animal AQP 4, 1, 0, 2, 6 and 5 may exist (Fig. 1), we searched for conserved sites that could be shared by the proposed groups of orthology. Three sites added some clues to this phylogenetic question: Phe43 (in AqpZ), which is part of the ar/R filter, is found as His in AQP8 and TIPs, and as Phe in PIPs and classic animal AQPs; His174 (in AqpZ), also belonging to the ar/R filter, is replaced by Ile in TIPs and AQP8s and remains as His in PIPs and classic animal AQPs; and Ile178 (in AqpZ) is found as Gly in TIPs and AQP8, and Ile in PIPs and classic animal AQPs. Interestingly, these three sites are very close in the 3D structure (5.3–7.8 Å), and might constitute a molecular synapomorphy of AQP8 and TIPs, in support of a most recent common ancestor for both subfamilies. Regarding the potential close relationship of PIPs and classic animal AQPs, the conserved sites described above represent the ancestral state of the whole AQP family, and hence provide no support for a deep orthology relationship of both subfamilies.

4. Functions of MIPs in an evolutionary context: bacteria, flowering plants and vertebrates

Since their discovery, isolation of different MIPs has been followed by functional characterization through e.g., in vitro functional assays of substrate transport in *Xenopus* oocytes [5]. Once MIP proteins were characterized, their function and tissue localization were used to infer their potential physiological roles [58,59]. At present, many MIPs are identified directly from genome projects through comparative genomic techniques, and hints of their function could be inferred based on microarray and/or RNA-seq data. Our aim here is not to review what is known about MIP functions (see e.g., [4,58]) but rather to put MIP genomic and gene expression data into an evolutionary context to better understand the diversification of MIPs across different phyla.

In bacteria and archaea, MIPs have not diversified as much as in eukaryotes, providing a simpler functional (ancestral) scenario: one paralog is in charge of water transport and the other of glycerol transport. As an exception to this rule, in thermophilic archaea and bacteria MIPs are absent, suggesting that their function might be negligible at high temperatures or that alternative mechanisms are used for water and glycerol transport. Moreover, other archaea have evolved AQPs (AqpM) that are able to transport glycerol, presumably with adaptive advantages (Fig. 2; [46]). We searched for gene-neighborhood conservation around AQPs and GLPs in bacterial and archaeal genomes using the String web-server [60] as a mean to identify potential functional relationships with other genes [61]. GLPs were frequently found close to glycerol kinase, glycerol-3-phosphate dehydrogenase, and glycerophosphoryl diester phosphodiesterase genes, both in bacteria and in the few GLP-containing archaea. This suggests functional relationships among these genes, and furthermore microarray data supports their coexpression. AQPs do not show such an obvious pattern of neighborhood conservation. In some cases, AQPs from diverse lineages (e.g. *Rhizobium*, *Methylobacterium*, *Polaromonas*, *Methylovorus*, *Gloeobacter*, *Deinococcus*, and archaeal *Nitrosopumilus*) are found close to a protein tyrosine phosphatase (COG0394) and a transcriptional regulator of the ArsR family. Archeal AqpM from Euryarchaeota is frequently found close to two genes coding for hypothetical proteins that posses archaeal-specific Pfam domains of unknown function: DUF2193 and DUF2180 (probably a Zn-finger). Conservation of gene context between non-related lineages of bacteria and archaea possibly reflects HGT, further explaining the mixture of archaeal and bacterial AQPs in the phylogeny (Fig. 2).

In contrast to the pattern described in bacteria and archaea, MIPs have diversified largely in eukaryotes by lineage-independent gene duplications. This recurrent pattern of MIP expansions supports the adaptive value of these proteins and points towards functional diversification of paralogs. However, some degree of functional redundancy among paralogs is also suspected [62], indicating that a great number of copies might have also a gene-dosage role. To further understand the relationship between expansion by gene duplication and functional divergence, we analyzed the patterns of tissue expression in humans and *Arabidopsis* in an evolutionary context, as well as the strength of purifying selection acting on each paralog (Figs. 8 and 9). In humans, we used RNA-seq data from the comprehensive Illumina Human Body Map, which comprises up to 16 tissues: adipose, adrenal, blood, brain, breast, colon, heart, kidney, liver, lung, lymph nodes, ovary, prostate, skeletal muscle, testes, and thyroid (E-MTAB-513 in ArrayExpress; [63]). Reads aligned against the human GRCh37/hg19 assembly were retrieved from Ensembl [64], and uniquely mapping reads against exons of principal gene isoforms (as identified by Appris; [65]) were counted using bedtools [66]. Subsequently, the corresponding RPKM (reads per kilobase of exon model per million mapped reads) values were calculated as a measure of the expression level. The expression heatmap (Fig. 8) shows that AQP3 and AQP1 are the most ubiquitously expressed MIPs within the GLP and AQP lineages, respectively. Consistently, AQP3 and AQP1 are under strong purifying selection. AQP7, AQP11 and AQP4 are also broadly expressed, although at lower levels (Fig 8). Despite its broad pattern of expression, AQP7 is evolving under weak purifying selection, which, interestingly, could be related to the presence of several AQP7 pseudogenes (at least four in human, all along chromosome 9). AQP 9 and 10 are also evolving under weak purifying selection, AQP10 being poorly expressed, and AQP9 being mostly expressed in blood (Fig. 8). Within AQPs, AQP2, 5 and 6 are located in tandem in chromosome 12 (q13.12), and have highly specific patterns of expression, with AQP2 and AQP6 predominating in kidney (Fig 8) and AQP5 in the testes, lung and thyroid. AQP4 is clearly the most expressed paralog in brain (Fig 8). These patterns are in agreement with what was known about AQP expression in mammals and chicken [67]. AQP8, AQP0, and AQP12A and B showed no expression in this collection of tissues, although AQP8 is known to express in the digestive tract [68], AQP0 in eye lens [69], and AQP12 intracellularly in pancreas [70]. The data obtained by RNA-seq were in agreement with data from a barcodized collection of microarray experiments obtained from the Gene Expression Barcode (see Appendix A; [71]). The only exceptions were the AQP12 paralogs, which show ubiquitous expression in this microarray platform (HGU133plus2), although we noticed that the microarray probes assigned to AQP12 genes (1559575_a_at and 1554344_s_at) were too unspecific. Overall, these results indicate that subfuncionalization associated to tissue specificity has been the main driving force for the diversification of MIPs in vertebrates.

The diversity of vertebrate MIPs could be explained under the following evolutionary scenario: GLP and AQP lineages separated at the LUCA ancestor, the origin of some major groups might be related to deep orthologies (as discussed above for the putative orthology of AQP8 and plant TIPs), and most of the diversification could be explained by the rounds of whole genome duplication experienced by vertebrates early in their evolutionary history [19]. However, additional more recent gene duplications need to be invoked to fully explain observed MIP diversity in vertebrates. We identified several tandem duplications when locating MIP genes in the human genome, two comprising AQP5, 2 and 6 that trace back to the ancestor of tetrapods; one specific of the hominid lineage resulting in AQP12A and B; and one for AQP3 and 7. For these more recent gene duplications, sequence divergence is

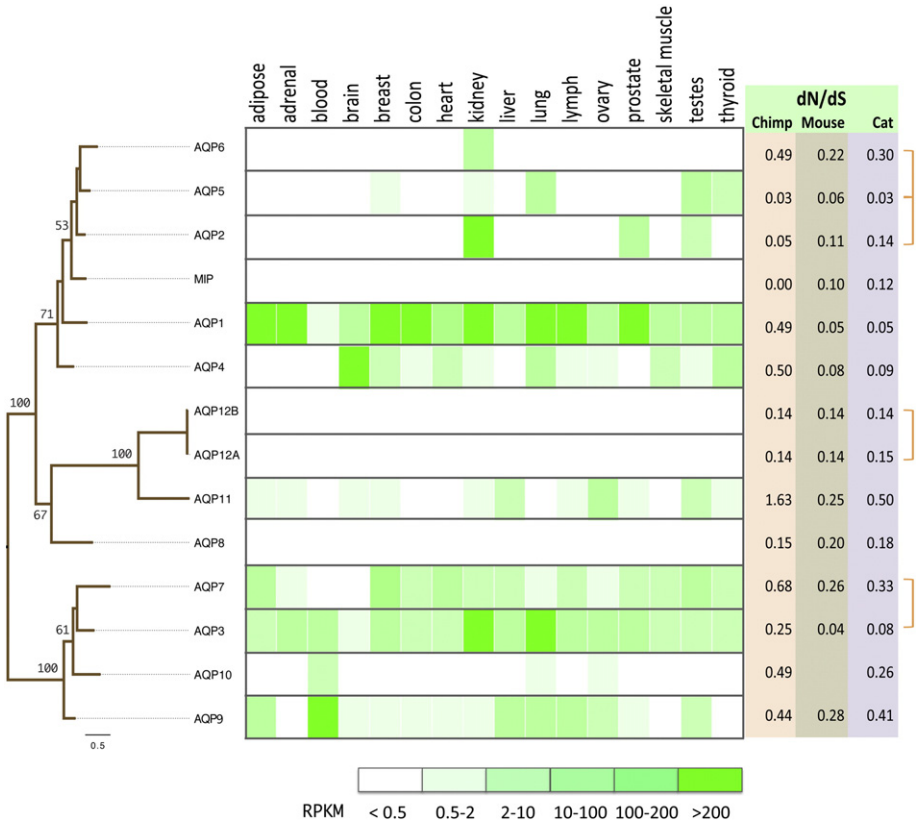

Fig. 8. Expansion, functional diversification and selective constraints of MIP proteins in humans. The figure displays the phylogeny of the 14 human paralogs with numbers above branches indicating bootstrap support (in percentage; 1000 pseudoreplicates). The heatmap shows gene expression levels for each paralog for 16 different tissues. Expression levels are measured as RPKM (reads per kilobase of exon model per million mapped reads). In addition, ratios of non-synonymous substitutions per non-synonymous site versus synonymous substitutions per synonymous site (dN/dS) are shown between human and chimp, mouse and cat orthologs, hence covering diverse times of divergence (chimp ~6.3 Mya, mouse ~91 Mya, and cat ~97.4 Mya according to [75]). Orange brackets to the right indicate tandem duplications.

not high (what may be explained by paralogous gene conversion), indicating that functional diversity may be partly acquired through changes at the regulatory level.

In *Arabidopsis*, as in other plants, MIPs have diversified extremely, especially if we consider that in this species there are no members of the GLP lineage, and that XIP and HIP subfamilies were lost in *Arabidopsis* and seed plants, respectively. Moreover, the ar/R selective filters of *Arabidopsis* MIPs show several ar/R combinations different from those found in orthodox AQPs and GLPs, supporting that plant MIPs have additional transport functions besides water and glycerol [45]. We obtained mean-normalized gene expression intensities from a microarray-based gene expression map of *Arabidopsis thaliana* development [72] available at AtGenExpress. Some general patterns were evident (Fig. 9): first, plant AQPs are more abundantly expressed in roots and seeds (as could be expected given their water-transport function), and second, in the NIP, SIP and TIP subfamilies, there are at least one paralog specifically expressed in roots and one in seeds. In contrast, no paralog of the PIP subfamily showed clear seed-specific expression. Interestingly, PIPs are the only Plant MIPs that exclusively transport water and no other solute, and PIP genes are overexpressed after drought stress in leaves [73]. In contrast TIPs transport other substrates than water including urea, ammonia, and hydrogen peroxide, and TIP genes become underexpressed after drought stress [73]. In contrast to human AQPs, no clear differences in the strength of purifying selection could be appreciated in plants MIPs. However in many cases it was not possible to calculate dN/dS because one-to-one orthology relationships were usually broken due to the high frequency of gene duplications and losses. Two sister group relationships showed clearly divergent expression patterns (Fig. 9). NIP11 and 12, which probably duplicated recently at the ancestor of Brassicaceae, were basically expressed in roots and seeds (the latter also in flowers), respectively. The same pattern occurred with SIP11 and SIP12. Interestingly, SIP21, which outgroups SIP11 and SIP12, has a pattern of expression very similar to SIP11, indicating that SIP12 could be the paralog under selection after gene duplication. However, considering that SIP1 and SIP2 duplication likely traces back to the ancestor of seed plants, prompts for cautious interpretation of this pattern. Less marked but still evident patterns of gene expression divergence are observed for the recent duplications involving TIP11–TIP12 and PIP25–PIP26 (Fig. 9). In TIPs, there are three broadly expressed paralogs (TIP21, TIP12 and TIP11), two that are seed-specific (TIP31, TIP13), three restricted to root (TIP23, TIP22, TIP41), and two specific of flowers (TIP13, TIP51). XIPs, which are absent in *Arabidopsis* and monocots, were reported to express widely in poplar with no particular tissue specificity [47].

Regarding the evolutionary events that led to the actual MIP diversity in flowering plants, the following scenario could be posed. While the principal subfamilies of plant AQPs arose by gene duplications at the ancestor of land plants (or possibly before in the case of TIPs, see above), NIPs were probably acquired through HGT from bacteria (Fig 6; [26]). Later on, many lineage-independent gene duplications occurred. For example, the NIP1 subfamily expanded independently in Brassicaceae and Poaceae. In contrast to vertebrates, scarce evidence of tandem duplications is found in plants, and whole-genome duplications and/or polyploidization might be in part responsible of plant MIPs expansion. As in animals, sequence divergence is not high between many paralogs, and functional diversity may occur through changes at the regulatory level [74].

5. Conclusions

MIPs are the focus of numerous studies due to their key function as water channels and their spectacular functional and structural

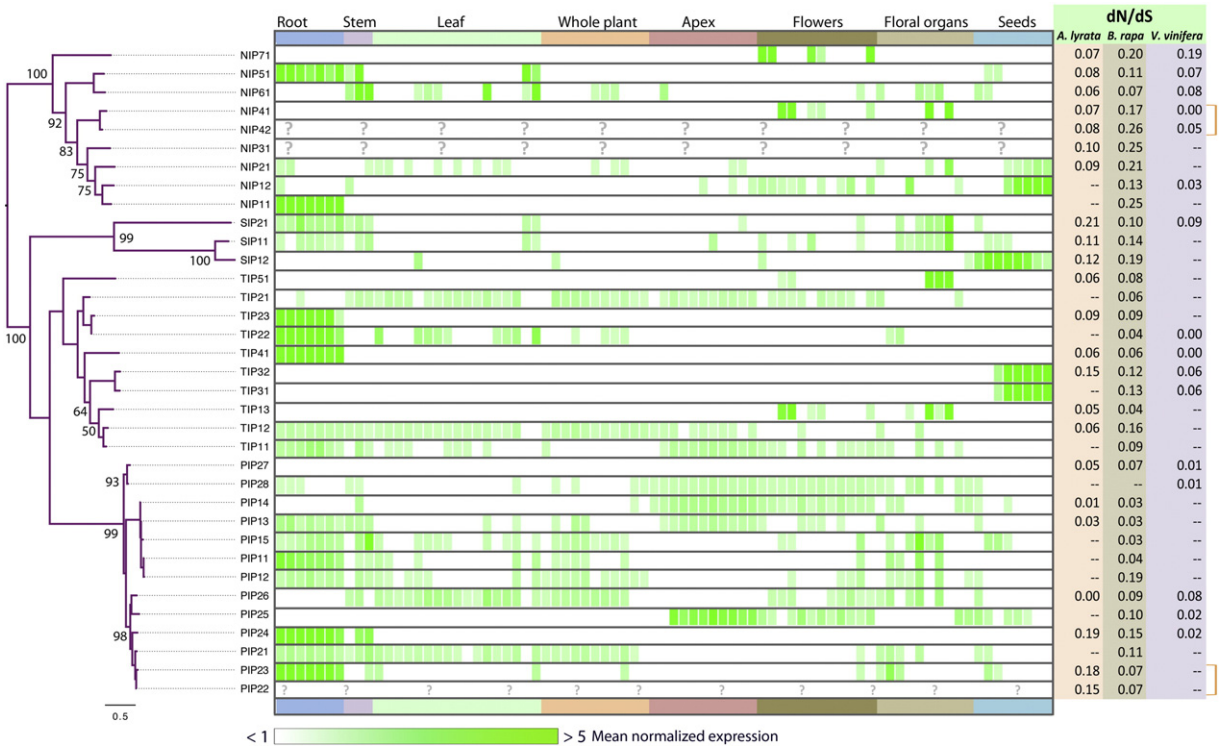

Fig. 9. Expansion, functional diversification and selective constraints of MIP proteins in the lineage to *Arabidopsis*. The figure shows the phylogeny of the 35 paralogs of *A. thaliana*, with numbers above branches indicating bootstrap support (in percentage; 1000 pseudoreplicates). The heatmap shows normalized gene expression intensities of eight different organs. In addition, ratios of non-synonymous substitutions per non-synonymous site versus synonymous substitutions per synonymous site (dN/dS) are shown between *A. thaliana* and *Arabidopsis lyrata*, *Brassica rapa* and *Vitis vinifera* orthologs, hence covering diverse times of divergence (*A. lyrata* ~5.4 Mya, *B. rapa* ~16.4 Mya, and *V. vinifera* ~113.3 Mya; [75]). Orange brackets to the right indicate tandem duplications.

diversification. In addition, as shown here, MIPs constitute a model system for studying molecular evolution but also embrace a great challenge for phylogenetic reconstruction. Although there are several conserved amino acids, the most prominent ones the two NPA boxes, which clearly identify members of the family, the extraordinary sequence divergence between subfamilies is enough to make phylogenetic inference of deep orthologies difficult due to the lack of enough shared derived residues. In addition to gene duplication coupled with sequence and functional divergence, diversity of the family, at least in plants, may have been acquired partly through several HGT events. Moreover, instances of functional convergence [2] often hinder phylogenetic relationships. In spite of all these challenges, it is possible to infer major evolutionary and genetic processes involved in the generation of diversity within the family. Our evolutionary analyses demonstrate a clear ancient origin of the XIP subfamily (at least tracing back to the common ancestor of Dyctiostellida + Ophistokonta + Plants), and suggest a potential deep orthology uniting animal AQP8 with HIPs, XIPs, and TIPs based on the reconstructed phylogeny, the existence of putative shared derived characters, and the nearly complete distribution of the group in eukaryotes. Moreover, our results suggest that seed plants co-opted NIPs to transport glycerol and concurrently lost GLPs due to functional redundancy. On the contrary, we find no support for a close relationship between plant SIPs and vertebrate AQP11 and 12 beyond their extremely divergent sequences that are responsible of a likely LBA artifact in the phylogenetic reconstruction. Yet, we were able to show that true SIPs orthologs are present in algae. After analyzing more than 1700 MIP sequences, two main evolutionary patterns are derived from our results: one is the major ancient split between proteins that ancestrally transported water (AQPs) and glycerol (GLPs), and the other is the great and recurrent expansions of both subfamilies in eukaryotes that have led to the structurally and functionally highly diverse members of the family with multiple physiological roles. Groups of paralogy were initially well documented in flowering plants and vertebrates, but now that data are available from many non-model organisms, it is becoming evident that rapid evolutionary turnovers of gene duplications and gene losses were widespread supporting the role of these proteins in adaptive processes. Furthermore, our analyses of expression profiles within a phylogenetic context show that subfunctionalization (linked to changes in the regulation of expression) after gene duplication is the main evolutionary mechanism underlying diversification. Neofunctionalization (intracellular AQPs) and co-option (NIPs) have also contributed to further generate this diversity.

Acknowledgements

We thank Susanna Törnroth-Horsefield as Editor of the special issue on aquaporins, and two anonymous reviewers for their insightful comments on a previous version of the manuscript. Phylogenetic analyses were performed in the supercomputer Altamira at the Institute of Physics of Cantabria (IFCA-CSIC), member of the Spanish Supercomputing Network. We thank J. Marco and L. Cabellos for providing access to and maintaining Altamira. II acknowledges the support of the Alexander von Humboldt Foundation (Fellowship for Postdoctoral Researchers) during the final stages of this work. This study was partly funded by project CGL2010-18216 of the Ministerio de Economía y Competitividad of Spain to RZ.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.bbagen.2013.12.001.

References

[1] J. Carbrey, P. Agre, Discovery of the aquaporins and development of the field, in: E. Beitz (Ed.), Aquaporins, vol. 190, Springer, Berlin Heidelberg, 2009, pp. 3–28.

[2] R. Hove, M. Bhave, Plant aquaporins with non-aqua functions: deciphering the signature sequences, Plant Mol. Biol. 75 (2011) 413–430.
[3] E. Kruse, N. Uehlein, R. Kaldenhoff, The aquaporins, Genome Biol. 7 (2006) 206.
[4] C. Maurel, L. Verdoucq, D.-T. Luu, V. Santoni, Plant aquaporins: membrane channels with multiple integrated functions, Annu. Rev. Plant Biol. 59 (2008) 595–624.
[5] G.M. Preston, T.P. Carroll, W.B. Guggino, P. Agre, Appearance of water channels in xenopus oocytes expressing red cell CHIP28 protein, Science 256 (1992) 385–387.
[6] J.B. Heymann, A. Engel, Aquaporins: phylogeny, structure, and physiology of water channels, Physiology 14 (1999) 187–193.
[7] T. Gonen, T. Walz, The structure of aquaporins, Q. Rev. Biophys. 39 (2006) 361–396.
[8] J.B. Heymann, A. Engel, Structural clues in the sequences of the aquaporins, J. Mol. Biol. 295 (2000) 1039–1053.
[9] G.P. Bienert, D. Cavez, A. Besserer, M.C. Berny, D. Gilis, M. Rooman, F. Chaumont, A conserved cysteine residue is involved in disulfide bond formation between plant plasma membrane aquaporin monomers, Biochem. J. 445 (2012) 101–111.
[10] K. Murata, K. Mitsuoka, T. Hirai, T. Walz, P. Agre, J.B. Heymann, A. Engel, Y. Fujiyoshi, Structural determinants of water permeation through aquaporin-1, Nature 407 (2000) 599–605.
[11] H. Sui, B.-G. Han, J.K. Lee, P. Walian, B.K. Jap, Structural basis of water-specific transport through the AQP1 water channel, Nature 414 (2001) 872–878.
[12] D. Fu, A. Libson, L.J.W. Miercke, C. Weitzman, P. Nollert, J. Krucinski, R.M. Stroud, Structure of a glycerol-conducting channel and the basis for its selectivity, Science 290 (2000) 481–486.
[13] H. Anderberg, J. Danielson, U. Johanson, Algal MIPs, high diversity and conserved motifs, BMC Evol. Biol. 11 (2011) 110.
[14] H.I. Anderberg, P. Kjellbom, U. Johanson, Frontiers in Plant Science, Annotation of *Selaginella moellendorffii* major intrinsic proteins and the evolution of the protein family in terrestrial plants, Front. Plant Sci. (2012) 3.
[15] J. Danielson, U. Johanson, Unexpected complexity of the aquaporin gene family in the moss *Physcomitrella patens*, BMC Plant Biol. 8 (2008) 1–15.
[16] J.H. Danielson, U. Johanson, Phylogeny of major intrinsic proteins, in: T. Jahn, G. Bienert (Eds.), MIPs and Their Role in the Exchange of Metalloids, vol. 679, Springer, New York, 2010, pp. 19–31.
[17] K. Ishibashi, S. Kondo, S. Hara, Y. Morishita, The evolutionary aspects of aquaporin family, Am. J. Physiol. Regul. Integr. Comp. Physiol. 300 (2011) R566–R576.
[18] G. Soto, K. Alleva, G. Amodeo, J. Muschietti, N.D. Ayub, New insight into the evolution of aquaporins from flowering plants and vertebrates: orthologous identification and functional transfer is possible, Gene 503 (2012) 165–176.
[19] R. Zardoya, Phylogeny and evolution of the major intrinsic protein family, Biol. Cell. 97 (2005) 397–414.
[20] R. Zardoya, S. Villalba, A phylogenetic framework for the aquaporin family in eukaryotes, J. Mol. Evol. 52 (2001) 391–404.
[21] A. Froger, D. Thomas, C. Delamarche, B. Tallur, Prediction of functional residues in water channels and related proteins, Protein Sci. 7 (1998) 1458–1468.
[22] J.S. Hub, B.L. de Groot, Mechanism of selectivity in aquaporins and aquaglyceroporins, Proc. Natl. Acad. Sci. 105 (2008) 1198–1203.
[23] U. Johanson, M. Karlsson, I. Johansson, S. Gustavsson, S. Sjövall, L. Fraysse, A.R. Weig, P. Kjellbom, The complete set of genes encoding major intrinsic proteins in *Arabidopsis* provides a framework for a new nomenclature for major intrinsic proteins in plants, Plant Physiol. 126 (2001) 1358–1369.
[24] F. Quigley, J. Rosenberg, Y. Shachar-Hill, H. Bohnert, From genome to function: the *Arabidopsis* aquaporins, Genome Biol. (2001) 3(research0001.0001-research0001.0017).
[25] K. Ishibashi, Aquaporin subfamily with unusual NPA boxes, Biochim. Biophys. Acta Biomembranes 1758 (2006) 989–993.
[26] R. Zardoya, X. Ding, Y. Kitagawa, M.J. Chrispeels, Origin of plant glycerol transporters by horizontal gene transfer and functional recruitment, Proc. Natl. Acad. Sci. 99 (2002) 14893–14896.
[27] R. Ekblom, J. Galindo, Applications of next generation sequencing in molecular ecology of non-model organisms, Heredity 107 (2011) 1–15.
[28] Z. Wang, M. Gerstein, M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics, Nat. Rev. Genet. 10 (2009) 57–63.
[29] D. Brawand, M. Soumillon, A. Necsulea, P. Julien, G. Csardi, P. Harrigan, M. Weier, A. Liechti, A. Aximu-Petri, M. Kircher, F.W. Albert, U. Zeller, P. Khaitovich, F. Grutzner, S. Bergmann, R. Nielsen, S. Paabo, H. Kaessmann, The evolution of gene expression levels in mammalian organs, Nature 478 (2011) 343–348.
[30] A. Bansal, R. Sankararamakrishnan, Homology modeling of major intrinsic proteins in rice, maize and *Arabidopsis*: comparative analysis of transmembrane helix association and aromatic/arginine selectivity filters, BMC Struct. Biol. 7 (2007) 27.
[31] S. Gustavsson, A.-S. Lebrun, K. Nordén, F. Chaumont, U. Johanson, A novel plant major intrinsic protein in *Physcomitrella patens* most similar to bacterial glycerol channels, Plant Physiol. 139 (2005) 287–295.
[32] A. Tingaud-Sequeira, M. Calusinska, R. Finn, F. Chauvigne, J. Lozano, J. Cerda, The zebrafish genome encodes the largest vertebrate repertoire of functional aquaporins with dual paralogy and substrate specificities similar to mammals, BMC Evol. Biol. 10 (2010) 38.
[33] Uniprot Consortium, Update on activities at the Universal Protein Resource (UniProt) in 2013, Nucleic Acids Res. 41 (2013) D43–D47.
[34] P. Flicek, I. Ahmed, M.R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gil, C. García-Girón, L. Gordon, T. Hourlier, S. Hunt, T. Juettemann, A.K. Kähäri, S. Keenan, M. Komorowska, E. Kulesha, I. Longden, T. Maurel, W.M. McLaren, M. Muffato, R. Nag, B. Overduin, M. Pignatelli, B. Pritchard, E. Pritchard, H.S. Riat, G.R.S. Ritchie, M. Ruffier, M. Schuster, D. Sheppard, D. Sobral, K. Taylor, A. Thormann, S. Trevanion, S. White, S.P. Wilder, B.L. Aken, E. Birney, F. Cunningham, I. Dunham, J. Harrow, J. Herrero, T.J.P. Hubbard, N. Johnson, R. Kinsella, A. Parker, G. Spudich, A. Yates, A. Zadissa, S.M.J. Searle, Ensembl 2013, Nucleic Acids Res. 41 (2013) D48–D55.
[35] R.D. Finn, J. Tate, J. Mistry, P.C. Coggill, S.J. Sammut, H.-R. Hotz, G. Ceric, K. Forslund, S.R. Eddy, E.L.L. Sonnhammer, A. Bateman, The Pfam protein families database, Nucleic Acids Res. 36 (2008) D281–D288.
[36] S. Hunter, P. Jones, A. Mitchell, R. Apweiler, T.K. Attwood, A. Bateman, T. Bernard, D. Binns, P. Bork, S. Burge, E. de Castro, P. Coggill, M. Corbett, U. Das, L. Daugherty, L. Duquenne, R.D. Finn, M. Fraser, J. Gough, D. Haft, N. Hulo, D. Kahn, E. Kelly, I. Letunic, D. Lonsdale, R. Lopez, M. Madera, J. Maslen, C. McAnulla, J. McDowall, C. McMenamin, H. Mi, P. Mutowo-Muellenet, N. Mulder, D. Natale, C. Orengo, S. Pesseat, M. Punta, A.F. Quinn, C. Rivoire, A. Sangrador-Vegas, J.D. Selengut, C.J.A. Sigrist, M. Scheremetjew, J. Tate, M. Thimmajanarthanan, P.D. Thomas, C.H. Wu, C. Yeats, S.-Y. Yong, InterPro in 2011: new developments in the family and domain prediction database, Nucleic Acids Res. 40 (2012) D306–D312.
[37] K. Katoh, D.M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability, Mol. Biol. Evol. 30 (2013) 772–780.
[38] S. Capella-Gutiérrez, J.M. Silla-Martínez, T. Gabaldón, trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses, Bioinformatics 25 (2009) 1972–1973.
[39] A. Stamatakis, RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models, Bioinformatics 22 (2006) 2688–2690.
[40] F. Abascal, R. Zardoya, D. Posada, ProtTest: selection of best-fit models of protein evolution, Bioinformatics 21 (2005) 2104–2105.
[41] D. Darriba, G.L. Taboada, R. Doallo, D. Posada, ProtTest 3: fast selection of best-fit models of protein evolution, Bioinformatics 27 (2011) 1164–1165.
[42] S. Guindon, O. Gascuel, A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood, Syst. Biol. 52 (2003) 696–704.
[43] J. Felsenstein, Confidence limits on phylogenies: an approach using the bootstrap, Evolution 39 (1985) 783–791.
[44] I.S. Wallace, W.-G. Choi, D.M. Roberts, The structure, function and regulation of the nodulin 26-like intrinsic protein family of plant aquaglyceroporins, Biochim. Biophys. Acta Biomembranes 1758 (2006) 1165–1175.
[45] I.S. Wallace, D.M. Roberts, Homology modeling of representative subfamilies of *Arabidopsis* major intrinsic proteins. classification based on the aromatic/arginine selectivity filter, Plant Physiol. 135 (2004) 1059–1068.
[46] D. Kozono, X. Ding, I. Iwasaki, X. Meng, Y. Kamagata, P. Agre, Y. Kitagawa, Functional expression and characterization of an archaeal aquaporin: AqpM from methanothermobacter marburgensis, J. Biol. Chem. 278 (2003) 10649–10656.
[47] A. Gupta, R. Sankararamakrishnan, Genome-wide analysis of major intrinsic proteins in the tree plant *Populus trichocarpa*: characterization of XIP subfamily of aquaporins from evolutionary perspective, BMC Plant Biol. 9 (2009) 134.
[48] S. Dietz, J. von Bülow, E. Beitz, U. Nehls, The aquaporin gene family of the ectomycorrhizal fungus *Laccaria bicolor*: lessons for symbiotic functions, New Phytol. 190 (2011) 927–940.
[49] M. Maeshima, F. Ishikawa, ER membrane aquaporins in plants, Pflugers Arch.—Eur. J. Physiol. 456 (2008) 709–716.
[50] U. Johanson, S. Gustavsson, A new subfamily of major intrinsic proteins in plants, Mol. Biol. Evol. 19 (2002) 456–461.
[51] F. Leliaert, D.R. Smith, H. Moreau, M.D. Herron, H. Verbruggen, C.F. Delwiche, O. De Clerck, Phylogeny and molecular evolution of the green algae, Crit. Rev. Plant Sci. 31 (2012) 1–46.
[52] G. Celniker, G. Nimrod, H. Ashkenazy, F. Glaser, E. Martz, I. Mayrose, T. Pupko, N. Ben-Tal, ConSurf: using evolutionary data to raise testable hypotheses about protein function, Israel J. Chem. 53 (2013) 199–206.
[53] M. Ikeda, A. Andoo, M. Shimono, N. Takamatsu, A. Taki, K. Muta, W. Matsushita, T. Uechi, T. Matsuzaki, N. Kenmochi, K. Takata, S. Sasaki, K. Ito, K. Ishibashi, The NPC motif of aquaporin-11, unlike the NPA motif of known aquaporins, is essential for full expression of molecular function, J. Biol. Chem. 286 (2011) 3342–3350.
[54] V. Lagrée, A. Froger, S. Deschamps, J.-F. Hubert, C. Delamarche, G. Bonnec, D. Thomas, J. Gouranton, I. Pellerin, Switch from an aquaporin to a glycerol channel by two amino acids substitution, J. Biol. Chem. 274 (1999) 6817–6819.
[55] D.F. Savage, J.D. O'Connell, L.J.W. Miercke, J. Finer-Moore, R.M. Stroud, Structural context shapes the aquaporin selectivity filter, Proc. Natl. Acad. Sci. 107 (2010) 17164–17169.
[56] X. Lin, T. Hong, Y. Mu, J. Torres, Identification of residues involved in water versus glycerol selectivity in aquaporins by differential residue pair co-evolution, Biochim. Biophys. Acta Biomembranes 1818 (2012) 907–914.
[57] J.S. Jung, G.M. Preston, B.L. Smith, W.B. Guggino, P. Agre, Molecular structure of the water channel through aquaporin CHIP. The hourglass model, J. Biol. Chem. 269 (1994) 14648–14654.
[58] A.S. Verkman, Mammalian aquaporins: diverse physiological roles and potential clinical significance, Expert Rev. Mol. Med. 10 (2008) e13.
[59] K. Forrest, M. Bhave, Major intrinsic proteins (MIPs) in plants: a complex gene family with major impacts on plant phenotype, Funct. Integr. Genomics 7 (2007) 263–289.
[60] A. Franceschini, D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, J. Lin, P. Minguez, P. Bork, C. von Mering, L.J. Jensen, STRING v9.1: protein–protein interaction networks, with increased coverage and integration, Nucleic Acids Res. 41 (2013) D808–D815.
[61] W.C. Lathe 3rd, B. Snel, P. Bork, Gene context conservation of a higher order than operons, Trends Biochem. Sci. 25 (2000) 474–479.
[62] D. Cohen, M.-B. Bogeat-Triboulot, S. Vialet-Chabrand, R. Merret, P.-E. Courty, S. Moretti, F. Bizet, A. Guilliot, I. Hummel, Developmental and environmental regulation of *Aquaporin* gene expression across *Populus* species: divergence or redundancy? PLoS ONE 8 (2013) e55506.
[63] G. Rustici, N. Kolesnikov, M. Brandizi, T. Burdett, M. Dylag, I. Emam, A. Farne, E. Hastings, J. Ison, M. Keays, N. Kurbatova, J. Malone, R. Mani, A. Mupo, R. Pedro Pereira, E. Pilicheva, J. Rung, A. Sharma, Y.A. Tang, T. Ternent, A. Tikhonov, D.

Welter, E. Williams, A. Brazma, H. Parkinson, U. Sarkans, ArrayExpress update—trends in database growth and links to data analysis tools, Nucleic Acids Res. 41 (2013) D987–D990.
[64] P. Flicek, M.R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gil, L. Gordon, M. Hendrix, T. Hourlier, N. Johnson, A.K. Kähäri, D. Keefe, S. Keenan, R. Kinsella, M. Komorowska, G. Koscielny, E. Kulesha, P. Larsson, I. Longden, W. McLaren, M. Muffato, B. Overduin, M. Pignatelli, B. Pritchard, H.S. Riat, G.R.S. Ritchie, M. Ruffier, M. Schuster, D. Sobral, Y.A. Tang, K. Taylor, S. Trevanion, J. Vandrovcova, S. White, M. Wilson, S.P. Wilder, B.L. Aken, E. Birney, F. Cunningham, I. Dunham, R. Durbin, X.M. Fernández-Suarez, J. Harrow, J. Herrero, T.J.P. Hubbard, A. Parker, G. Proctor, G. Spudich, J. Vogel, A. Yates, A. Zadissa, S.M.J. Searle, Ensembl 2012, Nucleic Acids Res. 40 (2012) D84–D90.
[65] J.M. Rodriguez, P. Maietta, I. Ezkurdia, A. Pietrelli, J.-J. Wesselink, G. Lopez, A. Valencia, M.L. Tress, APPRIS: annotation of principal and alternative splice isoforms, Nucleic Acids Res. 41 (2013) D110–D117.
[66] A.R. Quinlan, I.M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features, Bioinformatics 26 (2010) 841–842.
[67] R. Isokpehi, R. Rajnarayanan, C. Jeffries, T. Oyeleye, H. Cohly, Integrative sequence and tissue expression profiling of chicken and mammalian aquaporins, BMC Genomics 10 (2009) S7.
[68] G. Calamita, A. Mazzone, A. Bizzoca, A. Cavalier, G. Cassano, D. Thomas, M. Svelto, Expression and immunolocalization of the aquaporin-8 water channel in rat gastrointestinal tract, Eur. J. Cell Biol. 80 (2001) 711–719.
[69] A. Chepelinsky, Structural function of MIP/aquaporin 0 in the eye lens; genetic defects lead to congenital inherited cataracts, in: E. Beitz (Ed.), Aquaporins, vol. 190, Springer, Berlin Heidelberg, 2009, pp. 265–297.
[70] T. Itoh, T. Rai, M. Kuwahara, S.B.H. Ko, S. Uchida, S. Sasaki, K. Ishibashi, Identification of a novel aquaporin, AQP12, expressed in pancreatic acinar cells, Biochem. Biophys. Res. Commun. 330 (2005) 832–838.
[71] M.N. McCall, K. Uppal, H.A. Jaffee, M.J. Zilliox, R.A. Irizarry, The gene expression barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes, Nucleic Acids Res. 39 (2011) D1011–D1015.
[72] M. Schmid, T.S. Davison, S.R. Henz, U.J. Pape, M. Demar, M. Vingron, B. Scholkopf, D. Weigel, J.U. Lohmann, A gene expression map of *Arabidopsis thaliana* development, Nat. Genet. 37 (2005) 501–506.
[73] E. Alexandersson, J.Å.H. Danielson, J. Råde, V.K. Moparthi, M. Fontes, P. Kjellbom, U. Johanson, Transcriptional regulation of aquaporins in accessions of *Arabidopsis* in response to drought stress, Plant J. 61 (2010) 650–660.
[74] W. Park, B. Scheffler, P. Bauer, B.T. Campbell, Identification of the family of aquaporin genes and their expression in upland cotton (*Gossypium hirsutum* L.), BMC Plant Biol. 10 (2010) 142.
[75] S.B. Hedges, J. Dudley, S. Kumar, TimeTree: a public knowledge-base of divergence times among organisms, Bioinformatics 22 (2006) 2971–2972.
[76] R.L. Tatusov, E.V. Koonin, D.J. Lipman, A genomic perspective on protein families, Science 278 (1997) 631–637.
[77] S. Eddy, HMMER User's Guide. Biological sequence analysis using profile hidden Markov models, 2001.