





AQPX-cluster aquaporins and aquaglyceroporins are asymmetrically distributed in trypanosomes

Fiorella Carla Tesan ^{1,2}, Ramiro Lorenzo ³, Karina Alleva ^{1,2,4}✉ & Ana Romina Fox ^{3,4}✉

Major Intrinsic Proteins (MIPs) are membrane channels that permeate water and other small solutes. Some trypanosomatid MIPs mediate the uptake of antiparasitic compounds, placing them as potential drug targets. However, a thorough study of the diversity of these channels is still missing. Here we place trypanosomatid channels in the sequence-function space of the large MIP superfamily through a sequence similarity network. This analysis exposes that trypanosomatid aquaporins integrate a distant cluster from the currently defined MIP families, here named aquaporin X (AQPX). Our phylogenetic analyses reveal that trypanosomatid MIPs distribute exclusively between aquaglyceroporin (GLP) and AQPX, being the AQPX family expanded in the Metakinetoplastina common ancestor before the origin of the parasitic order Trypanosomatida. Synteny analysis shows how African trypanosomes specifically lost AQPXs, whereas American trypanosomes specifically lost GLPs. AQPXs diverge from already described MIPs on crucial residues. Together, our results expose the diversity of trypanosomatid MIPs and will aid further functional, structural, and physiological research needed to face the potentiality of the AQPXs as gateways for trypanocidal drugs.

¹Universidad de Buenos Aires, Facultad de Farmacia y Bioquímica, Departamento de Fisicomatemática, Cátedra de Física, Buenos Aires, Argentina.

²CONICET-Universidad de Buenos Aires, Instituto de Química y Fisicoquímica Biológicas (IQUIFIB), Buenos Aires, Argentina. ³Laboratorio de Farmacología, Centro de Investigación Veterinaria de Tandil (CIVETAN), (CONICET-CICPBA-UNCPBA) Facultad de Ciencias Veterinarias, Universidad Nacional del Centro de la Provincia de Buenos Aires, Tandil, Argentina. ⁴These authors contributed equally: Karina Alleva, Ana Romina Fox. ✉email: kalleva@ffyb.uba.ar; rfox@vet.unicen.edu.ar

The Trypanosomatida order of kinetoplastids (Euglenozoa, Discoba) gathers a vast diversity of parasitic protozoans that cause worldwide health problems infecting humans and livestock^{1–3}. Drugs preferential uptake and the presence of pathogen-specific enzymes determine the selectivity and toxicity of currently available drugs for disease control^{4,5}. In this regard, Major Intrinsic proteins (MIP) mediate the internalization of drugs that are the first choice against *Trypanosoma brucei* and *Leishmania* spp. (i.e., pentamidine and antimonial compounds, respectively)^{6,7}. Those findings support MIPs as potential drug targets against protozoan parasites⁸. Regardless, channels with very different pore properties build up the MIP superfamily, and comprehensive analysis of trypanosomatid MIPs diversity is still missing.

MIPs facilitate the diffusion of water and a variety of relatively small solutes through biological membranes⁹. Even with considerable sequence divergence, inside the MIP superfamily, its members preserve a typical three-dimensional structure, and they organize as tetramers having each monomer an individual transporting pore¹⁰. Two NPA (Asn-Pro-Ala) motifs in the middle part of the pore, regulate water conductance and operate as a barrier for the passage of inorganic cations (such as Na⁺ and K⁺)^{11,12}, and also participate in proton filtration^{13,14}. Still, protons are fully blocked at the selectivity filter^{11,12,15}, known as aromatic/Arginine (ar/R), which also executes a primary permeation role. The residues of this filter are related to the functional properties of the channel^{16,17} and, interestingly, play a central role in trypanosomatid drug uptake, i.e., their mutation may lead to drug resistance events¹⁸. Finally, five amino acid residues, designated as Froger positions, are involved in the discrimination between water or glycerol transport¹⁹.

The pioneer studies on MIPs diversity proposed that Eukarya isoforms derived from two bacterial channels: glycerol facilitators or aquaglyceroporins (GLP) and water channels or aquaporins (AQP)^{20,21}. Subsequent studies revealed an unexpected diversity of MIPs in the three domains of life and that first distinction AQP versus GLP remained insufficient to describe MIPs phylogeny. Consistently, the nomenclature of MIPs became more complex. Today, four clusters of prokaryotic MIPs have been described, named as grades to point to the polyphyletic nature of the superfamily (AqpM, AqpN, AqpZ, and GlpF)²². In Eukarya, there are up to seven recognized families of land plant MIPs: plasma membrane intrinsic protein (PIP), tonoplast intrinsic protein (TIP), Nodulin 26-like intrinsic protein (NIP), small basic intrinsic protein (SIP), X or uncharacterized intrinsic protein (XIP), hybrid intrinsic protein (HIP), and GlpF-like intrinsic protein (GIP)^{23–26}, while green algae have PIPs and GIPs but also other five subfamilies (named MIP A–E) not found in land plants²⁴. Animalia has four MIP families (AQP1-like, AQP8-like, AQP3-like, and AQP11-like)^{26,27}. Phylogenetic studies including plants and animals cluster PIPs with AQP1-like (considered the classical AQPs), TIPs with animal Aqp8-like, and SIPs with AQP11-like^{27,28}. There are different hypotheses regarding the origin of NIPs and AQP3-like^{27–29}, which is still an unresolved issue. Nevertheless, there is currently no disagreement about the existence of a common ancestor among Eukarya AQP3-like and Bacteria GlpFs so, the term GLP refers to this monophyletic group. On the other hand, the term AQP, when used, refers to a polyphyletic group.

As it is noticeable from the previous paragraphs, most of the described MIPs belong to two Eukarya supergroups (Archaeplastida and Amorphea -specifically Animalia-). In contrast, little is reported regarding other supergroups, such as Discoba, TSAR (Telonemia, Stramenopila, Alveolata, and Rhizaria), and Haptista. Still, the available data points to a quite diversified scenario in these supergroups. 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)³⁰. Also, there is no uniformity concerning MIP diversity among protozoans²⁸, while *Plasmodium* spp. (TSAR) carry a single MIP gene, up to five have been identified in the genomes of *T. brucei*, *T. cruzi*, and *L. major* (kinetoplastids, Discoba)³¹. *T. brucei* MIPs were previously set as GLPs and *T. cruzi* MIPs as AQPs, whereas *L. major* MIPs were described in both groups^{28,32}. Additionally, *L. major* and *T. cruzi* AQPs were regarded as TIP-related AQPs^{31,33}. However, none of those studies focused specifically on the phylogeny of the Kinetoplastea class MIPs. Today, the increased availability of genomes and transcriptomes of kinetoplastid species³⁴ provides the tools needed for a deep evolutionary study of MIPs diversity in this class.

Studies elucidating phylogenetic relationships among MIPs have opened ways to understand and predict relevant structure–function relationships in the evolution of utterly different organism lineages, such as tetrapods²², insects³⁵, and plants²⁷. In this work, we show that two MIP families expanded among trypanosomatids: GLP and a MIP family previously undescribed as such, named here AQPX. GLPs were not found in other kinetoplastid orders, whereas AQPXs were found in early-branching kinetoplastids. The AQPX family expanded in the Metakinetoplastina common ancestor before the origin of the parasitic order Trypanosomatida and extant trypanosomes hold up to four AQPX paralogs. Additionally, MIPs distribute asymmetrically inside the genus *Trypanosoma*: African trypanosomes specifically lost AQPXs and kept GLPs, whereas American trypanosomes specifically lost GLPs. This in-depth analysis of parasite MIPs may help understand the relevance of these channels in the physiology of the different parasites and assess their potential as drug targets.

Results and discussion

Kinetoplastid MIPs are either GLPs or non-orthodox AQPs.

We built a sequence similarity network (SSN) to explore where and how kinetoplastid MIPs localize in the superfamily sequence–function space. The starting point was a group of 52,453 MIPs retrieved from the Uniprot database. After clustering to 85% amino acid sequence identity and filtering by length, 16,170 representative accessions composed the network's nodes. The threshold for connecting nodes was set in an alignment score of 35 (corresponding to 35–40% pairwise sequence similarity) and rendered 10 clusters (Fig. 1a). Nearly half of the SSN nodes are from bacteria and the other half from eukaryotes, pointing to an expansion and diversification of the MIP superfamily that is similar in magnitude in both domains of life (Fig. 1b).

Already characterized MIPs that belong to different phylogenetic groups and with different permeation properties cluster separately in our SSN. Holding 80% of the nodes, Cluster 1 has a domain contribution similar to the full network, and the other smaller clusters are almost specific to Bacteria or Eukarya (Fig. 1b). MIPs with more divergent primary amino acid sequences localize in smaller clusters. That is the case for the plant XIPs and SIPs, the metazoan AQP11–12 group, algae MIPs (cluster 3, 5, 6, and 9, respectively), and other still uncharacterized divergent clusters (2, 4, 7, 8, and 10) (Fig. 1a). Figure 2 displays a detailed view of Cluster 1. Three main subclusters compose this cluster: (i) AQP_{SSN} (also internally structured allowing us to distinguish plant PIPs, TIPs, and NIPs, metazoans AQP1-like and AQP8-like and, prokaryotic AqpZs, AqpNs and AqpMs); (ii) GLP_{SSN} (where *T. brucei* and *T. evansi* MIPs localize among the Eukarya nodes), and (iii) AQPX_{SSN} (a small subcluster of mostly uncharacterized MIPs). The subindex SSN highlights that

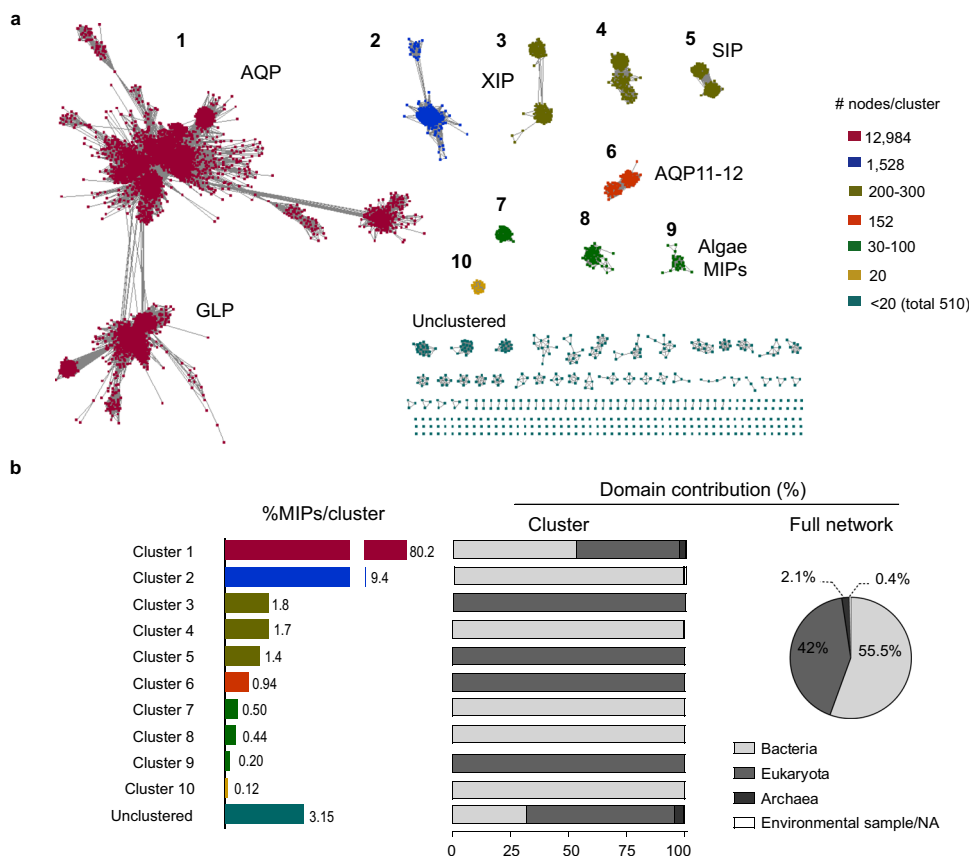


Fig. 1 Sequence similarity network (SSN) of the MIP superfamily. **a** SSN is composed of 16,170 nodes (squares), which represent proteins sharing >85% sequence identity, connected by edges with an average pairwise alignment score of at least 35. Edge length is a measure of the relative dissimilarity of each pair of sequences. Nodes are colored by the size of each cluster (numbered 1-10). Nodes found in clusters with <20 nodes were arbitrarily considered to uncluster. Names indicate clusters containing already characterized proteins. **b** MIPs distribution among clusters and taxonomic composition of each cluster according to the three-domain system of classification (Archaea, Bacteria, and Eukarya).

these groups arise from the network analysis and do not imply phylogenetic relations, even if both analyses can be congruent. Interestingly, many kinetoplastid (Discoba supergroup) MIPs are part of AQPX_{SSN}, a still uncharacterized subcluster that is far away from well-known MIPs.

Kinetoplastid MIPs are abundant among AQPX_{SSN} and AQPX is a MIP family. The AQPX_{SSN} subcluster is less crowded than the other two subclusters. Only 3% of Cluster 1 nodes are in this group. Long edges connect AQPX_{SSN} with AQP_{SSN}, whereas none edges connect it to the GLP_{SSN} (Figs. 1 and 2). AQPX_{SSN} is composed of uncharacterized prokaryotic and eukaryotic MIPs. Almost 75 and 10% of the AQPX_{SSN} nodes are from the Bacteria and Archaea domain of life, respectively (Fig. 2). The kinetoplastid MIPs, present in AQPX_{SSN}, have a unique closeness to prokaryotic uncharacterized MIPs. Thus, to investigate the putative origin of the MIPs that belong to the AQPX_{SSN} subcluster, we performed a phylogenetic analysis of the prokaryotic MIPs. The study included those bacterial MIPs present in AQPX_{SSN} (named AqpX) and the currently described four prokaryotic MIP grades (i.e., AqpM, AqpN, AqpZ, and Glp)²². Supplementary Data 1 details sequence data. Our study showed that AqpXs integrate a well-supported grade among prokaryotic MIPs. Therefore, this is evidence of AQPX being a grade of MIPs whose origin can be placed before the emergence of the Eukarya domain of life. Detailed analysis and discussion of this task are available in the Supplementary Results and Discussion, in Supplementary Fig. 1 and 2.

Regarding the eukaryotic MIPs present in AQPX_{SSN}, all nodes belong exclusively to unicellular organisms, 72% corresponding to the Kinetoplastea class (Discoba supergroup) and 16% to the TSAR supergroup (Fig. 2). The SSN exposed that Discoba and TSAR supergroups have different MIPs distribution as already suggested²⁸. Discoba MIPs distribute principally among the GLP_{SSN} and AQPX_{SSN} subclusters (41 and 57%, respectively), whereas TSAR MIPs are mainly from the GLP_{SSN} subcluster (93%) with a low percentage of isoforms distributed among the AQP_{SSN} and AQPX_{SSN} subclusters (5 and 2%, respectively). Altogether, this data points to an important presence of AQPX isoforms inside the Discoba supergroup and not in other Eukarya supergroups.

Asymmetric distribution of MIP repertoire among kinetoplastids. It has been previously described that *T. cruzi* and *T. brucei* do not share any MIP ortholog, whereas parasites of the genus *Leishmania* share MIP orthologs with the former two²⁸. Here, our SSN data stands out for the presence of AQPXs among trypanosomatids. To put all these data together and propose a hypothesis for the origin/s of trypanosomatid MIPs, we reconstructed MIPs phylogenetic history for the full Kinetoplastea class. We performed an intensive search of MIPs in publicly available databases and stumbled upon heterogeneous genome sequence availability (detailed in Supplementary Data 2). Trypanosomatida is the most studied order within the Kinetoplastea class with many genomes available, whereas the Prokinetoplastina subclass (*Ichthyobodo*, *Perkinsela*, PhM-4, and PhF-6) or

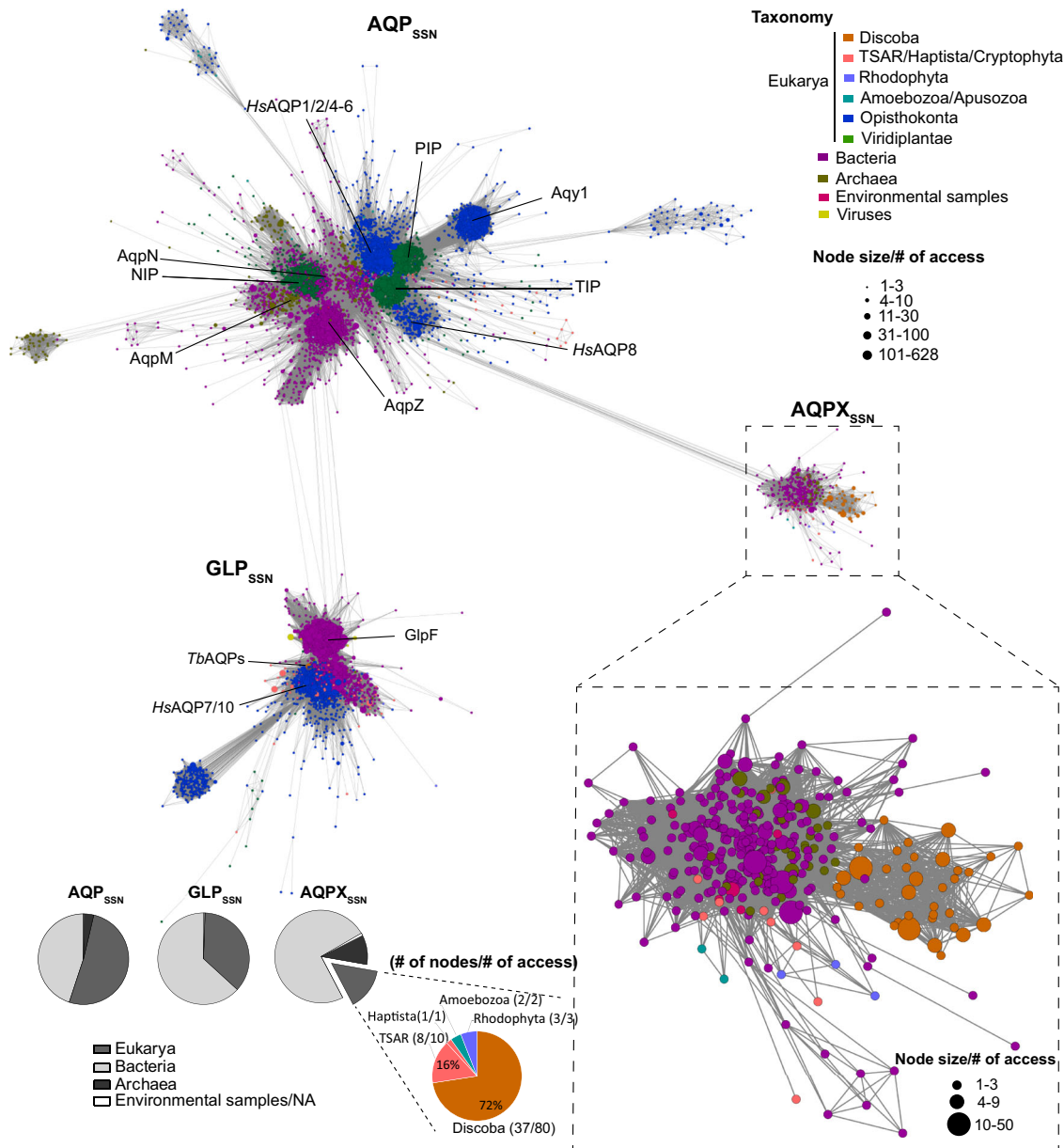


Fig. 2 Close up view of Cluster 1 of the Sequence similarity network (SSN). The cluster is composed of three main subclusters: AQP_{SSN}, GLP_{SSN}, and AQPX_{SSN}. Kinetoplastid AQPs localize in the AQPX subcluster. Nodes are colored to distinguish among taxonomic groups. The node size reflects the number of accessions that are grouped with 85% identity in that node. The pie charts show the taxonomic composition of each subcluster according to the three-domain system.

bodonida order have far fewer sequences available. Thus, we included transcriptome retrieved sequences to increase our data set of MIPs. In the specific case of *Parabodo caudatus* and *Pro-cryptobia sorokinii*, we retrieved their MIP sequences from studies where the bodonids were prey (Supplementary Data 3). We found no MIP sequences encoded in the genomes of two early-branching parasites/commensals (*Perkinsela* sp. and *Trypanoplasma borrelli*). Parasitism/commensalism evolved several times independently among kinetoplastids³⁶ (Fig. 3a) and, it seems that there is no relationship between this process and the MIP presence or absence in kinetoplastid genomes since, in opposition to *Perkinsela* sp. and *T. borrelli*, trypanosomatid parasites had many MIPs. Besides, the absence of MIP genes in a eukaryotic organism is a rare event that was only reported in three other protozoans: *Cryptosporidium parvum* (TSAR)³¹, *Tetrahymena thermophila* (TSAR), and *Giardia intestinalis* (Metamonada)²⁸. We also

searched for MIPs on species commonly used as outgroups in phylogenetic studies of kinetoplastids (i.e., euglenids or diplomids). The complete list of MIPs here analyzed is reported in Supplementary Data 4. Curiously, the sequence identity among kinetoplastid MIPs and diplomemid or euglenid MIPs is low (Supplementary Data 5). Therefore, we searched for MIP sequences within the complete Discoba supergroup (which includes Jakobids, Heterolobosea, and Euglenozoa) to observe the big picture by constructing a preliminary phylogenetic tree. This tree, which also included bacterial MIPs as reference for each already described grade, was built by the Maximum likelihood method, and was rooted in the long and fully supported branch that separated GLPs from other MIPs (Fig. 3b). Thus, our tree displays two primary branches at first sight, generally referred to as GLP and AQP (Fig. 3b). Notwithstanding this central division, we acknowledge the polyphyletic nature of the AQPs, further

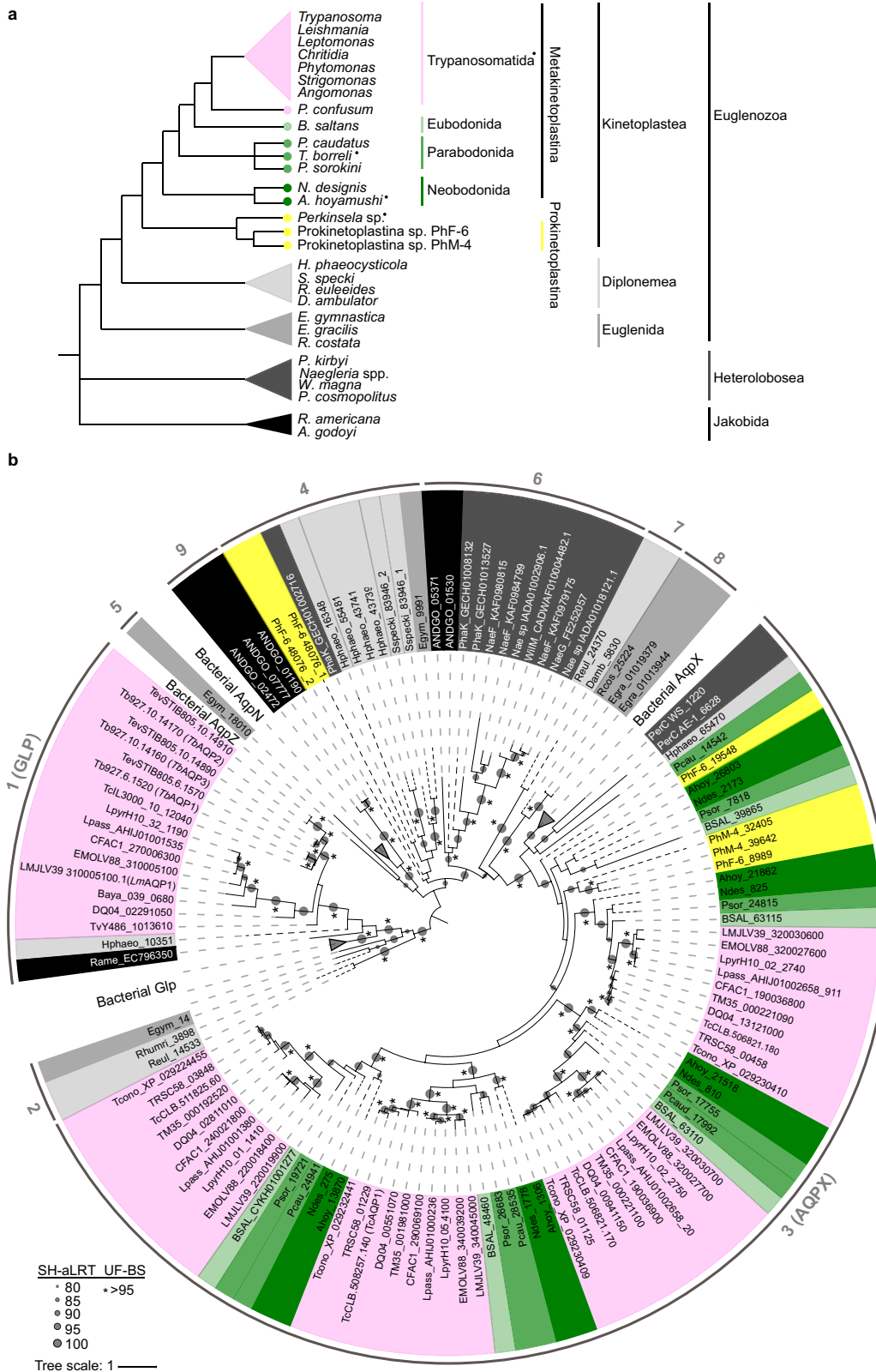


Fig. 3 Preliminary phylogenetic tree of the Discoba Major intrinsic proteins (MIPs) superfamily. a Cladogram showing the current accepted evolutionary relations inside the Discoba supergroup^{36,77,78}. Genera and species present in this cladogram correspond to those with genome/transcriptome data analyzed in the current work. Parasitic species or genera are indicated by full black circles superscripts. **b** Discoba MIPs preliminary phylogenetic tree reconstructed by maximum likelihood. Branch support was assessed by the ultrafast bootstrap (UF) approximation with 10,000 replicates and the SH-aLRT with 1000 replicates. Accessions were shaded following the same color code of the cladogram in **a**. Dashed lines represent transcriptome retrieved sequences and each MIP clade is referenced by a number.

explained over the text by describing each AQP group found (referenced by consecutive numbers, 3–9, in Fig. 3b) and focusing later precisely on AQPXs.

Inside the Euglenida group, only one phototrophic organism (the freshwater *Euglena gracilis*) has a sequenced genome and transcriptome available, and other two organisms (the phototrophic *Eutreptiella gymnastica* and the heterotrophic *Rhabdomonas costata* have transcriptomes available. *E. gymnastica* has two AQPs (Fig. 3b, groups 4 and 5) that do not group with the other euglenid isoforms (Fig. 3b, group 8). Besides, *R. costata* and *E. gracilis* AQPs, located in a long branch of the tree, have unique MIP structural determinants (Supplementary Data 6) and low sequence identity to all the other Discoba MIPs (under the 20%) (Supplementary Data 5). Three species are not enough to build up conclusions about the entire group, but it allows us to expose that lineage-specific MIPs evolved among euglenids and, none of them are ancestors of kinetoplastid MIPs. *Andalucia godoyi* (Jakobids) is the unique organism that we found to have MIPs grouping with AqpNs (Fig. 3b, group 9). Also, AQPs from *A. godoyi* and heterolobosean species integrate a supported group with low amino acid sequence identity to the other Discoba MIPs (Fig. 3b, group 6) and with >40% sequence identity to plant TIPs (NCBI BLAST results, 65–70% coverage). Group 4, even without significant statistical support, clusters Heterolobosea and Euglenozoa MIPs, keeping structural determinants that resemble AQP1-like channels or plant PIPs (Supplementary Data 6, selectivity filter), both proposed to derive from a common eukaryotic ancestor²⁷. Interestingly, just AQPs from Prokinetoplastina and none of the trypanosomatid MIPs form part of those previously described Discoba AQP groups (4–9). Instead, all trypanosomatid MIPs from the AQP branch belong to the AQPX cluster with Bacteria AqpXs (Fig. 3b, group 3). There exists the possibility that a trypanosomatid ancestor acquired an AQPX by lateral gene transfer, an event already described for several trypanosomatid genes³⁷. But the AQPXs were present in early-branching kinetoplastids (Prokinetoplastina), and therefore are ancestral kinetoplastid genes. Thus, if the acquisition of AQPXs occurred by lateral gene transfer, it happened before the kinetoplastid lineage emerged.

In opposition to the vast number of AQPXs, our analysis revealed a small number of GLPs among kinetoplastids (Fig. 3b, group 1). Moreover, we found only trypanosomatid GLP isoforms, and we found no bodonid, nor prokinetoplastina GLPs, suggesting an asymmetric MIPs repertoire among kinetoplastids. Without considering trypanosomatids, we found only five GLPs. One diplomemid (*Hemistasia phaeocysticola*) and one jakobid (*Reclinomonas americana*) isoform showed 19–29% identity to trypanosomatids GLPs and similar structural determinants (Supplementary Data 5 and Supplementary Data 6). While other two GLPs from diplomemids (*Rhynchopus* spp.) and one from a euglenid (*E. gymnastica*) (Fig. 3b, group 2) had lower sequence identity to trypanosomatids GLPs (15–24%) and different structural determinants (Supplementary Data 5 and Supplementary Data 6). Comparing among the trypanosomatid species, we observed that African trypanosomes (*T. evansi*, *T. congolense*, *T. vivax*, and *T. brucei*) have only GLP representatives and no AQPXs. Also, outside the Trypanosoma genus, the genome of *Blechnomonas ayalai* codified only for a GLP. On the contrary, American trypanosomes (*T. theileri*, *T. rangeli*, *T. conorhini*, and *T. cruzi*) have four MIPs, all of which are AQPXs, and none GLP. *T. grayi* remains an exception to this matter as its genome codes for the four AQPXs and one GLP, similar to the genomes of *Leishmania* spp.

Finally, to evaluate the reliability of the heterogeneous sources of Discoba MIPs we analyzed the completeness of the genome and transcriptome assemblies using the tool Benchmarking

Universal Single-Copy Orthologs (BUSCO). Most of them showed good levels of completeness (Supplementary Data 2, analyzed in Supplementary Results and Discussion). Additionally, most of the transcriptomes here analyzed were already used to successfully carry out a comparative analysis of euglenozoans metabolic enzymes and molecular features (DNA pre-replication complex, kinetochore machinery)³⁴. Altogether, this indicates that a reliable set of assemblies was used in our MIPs searches. Still, it is worth mentioning that a different picture might be reconstructed once more Discoba organisms have their genomes sequenced and can be included in the study.

Origin of the trypanosomatid AQP α - δ clades in the Metakinetoplastina group. Our preliminary analysis showed that the GLP grade was less crowded than the AQP group, as if an expansion among AQP grades had occurred. This burst can be seen specifically in the AQPX family, populated by trypanosomatids. Thus, to better understand kinetoplastid AQPXs' evolutionary history, we built a phylogenetic tree for the Discoba supergroup analyzing a wider diversity of trypanosomatids. We added the early-branching trypanosomatid, *Paratrypanosoma confusum*, the plant infecting *Phytomonas*, and the monoxenous genera *Angomonas* and *Strigomonas*. The AQPX isoforms of the early-branching Discoba organism *Percolomonas cosmopolitus* (Heterolobosea) served as root.

In this tree, trypanosomatid AQPX isoforms segregate together with bodonid MIPs in four very well-supported orthologous clusters: α , β , γ , and δ (named after *T. cruzi* and *L. major* aquaporins³¹) (Fig. 4). Each cluster is internally congruent with the organismal tree at species levels and, within each one, sequence identities go from 50 to 90% (Supplementary Data 7). AQPXs from Prokinetoplastina, early-branching kinetoplastids, compose a sister clade of these α - δ clades. AQPXs of free-living bodonid (eu-, para-, and neo-bodonids), and the only diplomemid AQPX found, form a more distant clade from trypanosomatid AQPXs, but this node is not statistically supported (Fig. 4). Altogether, we propose that the α - δ loci appeared through gene duplication from a single ancestral locus in the genome of an ancestral metakinetoplastid before the diversification of extant genera.

Gains and losses of MIPs in Trypanosome genomes. Inside the Trypanosomatida order, the genomes are highly syntenic³⁸, even though our phylogenetic analysis showed important differences in the displayed MIPs repertoire exposed by its members. Thus, to get more clues about trypanosomatid MIPs history, we compared the genomic neighborhood of these channels among representative trypanosomatids and their closest known non-parasitic relative, *B. saltans* (Fig. 5). We analyzed nine genomes, four of them are assembled at the chromosome level (*T. cruzi*, *T. brucei*, *T. congolense*, and *L. major*), two at the supercontig level (*P. confusum* and *B. saltans*) and three at the contig level (*T. grayi*, *T. theileri* and *B. ayalai*) (Supplementary Table 1). Overall, the quality of the assemblies, even if not homogeneous, is undoubtedly good. The genome coverages for the studied regions are among 41X and 200X (Supplementary Data 8). The coverage and undefined regions (Ns) are available in Supplementary Figs. 3–8. In Fig. 6, we summarized the accumulated knowledge relative to Discoba MIPs diversity. Inside the Kinetoplastea class, we propose a scheme of gains and losses compatible with our phylogenetic and syntenic data (Fig. 6a).

There is conserved synteny for the α - δ AQPXs of trypanosomatids and *B. saltans* (Fig. 5a and Supplementary Figs. 3–5) even when this bodonid genome only showed ~10% co-linearity with trypanosomatid genomes³⁹. A fifth AQPX in *B. saltans*, with low

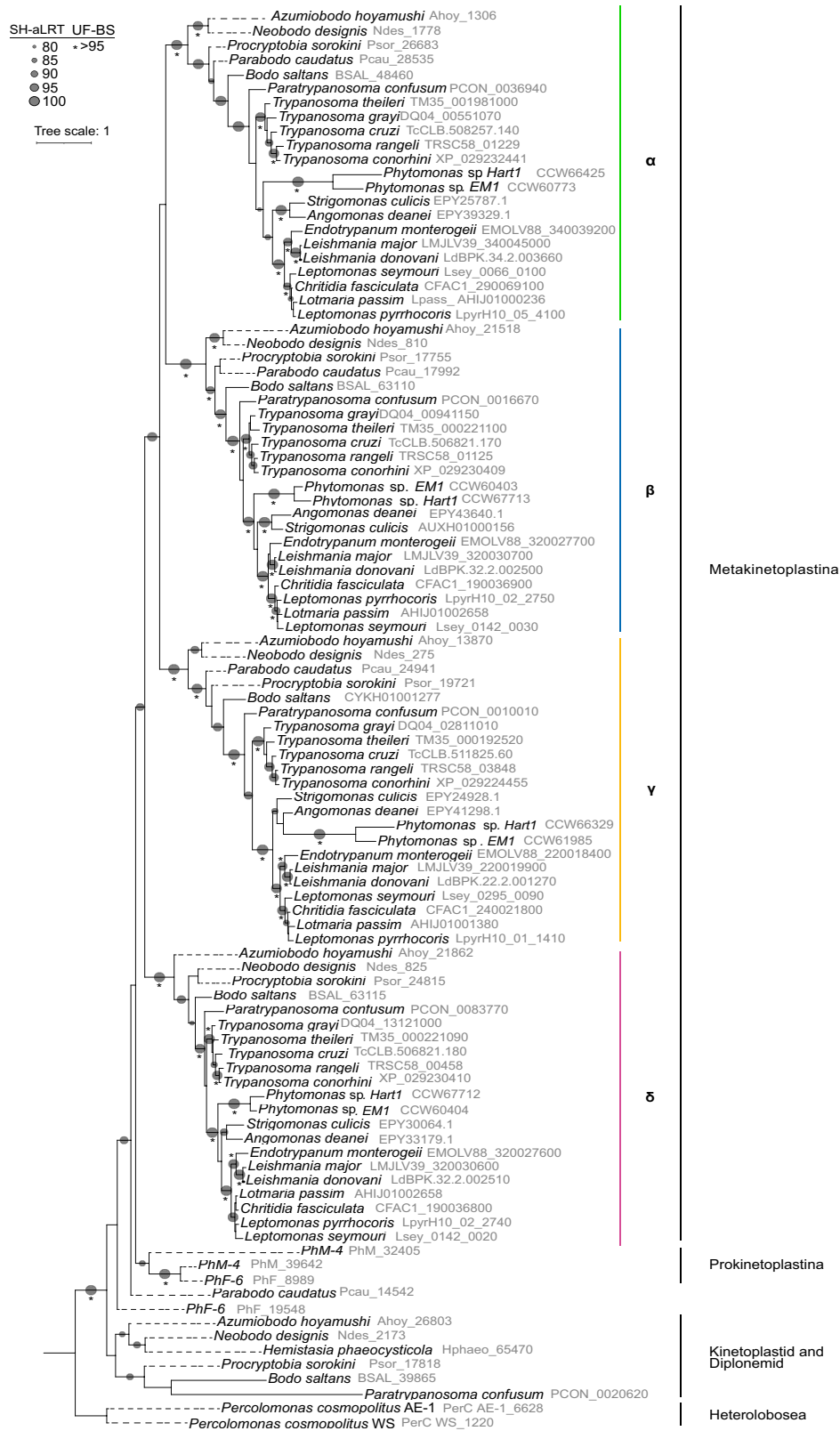


Fig. 4 AQPX phylogenetic tree of the Discoba supergroup. Phylogeny of AQPX proteins was reconstructed by maximum likelihood. Branch support was assessed by the ultrafast bootstrap (UF) approximation with 10,000 replicates and the SH-aLRT with 1000 replicates. Dashed lines represent transcriptome retrieved sequences.

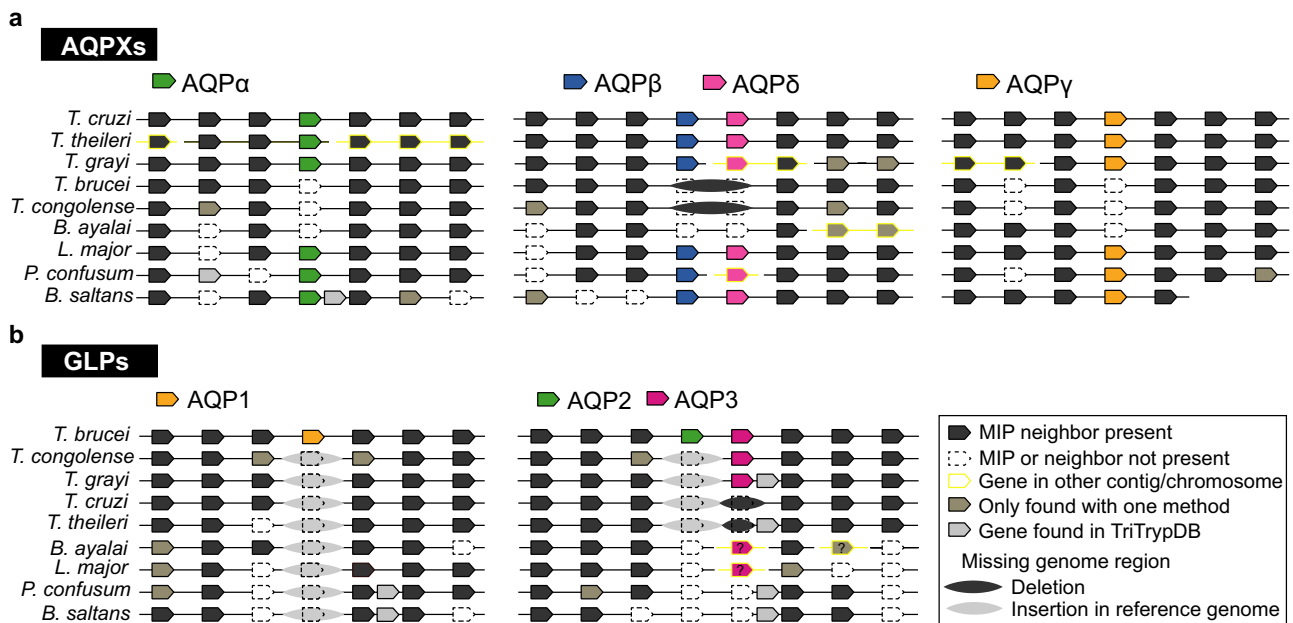


Fig. 5 Synteny analysis of Trypanosomatid MIPs. Comparative gene organization of the regions where **a** AQPXs and **b** GLPs are found. The analyzed regions comprise 10 Kb down and up-stream of each MIP in the reference genomes (*T. cruzi* or *T. brucei*) and the equivalent syntenic regions in the other genomes. For clarity, in this figure only the first three genes for down and up-stream regions are shown. Homologous genes are vertically aligned. The graph shows when the syntenic genes are observed using both SimpleSynteny analysis and TriTrypDB (black arrow box) or one of these methods (dark gray arrow box). Genes absent in the reference genome (*T. cruzi* or *T. brucei*) were not detected by SimpleSynteny but were observed in TriTrypDB (light gray arrow box).

sequence identity with all the other AQPXs, localizes in a genomic region non-syntenic with parasite genomes, neither with the fifth AQPX of *P. confusum* (Supplementary Fig. 6). Then, the genome region coding for this *B. saltans* AQPX probably was lost in the trypanosomatids ancestor during the genome rearrangement in the transition from free-living to parasitism.

Among trypanosomatids, α - δ AQPXs seem to have been lost two times in different branches of the evolutionary tree (in African trypanosomes and *B. ayalai*, Fig. 6a).

Even when the AQPXs are missing in these two groups, the flanking genes are conserved (Fig. 5a). In the particular case of α and γ AQPXs, the accumulation of mutations seems to be the mechanism of gene losses, as the size of the intergenic region among flanking genes is close to 1 Kb, the expected size for these AQPXs, (Supplementary Figs. 3–5). The β and δ AQPXs localize in tandem in trypanosomatids and *B. saltans* (not in *P. confusum*), and different mechanisms seem to be after these gene losses. In African trypanosomes, β and δ AQPXs losses seem a consequence of a deletion in their most recent common ancestor genomic region. In contrast, their losses in *B. ayalai* seem not to be associated with genomic deletions but with the accumulation of mutations (Fig. 5a, and detailed synteny data in Supplementary Fig. 4).

The closely related species *T. brucei* and *T. evansi* are the only two trypanosomatids carrying three GLPs (Fig. 6b). *TbAQP1* neighbor genes are conserved inside the Trypanosomatinae subfamily. In contrast, none of *TbAQP1*'s orthologs appear in that syntenic region. Thus, *TbAQP1* (and its ortholog in *T. evansi*) appears to be a recent acquisition, via transpositive duplication, in their last common ancestor (Fig. 5b). *TbAQP2* and *TbAQP3* localize in tandem in *T. brucei* chromosome 10 and, *TbAQP2* seems to be a consequence of a recent duplication event as *TbAQP3* has a higher sequence identity with the GLPs of the other trypanosomatids than *TbAQP2* (*T. congolense*, *T. grayi*, *L. major*, and *B. ayalai*). The *TbAQP2-3* genomic region is syntenic

within the subfamily Trypanosomatinae, missing the GLPs only in American trypanosomes (Fig. 5b). Nevertheless, synteny is not conserved in this region among the subfamilies Leishmaniae and Trypanosomatinae. That is congruent with the analysis reported by El-Sayed et al.³⁸ of the syntenic blocks among *L. major* and *T. brucei* (neither *TbAQP2-3* nor *LmAQP1* genome regions are in the described syntenic blocks). Besides, no GLP genes were found in *P. confusum* or *B. saltans* genomes. The orthologs of *TbAQP3* flanking genes are retained but the intergenic region among those genes is large in *B. saltans* (near to 4 Kb) and even larger in *P. confusum* (near 35 Kb) (Supplementary Fig. 7). Moreover, this large region of *P. confusum* is undefined and therefore we cannot exclude the presence of a GLP in there. Therefore, we assembled transcriptomes available for *P. confusum* (Supplementary Data 2) and searched for GLPs, finding none. To complete the analysis, we also searched for GLPs in *B. saltans* transcriptomes (Supplementary Data 2), and we found none either. We can think that *TbAQP3* orthologous genes were specifically lost in these species. But, outside trypanosomatids, kinetoplastids lack GLPs, and the scenario of GLP loss in every lineage is very improbable. The most parsimonious scenario is the acquisition of a GLP in the common ancestor of trypanosomatids (after *P. confusum* branched at the Trypanosomatidae family base) which was then lost precisely two times: in American trypanosomes and the subfamily Strigomonadinae (Fig. 6a, b).

So, genera and species-specific gene gains and losses resulted in an asymmetric repertoire of MIPs in extant trypanosomatid parasites. Such processes are usual in the evolutionary history of other protein families among *T. brucei*, *T. cruzi*, and *Leishmania* species (i.e. cathepsins, amastins, nucleoside, and amino acid transporters)^{39,40}. Utterly different lifestyles and hosts might relate to species-specific gene expansions and losses. For example, amastin diversity remained unchanged until the origin of *Leishmania*. So, the specific δ -amastin expansion that occurred in this species was speculated to relate to *Leishmania*'s vertebrate

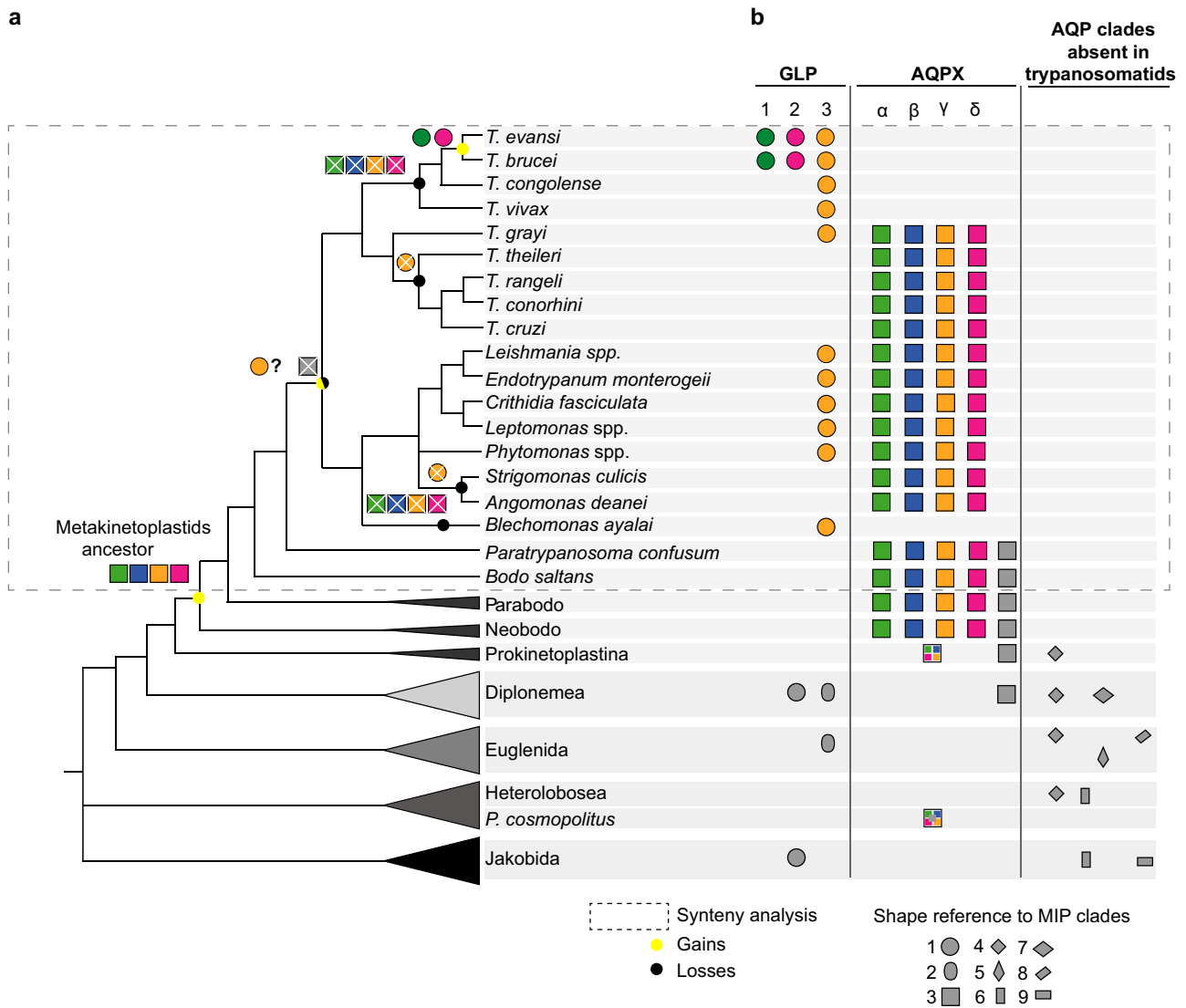


Fig. 6 Proposed evolutionary history of Trypanosomatid MIPs. **a** Cladogram showing the current accepted evolutionary relations inside the Discoba supergroup for the species analyzed in this work. **b** The MIPs repertoire for each species or group of species is represented by geometric shapes (proposed homologous genes share the same shape). Each shape is linked to a number in the references, and a number connects to the preliminary phylogenetic tree from Fig. 3. A dashed line square delimits the taxonomic groups represented by the syntenic analysis. Inside Metakinetoplastina, orthologs are represented in the same color. Inside the AQPX clade, for simplicity and because the phylogenetic relationships of some AQPXs are not fully resolved, gray squares may represent more than one homologous isoform. Multicolor squares represent ancestral homologs of trypanosomatid AQPXs. Over the cladogram, a proposed scheme of gains and losses, of GLPs and AQPXs, is depicted.

parasitism given the absence of this gene family in related monoxenous species (insect-restricted parasitism)⁴⁰. Regarding the MIP superfamily, biological relevance of each family (GLP and AQPX) in trypanosomes still remains obscure though the asymmetric pattern is coherent with the proposal of an evolutionary relationship between the loss of AQPs and consequent expansion of GLPs (or the other way around) based on observations of other unicellular organisms like Oomycetes, that hold numerous GLP isoforms and none AQP2s²⁸.

Key structure determinants of kinetoplastids AQPXs. To gather evidence of the putative role of MIPs in the evolution of kinetoplastids, we analyzed those key residues known to be related to the function and selectivity of the channels (i.e., the two signature NPA motifs, the selectivity filter and the Froger positions).

When GLPs are analyzed, it emerges that most of the trypanosomatids hold the same amino acids in NPA, selectivity

filter, and Froger Positions (except the extremely variable P5) (Fig. 7a). Among these isoforms, some have been functionally characterized as permeable to several solutes (Supplementary Data 6). For example, *LmAQP1* facilitates the diffusion of water and many non-ionic solutes (methylglyoxal, glycerol, dihydroxyacetone, glyceraldehyde, erythritol, and adonitol) but not urea⁴¹. Also, this GLP acts as a metalloid (As and Sb) gateway with implications in therapeutic interventions⁴².

The most recently acquired GLP of *T. brucei* and *T. evansi* (AQP2) present utterly divergent key MIP residues from the other GLPs (Fig. 7a). These AQP2s are the only GLPs with non-canonical NPA motifs (NSA and NPS). Importantly, the N in the first position of the motifs that have been proved to be important for cation blockage^{11,12} is conserved in *T. brucei* and *T. evansi* AQP2. Interestingly, functional consequences of the absence of both classical NPA motifs in *TbAQP2* are related to pentamidine sensitivity since the restitution of the NPA-NPA blocked the uptake of the drug⁷.

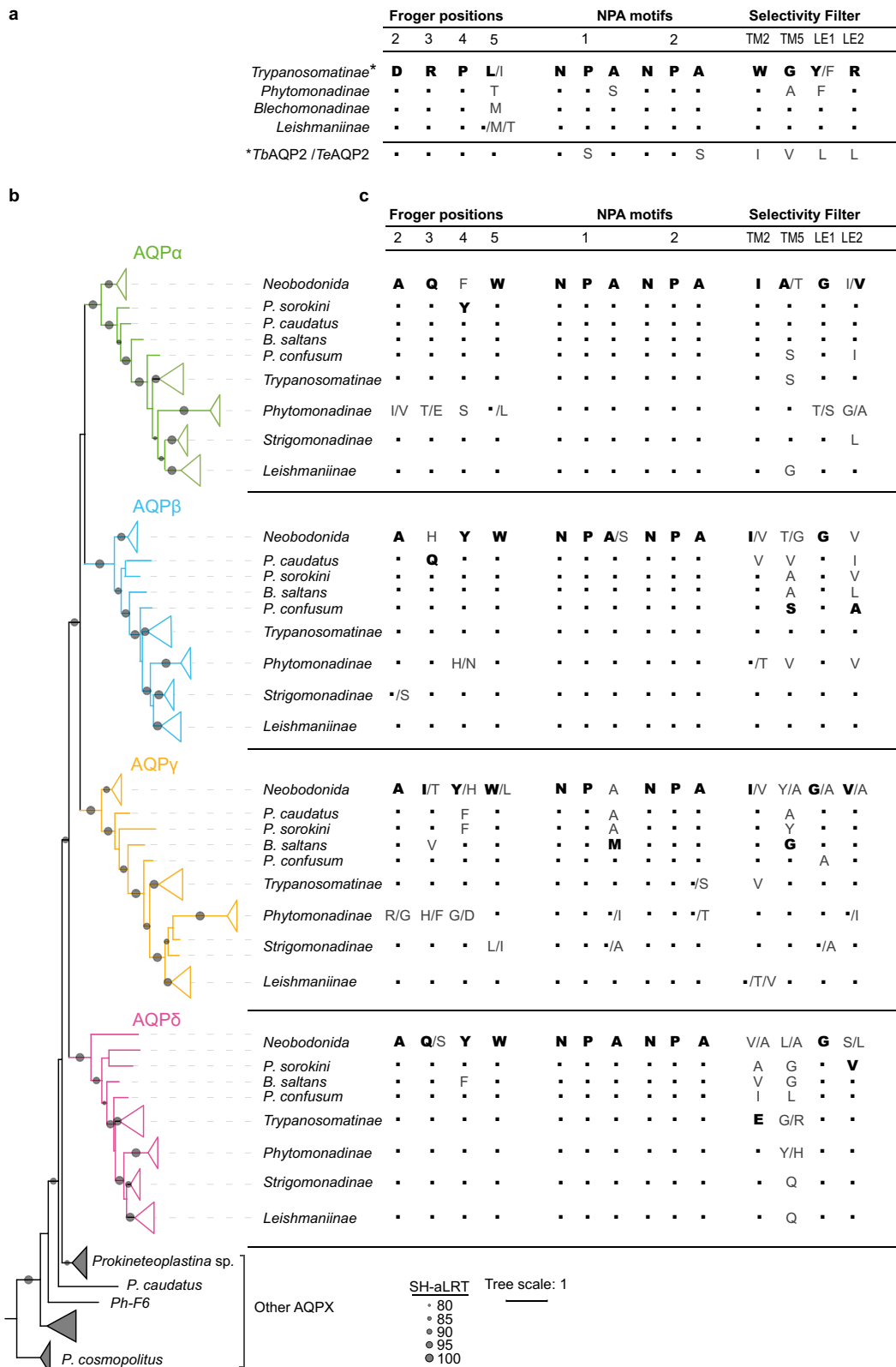


Fig. 7 Key MIP residues from GLP and AQPX subfamilies. Froger positions, NPA motifs and Selectivity filter residues of Discoba **a** GLPs and **b**, **c** AQPXs. The most frequent amino acid of a particular position in each clade is shown in bold first and then with dots. **b** Collapsed Discoba AQPXs phylogenetic tree reconstructed by maximum likelihood. **c** Analysis of residues.

Regarding the selectivity filter, these AQP2 carry a rare signature (IVLL), which is drastically different from the fully conserved selectivity filter of other trypanosomatids, or even *Discoba* GLPs (WGYP)^{7,43} (Fig. 7a and Supplementary Data 6). *TbAQP2* selectivity filter is wider and more aliphatic than others. A first hypothesis sustains that this feature contributes to pentamidine passing through^{7,44} and a second one that the unique selectivity filter in combination with a consequently exposed Asp (D265, Froger position P2), allows a high affinity binding of pentamidine followed by endocytosis⁴⁵. It is vital to bear that no other *T. brucei* MIP participates in pentamidine uptake (*TbAQP1* nor *TbAQP3*), whereas all *T. brucei* MIPs facilitate the diffusion of water, glycerol, and metalloids in a similar way^{46,47}. *TbAQP2* also presents a very different expression pattern compared to *TbAQP1* (the most abundant MIP in *T. brucei*) and *TbAQP3* (only present in blood stages)⁴⁶. Also, *TbAQP2* and *TbAQP3* have different subcellular localization and might play different roles in permeating water, glycerol and still undiscovered solutes^{46,48,49}. *TbAQP2* changes in key residues plus the different localization and transcription levels among paralogs point to their neofunctionalization in the last common ancestor of *T. brucei* and *T. evansi*.

On a different note, AQPXs distribute among kinetoplastids. All orders (Trypanosomatida, Bodonida, and Prokinetoplastina) have at least two AQPXs. We analyzed key MIP residues, sorted by kinetoplastid orders and even subfamilies (Fig. 7b). AQPXs display generally conserved Froger positions (AQYW from P2 to P5) (Fig. 7c) with AQP-like residues occupying them. Regarding the NPA motifs, while α , β , and δ AQPXs present well-conserved NPAs, AQPys present the first (N-terminal) motif as NPM (Fig. 7c and Supplementary Data 6). Interestingly, this substitution is absent in prokinetoplastina, para- and neo-bodonids, suggesting that it occurred in the common ancestor of *B. saltans* and trypanosomatids. Two isoforms from the γ clade carry neither classical NPA motifs: *TcAQP γ* (NPM-NPS) and *AQP γ* from *Phytomonas* sp. EM1 (NPI-NPT) (Supplementary Data 6). Currently, there is no data regarding the permeation capabilities of these last two AQP γ , but there is some information about other members of the γ clade with an N-terminal NPM motif. Homology modeling studies suggest that *LmAQP γ* maintains a well-conserved core structure⁵⁰, and functional studies showed that the *LdAQP γ* is so far the only AQPX of this parasite that facilitates water permeation³². Confirmation of these results for other clade members could reveal a neofunctionalization of this AQPX in the last common ancestor among the free-living *B. saltans* and trypanosomatids.

AQPXs have a rare pattern that resembles none of the previously described selectivity filter for the different families of AQPs (FHTR, PIPs; FHCR, AQP1-like; HIA/GR/V, TIPs; HIA/GR, AQP8-like; and T/PL/VAL, unorthodox AQPs)⁵¹. Compared with selectivity filter in classical water channels AQP1-likes and PIPs, the selectivity filter of AQPXs do not keep the R in Loop E (LE2), nor the aromatic amino acids in TM2, having, instead, aliphatic residues (Fig. 7c). That may give place to more hydrophobic and broader filters. Though their selectivity filter is aliphatic, they also hold an aliphatic uncharged residue (an A) where *TbAQPs* have an acidic amino acid (Froger position P2) and the impact of these differences and the eventual exposure of other AQPX residues affecting permeation or selectivity needs to be addressed by further structural and functional research. Finally, many AQPXs (except the β orthologs) have a V in the LE2 position. The presence of a V in this position was reported as a signature for subcellular MIPs⁵². Consistently, *TcAQP α* is present in acidocalcisomes and a vacuolar structure near the flagellar pocket^{53,54}, *LdAQP α* and δ in subcellular structures³². From a functional aspect, none of those mentioned MIPs

(*TcAQP α* , *LdAQP α* , *LdAQP γ* , and *LdAQP δ*) allow glycerol permeation^{32,53}. This functional data was not expected because, as already mentioned, a wider selectivity filter seems to be present in AQPXs⁵⁰.

Recently proposed permeation mechanisms through *TbAQP2*⁷ allow us to ask whether AQPXs might be capable of facilitating the uptake of larger solutes. However, as mentioned above, they have so far poor water or glycerol permeability. They may present a different solute selectivity profile given their rare selectivity filter, they might have additional undescribed pore constrictions, or there might be still unknown regulatory factors stabilizing their open or closed states influencing heterologous expression results and conclusions. It is worth mentioning that conclusions based on MIP motifs and their respective consequences on pore sizes and selectivity profiles can only be reached on the bases of structural results. Crystallization or ab initio/homology combined models need to be pursued to elucidate Kinetoplastid AQPX structures given their low identity with already crystallized MIPs.

In conclusion, we depicted here the complex universe of MIPs through a SSN, clearly exposing that trypanosomatids carry GLPs and AQPXs. AQPXs compose a cluster far away from the already characterized MIPs and, our phylogenetic studies support that they integrate, to the best of our knowledge, a newly defined MIP family. We got an insight into the phylogenetic study of these channels in kinetoplastids. We found that the α - δ clades appear in the common ancestor of bodonids and trypanosomatids. Curiously, African trypanosomes lost all the AQPX isoforms. Instead, these trypanosomes hold GLPs that we proposed to be acquired in a trypanosomatid ancestor and specifically lost in American trypanosomes. Was this change of MIPs repertoire inside the Trypanosomatinae subfamily a gene replacement process among GLPs and AQPXs? AQPXs hold selectivity filter residues that allow us to speculate that they have a more hydrophobic and wider selectivity filter than classical AQPs. Then, can the solutes permeated by AQPXs possibly be similar to GLPs permeated ones? As already exposed, AQPX do not seem to have good glycerol permeability. Nonetheless, the nature of biologically relevant solutes that permeate these channels is still elusive. Future research on the permeation capability and structure of GLPs and AQPXs will help understand their importance in the parasite's physiology. That, together with the knowledge on MIP repertoire and evolutionary history are crucial steps to unveil possible drug sensitivity/resistance mechanisms in the treatment of trypanosomiasis.

Methods

Construction of sequence similarity network. The SSN of the MIP superfamily was generated using the EFI-EST server⁵⁵. The full-size Pfam PF00230 database was downloaded from UniProt (version 2020-02). Proteins were clustered at 85% amino acid sequence identity using h-cd-hit⁵⁶ and filtered by length (200–500 residues). A list of 16,170 accessions, representative of 52,453 sequences, was loaded in the EFI-EST server (Option D). An alignment score of 35 (corresponding to ~40% sequence identity) was used to generate the SSN. The resultant network of ~8 M edges was visualized in the open-source software Cytoscape 3.8⁵⁷ using a 64 GB RAM server (Supplementary Data 9).

Sequence retrieval and phylogenetic analysis. To build the prokaryotic MIP tree protein sequences already known to belong to specific MIP families (i.e., AqpM, AqpN, AqpZ, Glp) were retrieved from Pommerrenig et al. (2020)⁵⁸, and together with the prokaryotic AqpX sequences retrieved from our SSN analysis, were clustered at 60% amino acid sequence identity using h-cd-hit⁵⁶. Sequences were aligned using MAFFT⁵⁹ v7 and trimmed using TrimAl⁶⁰ (-g 0.8 -cons 65). The list of accessions is in Supplementary Data 1.

Protein sequences from the *Discoba* supergroup organisms were retrieved from the public databases TriTrypDB, NCBI, and iMicrobe. First, we included MIP sequences that were tagged as aquaporin in the database, and we used a tBLASTn strategy to expand our set of MIPs. When no available genome was found for a given organism, we searched within transcriptome, either by blasting within published and publicly available assemblies or by assembling the Sequence Read Archive (SRA) using the rnaSPAdes software⁶¹ in the Galaxy servers at usegalaxy.

org.au and usegalaxy.org⁶². The sequence assemblies from Butenko et al.³⁴ were provided by Dr. Lukeš lab. Parabodonid MIPs were retrieved from studies where *P. caudatus* and *P. sorokini* were prey. RNAseq from samples that contained the Parabodonida and other species (PhF-6, *Rhodolphis limneticus*, *Rhodolphis marinus*) and cleaned RNAseq from those species were compared. Sequences were considered as putative Parabodonida MIPs, analyzing their identity among different RNAseq (Supplementary Data 3) and observing their position in the phylogenetic tree. MIP sequences wrongly assigned to *Colpodella angusta* (NCBI) were confirmed to belong to its prey, *P. caudatus*. *C. angusta* supposed MIPs that were only partial were almost identical to the retrieved *P. caudatus* MIPs (sequence identity climbed to 98 and 99%). Additionally, no genomic nor transcriptomic data was found for *Ichtyobodo* (Prokinetoplastina), *Cryptobia* (Parabodonid, Metakinetoplastina), *Dimagistella* spp., *Klosteria*, *Rhynchobodo* sp., or *Actuariola* (Neobodonids, Metakinetoplastina). *Percolomonas cosmopolitus* cultures were fed with *Enterobacter aerogenes*. Thus, the presence of bacterial contaminating transcripts was tested for the strain WS. Megablast of *Percolomonas cosmopolitus* Strain WS assembly against BLAST nucleotide database (nt17-Apr-2014) showed that 524 of 11,058 query sequences had a match (cut off e-value 10^{-3}). When selecting only the first match for the 524 query sequences (hit lowest e-value), 26 query sequences were matching with bacterial sequences. Suggesting very low contamination with bacterial RNA (26/11,058) and none of the matches correspond to the MIPs found in the transcriptome. The quality and completeness of the proteomes, transcriptomes, and genomes used in this study were assessed by using BUSCO tool suite v5.0.0⁶³ in the Galaxy public servers at usegalaxy.org.au and usegalaxy.org⁶². The datasets selected to run BUSCO were the closest to the lineage of the species under study, eukaryota_odb10 and euglenozoa_odb10 datasets. The web resource SMART (Simple Modular Architecture Research Tool)⁶⁴ was used to corroborate the domain architecture of the putative MIPs. All sequences used for the phylogenetic analysis and the information about their accession and type of data are listed in Supplementary Data 2–4. Multiple sequence alignment (MSA) was performed with retrieved sequences using MAFFT, V7 (E-INS-i strategy, leaving gappy regions, Blossum62 as scoring matrix and MAFFT homologous option activated). Prokaryotic MIPs were included as they appeared to have a high amino acid sequence similarity (30%) to kinetoplastid MIPs (Supplementary Data 5) and appeared in BLAST searches when the Kinetoplastid class was excluded. Sequences were then trimmed using TrimAL (-g 0.8 -cons 50) to conserve only the more confidently aligned regions.

Phylogenetic trees were built using IQ-TREE⁶⁵ 2.0-rc2 and the evolutionary relationships among sequences were inferred by using the maximum likelihood (ML) method. The best-fit model was found using ModelFinder⁶⁶. Branch support was calculated with the ultrafast bootstrap test⁶⁷ (10,000 iterations) and the Shimodaira-Hasegawa approximate likelihood ratio test (SH-aLRT)⁶⁸ (1,000 iterations). The best-fit model was LG+R8 for the Prokaryotic MIPs analysis, LG+F+R7 for the Preliminary tree of Discoba MIPs and, LG+F+R6 for the Discoba AQPX tree. The phylogenetic tree files in newick format are provided in Supplementary Data 10–12. Trees were edited using the Interactive Tree of Life tool⁶⁹. A visually revised alignment based on the resultant tree topology was constructed by manually correcting alignment errors and the phylogenetic tree analysis was performed again.

Synteny analysis. Synteny analysis was conducted by using BLAST+ (version 2.10.1+⁷⁰), SimpleSynteny software⁷¹ and by exploring the TriTrypDB⁷² genome browser. First, tBLAST was performed using the MIP and surrounding proteins found 10 Kb upstream and 10 Kb downstream. *T. cruzi* was used as a reference for protein sequences of AQPX alpha-gamma. *T. brucei* as a reference for GLP sequences. The genomes used as subjects in tBLAST search were the ones from *T. cruzi* (TcruziCLBrennerNon-Esmeraldo-like), *T. brucei* (TbruceiTREU927), *T. theileri* (TheileriEdinburgh), *T. grayi* (TgrayiANR4), *T. congolense* (TcongolenseIL3000_2019), *B. ayalai* (BayalaiB08-376), *L. major* (LmajorLV39c5), *P. confusum* (PconfusumCUL13), and *B. saltans* (BsaltansLakeKonstanz). The assembly status and metrics of these genomes were calculated using Quast v5.0.2⁷³ and are reported in the Supplementary Table S1. To calculate the coverage of the regions used for synteny analysis, the raw reads used for the assemblies (Supplementary Data 8) were mapped to the corresponding assembled genome using Bowtie2 with default parameters⁷⁴, and then the coverage analysis was performed using SAMtools⁷⁵. These analyses were performed in the Galaxy public servers at usegalaxy.org.au and usegalaxy.org⁶². For *T. brucei* we recovered the MIPs region coverage from TriTrypDB genome browser (Jbrowser). For *T. congolense* we could not find the SRAs used for the assembly in any public database. So, we used reads of another WGS project of the same strain to estimate the coverage. Synteny was checked by manual inspection of the tBLAST result table. Genomic regions showing syntenic genes were selected, including 1 Kb before and after the first and the last gene in synteny, respectively. These regions were used as input for SimpleSynteny software. SimpleSynteny uses mainly two cutoff parameters to find syntenic genes, the e-value, and the query coverage, set to 0.01 and 10%, respectively.

MIP residue assessment. MSA was performed as described above, in this case, to identify typical MIP residues in specific alignment positions. We used Bioedit 7.2.5⁷⁶ to visualize and extract specific positions from the MSA. We specifically

gather the information regarding: (i) Froger Positions, from 2 to 5, P1 was left out given that it remained a conflictive position in the MSA; (ii) both canonical NPA motifs; and (iii) selectivity filter residues located the second and fifth transmembrane domains (TMM2 and TMM5 respectively) along with two residues in loop E (LE1 and LE2). Already characterized MIPs (*Escherichia coli* GlpF and AqpZ, TbaQP1, 2 and 3) were used to check the alignment and the identity of the defined positions. Residue assessment was shown related to phylogeny to confirm these critical positions within the evolutionary history of the analyzed MIPs.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Received: 24 February 2021; Accepted: 21 July 2021;

Published online: 10 August 2021

References

1. Steverding, D. Sleeping sickness and Nagana disease caused by *Trypanosoma brucei* in *Arthropod Borne Diseases* (ed. Marcondes, C. B.) 277–297 (Springer, Cham, 2017).
2. Stoco, P. H., Miletti, L. C., Picozzi, K., Steindel, M. & Grisard, E. C. Other major trypanosomiasis in *Arthropod Borne Diseases* (ed. Marcondes, C. B.) 299–324 (Springer, Cham, 2017).
3. Pérez-Molina, J. A. & Molina, I. Chagas disease. *Lancet* **391**, 82–94 (2018).
4. de Koning, H. P. The drugs of sleeping sickness: their mechanisms of action and resistance, and a brief history. *Trop. Med. Infect. Dis.* **5**, 1–23 (2020).
5. Caldas, I. S., Santos, E. G. & Novaes, R. D. An evaluation of benzimidazole as a Chagas disease therapeutic. *Expert Opin. Pharmacother.* **20**, 1797–1807 (2019).
6. Munday, J. C., Settimo, L. & de Koning, H. P. Transport proteins determine drug sensitivity and resistance in a protozoan parasite, *Trypanosoma brucei*. *Front. Pharmacol.* **6**, 1–10 (2015).
7. Alghamdi, A. H. et al. Positively selected modifications in the pore of TbAQP2 allow pentamidine to enter *Trypanosoma brucei*. *Elife* **9**, 1–33 (2020).
8. Julia von Bülow, J. & Beitz, E. Attacking aquaporin water and solute channels of human-pathogenic parasites: New routes for treatment? in *Aquaporins In Health And Disease: New Molecular Targets For Drug Discovery* (eds. Soveral, G., Nielsen, S. & Casini, A.) 233–246 (CRC Press, 2016).
9. Borgnia, M., Nielsen, S., Engel, A. & Agre, P. Cellular and molecular biology of the aquaporin water channels. *Annu. Rev. Biochem.* **68**, 425–458 (1999).
10. Verkman, A. & Mitra, A. Structure and function of aquaporin water channels. *Am. J. Physiol. Ren. Physiol.* **278**, F13–F28 (2000).
11. Wu, B., Steinbronn, C., Alsterfjord, M., Zeuthen, T. & Beitz, E. Concerted action of two cation filters in the aquaporin water channel. *EMBO J.* **28**, 2188–2194 (2009).
12. Wree, D., Wu, B., Zeuthen, T. & Beitz, E. Requirement for asparagine in the aquaporin NPA sequence signature motifs for cation exclusion. *FEBS J.* **278**, 740–748 (2011).
13. Murata, K. et al. Structural determinants of water permeation through aquaporin-1. *Nature* **407**, 599–605 (2000).
14. Tajkhorshid, E. et al. Control of the selectivity of the aquaporin water channel family by global orientational tuning. *Science* **296**, 525–530 (2002).
15. Eriksson, U. K. et al. Subangstrom resolution x-ray structure details aquaporin-water interactions. *Science* **340**, 1346–1349 (2013).
16. De Groot, B. L. & Grubmüller, H. The dynamics and energetics of water permeation and proton exclusion in aquaporins. *Curr. Opin. Struct. Biol.* **15**, 176–183 (2005).
17. Hub, J. S. & de Groot, B. L. Mechanism of selectivity in aquaporins and aquaglyceroporins. *Proc. Natl Acad. Sci. USA* **105**, 1198–1203 (2008).
18. Baker, N. et al. Aquaglyceroporin 2 controls susceptibility to melarsoprol and pentamidine in African trypanosomes. *Proc. Natl Acad. Sci. USA* **109**, 10996–11001 (2012).
19. Froger, A., Tallur, B., Thomas, D. & Delamarche, C. Prediction of functional residues in water channels and related proteins. *Protein Sci.* **7**, 1458–1468 (1998).
20. Park, J. H. & Saier, M. H. Phylogenetic characterization of the MIP family of transmembrane channel proteins. *J. Membr. Biol.* **153**, 171–180 (1996).
21. Zardoya, R. Phylogeny and evolution of the major intrinsic protein family. *Biol. Cell* **97**, 397–414 (2005).
22. Finn, R. N., Chauvigné, F., Hlidberg, J. B., Cutler, C. P. & Cerdà, J. The lineage-specific evolution of aquaporin gene clusters facilitated tetrapod terrestrial adaptation. *PLoS ONE* **9**, 1–38 (2014).

23. Danielson, J. Å. H. & Johanson, U. Unexpected complexity of the Aquaporin gene family in the moss *Physcomitrella patens*. *BMC Plant Biol.* **8**, 1–16 (2008).
24. Anderberg, H. I., Danielson, J. Å. H. & Johanson, U. Algal MIPs, high diversity and conserved motifs. *BMC Evol. Biol.* **11**, 1–15 (2011).
25. Gustavsson, S., Lebrun, A. S., Nordén, K., Chaumont, F. & Johanson, U. A novel plant major intrinsic protein in *Physcomitrella patens* most similar to bacterial glycerol channels. *Plant Physiol.* **139**, 287–295 (2005).
26. Danielson, J. Å. H. & Johanson, U. Phylogeny Of Major Intrinsic Proteins in *MIPs and Their Role In The Exchange Of Metalloids* (eds. Jahn, T. P. & Bienert, G. P.) 19–31 (Springer, 2010).
27. Soto, G., Alleva, K., Amodeo, G., Muschietti, J. & Ayub, N. D. New insight into the evolution of aquaporins from flowering plants and vertebrates: orthologous identification and functional transfer is possible. *Gene* **503**, 165–176 (2012).
28. Abascal, F., Irisarri, I. & Zardoya, R. Diversity and evolution of membrane intrinsic proteins. *Biochim. Biophys. Acta* **1840**, 1468–1481 (2014).
29. Finn, R. N. & Cerdà, J. Evolution and functional diversity of aquaporins. *Biol. Bull.* **229**, 6–23 (2015).
30. Khabudaev, K. V., Petrova, D. P., Grachev, M. A. & Likhoshway, Y. V. A new subfamily LIP of the major intrinsic proteins. *BMC Genomics* **15**, 1–7 (2014).
31. Beitz, E. Aquaporins from pathogenic protozoan parasites: structure, function and potential for chemotherapy. *Biol. Cell* **97**, 373–383 (2005).
32. Biyani, N. et al. Characterization of *Leishmania donovani* aquaporins shows presence of subcellular aquaporins similar to tonoplast intrinsic proteins of plants. *PLoS ONE* **6**, 1–14 (2011).
33. Quintana, J. F. & Field, M. C. Evolution, function and roles in drug sensitivity of trypanosome aquaglyceroporins. *Parasitology* **148**, 1137–1142 (2021).
34. Butenko, A. et al. Evolution of metabolic capabilities and molecular features of diplomonads, kinetoplastids, and euglenids. *BMC Biol.* **18**, 1–28 (2020).
35. Finn, R. N., Chauvigné, F., Stavang, J. A., Belles, X. & Cerdà, J. Insect glycerol transporters evolved by functional co-option and gene replacement. *Nat. Commun.* **6**, 1–7 (2015).
36. Lukeš, J., Skalický, T., Týč, J., Votýpka, J. & Yurchenko, V. Evolution of parasitism in kinetoplastid flagellates. *Mol. Biochem. Parasitol.* **195**, 115–122 (2014).
37. Opperdoes, F. R. & Michels, P. A. Horizontal gene transfer in trypanosomatids. *Trends Parasitol.* **23**, 470–476 (2007).
38. El-Sayed, N. M. et al. Comparative genomics of trypanosomatid parasitic protozoa. *Science* **309**, 404–409 (2005).
39. Jackson, A. P. et al. Kinetoplastid phylogenomics reveals the evolutionary innovations associated with the origins of parasitism. *Curr. Biol.* **26**, 161–172 (2016).
40. Jackson, A. P. The evolution of amastin surface glycoproteins in trypanosomatid parasites. *Mol. Biol. Evol.* **27**, 33–45 (2010).
41. Figarella, K. et al. Biochemical characterization of *Leishmania major* aquaglyceroporin LmAQP1: possible role in volume regulation and osmotaxis. *Mol. Microbiol.* **65**, 1006–1017 (2007).
42. Gourbal, B. et al. Drug uptake and modulation of drug resistance in *Leishmania* by an aquaglyceroporin. *J. Biol. Chem.* **279**, 31010–31017 (2004).
43. Munday, J. C. et al. *Trypanosoma brucei* aquaglyceroporin 2 is a high-affinity transporter for pentamidine and melaminophenyl arsenic drugs and the main genetic determinant of resistance to these drugs. *J. Antimicrob. Chemother.* **69**, 651–663 (2014).
44. Quintana, J. F. et al. Instability of aquaglyceroporin (Aqp) 2 contributes to drug resistance in *Trypanosoma brucei*. *PLoS Negl. Trop. Dis.* **14**, 1–26 (2020).
45. Song, J. et al. Pentamidine is not a permeant but a nanomolar inhibitor of the *Trypanosoma brucei* aquaglyceroporin-2. *PLoS Pathog.* **12**, 1–14 (2016).
46. Uzcátegui, N. L. et al. Cloning, heterologous expression, and characterization of three aquaglyceroporins from *Trypanosoma brucei*. *J. Biol. Chem.* **279**, 42669–42676 (2004).
47. Uzcátegui, N. L. et al. *Trypanosoma brucei* aquaglyceroporins facilitate the uptake of arsenite and antimonite in a pH dependent way. *Cell. Physiol. Biochem.* **32**, 880–888 (2013).
48. Bassarak, B., Uzcátegui, N. L., Schönfeld, C. & Duzenko, M. Functional characterization of three aquaglyceroporins from *Trypanosoma brucei* in osmoregulation and glycerol transport. *Cell. Physiol. Biochem.* **27**, 411–420 (2011).
49. Uzcátegui, N. L. et al. *Trypanosoma brucei* aquaglyceroporins mediate the transport of metabolic end-products: methylglyoxal, D-lactate, L-lactate and acetate. *Biochim. Biophys. Acta Biomembr.* **1860**, 2252–2261 (2018).
50. Neumann, L. S. M., Dias, A. H. S. & Skaf, M. S. Molecular modeling of aquaporins from *Leishmania major*. *J. Phys. Chem. B* **124**, 5825–5836 (2020).
51. Finn, R. N. & Cerdà, J. Aquaporin in *Encyclopedia of Signaling Molecules* (ed. Choi, S.) 374–390 (Springer, Cham, 2018).
52. Ishibashi, K. Aquaporin subfamily with unusual NPA boxes. *Biochim. Biophys. Acta* **1758**, 989–993 (2006).
53. Montalvetti, A., Rohloff, P. & Docampo, R. A functional aquaporin co-localizes with the vacuolar proton pyrophosphatase to acidocalcisomes and the contractile vacuole complex of *Trypanosoma cruzi*. *J. Biol. Chem.* **279**, 38673–38682 (2004).
54. Li, Z.-H. et al. Hyperosmotic stress induces aquaporin-dependent cell shrinkage, polyphosphate synthesis, amino acid accumulation, and global gene expression changes in *Trypanosoma cruzi*. *J. Biol. Chem.* **286**, 43959–43971 (2011).
55. Zallot, R., Oberg, N. & Gerlt, J. A. The EFI web resource for genomic enzymology tools: leveraging protein, genome, and metagenome databases to discover novel enzymes and metabolic pathways. *Biochemistry* **58**, 4169–4182 (2019).
56. Huang, Y., Niu, B., Gao, Y., Fu, L. & Li, W. CD-HIT suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **26**, 680–682 (2010).
57. Shannon, P. et al. Cytoscape: a software environment for integrated models. *Genome Res.* **13**, 2498–2504 (2003).
58. Pommerrenig, B. et al. Functional evolution of nodulin 26-like intrinsic proteins: from bacterial arsenic detoxification to plant nutrient transport. *New Phytol.* **225**, 1383–1396 (2020).
59. Katoh, K., Rozewicki, J. & Yamada, K. D. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief. Bioinform.* **20**, 1160–1166 (2019).
60. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
61. Bushmanova, E., Antipov, D., Lapidus, A. & Pribelski, A. D. RnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. *Gigascience* **8**, 1–13 (2019).
62. Afgan, E. et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* **46**, 537–544 (2018).
63. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
64. Schultz, J., Milpetz, F., Bork, P. & Ponting, C. P. SMART, a simple modular architecture research tool: Identification of signaling domains. *Proc. Natl Acad. Sci. USA* **95**, 5857–5864 (1998).
65. Nguyen, L. T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
66. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A. & Jermini, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
67. Minh, B. Q., Anh, M., Nguyen, T. & von Haeseler, A. Ultrafast approximation for phylogenetic bootstrap. *Mol. Biol. Evol.* **30**, 1188–1195 (2013).
68. Guindon, S. et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
69. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44**, W242–W245 (2016).
70. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2019). <https://doi.org/10.1186/1471-2105-10-421>
71. Veltri, D., Wight, M. M. & Crouch, J. A. SimpleSynteny: a web-based tool for visualization of microsynteny across multiple species. *Nucleic Acids Res.* **44**, W41–W45 (2016).
72. Aslett, M. et al. TriTrypDB: a functional genomic resource for the *Trypanosomatidae*. *Nucleic Acids Res.* **38**, D457–D462 (2010).
73. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
74. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–360 (2012).
75. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, 1–4 (2021).
76. Hall, T. A. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.* **41**, 95–98 (1999).
77. Garcia, H. A. et al. Pan-American *Trypanosoma (Megatrypanum) trinaperronei* n. sp. in the white-tailed deer *Odocoileus virginianus* Zimmermann and its deer ked *Lipoptena mazamae* Rondani, 1878: morphological, developmental and phylogeographical characterisation. *Parasites Vectors* **13**, 1–18 (2020).
78. Bradwell, K. R. et al. Genomic comparison of *Trypanosoma conorhini* and *Trypanosoma rangeli* to *Trypanosoma cruzi* strains of high and low virulence. *BMC Genomics* **19**, 1–20 (2018).

Acknowledgements

The authors acknowledge Dr. Julius Lukeš, who kindly provided PhM-4, PhF-6, *Hemistasia phaeocysticola*, *Rhynchopus rumis* and *Sulcionema specki* transcriptome assemblies, Dr. Vyacheslav Yurchenko, who kindly provided *Blechnomonas ayalai* SRAs, and Dr. Juan Pedro Liron for providing the computing power to build and visualize the SSN. This work was supported by the Agencia Nacional de Promoción Científica y Tecnológica (PICT 2017-0244 granted to K.A.).

Author contributions

Conceived and designed the experiments: F.C.T., K.A. and A.R.F. Performed the experiments: F.C.T., A.R.F. and R.L. Analyzed the data: F.C.T., R.L., K.A. and A.R.F. Wrote the paper: F.C.T., K.A. and A.R.F. Jointly supervised this study: K.A. and A.R.F. All authors revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s42003-021-02472-9>.

Correspondence and requests for materials should be addressed to K.A. or A.R.F.

Peer review information *Communications Biology* thanks Igor Cestari and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: Nico Fanget and Eve Rogers. Peer reviewer reports are available.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021