

SCIENTIFIC REPORTS



OPEN

Evolutionary analyses of myosin genes in trypanosomatids show a history of expansion, secondary losses and neofunctionalization

Denise Andréa Silva de Souza^{1,2}, Daniela Parada Pavoni^{1,2}, Marco Aurélio Krieger^{1,2,3} & Adriana Ludwig^{1,3}

Myosins are motor proteins that comprise a large and diversified family important for a broad range of functions. Two myosin classes, I and XIII, were previously assigned in Trypanosomatids, based mainly on the studies of *Trypanosoma cruzi*, *T. brucei* and *Leishmania major*, and important human pathogenic species; seven orphan myosins were identified in *T. cruzi*. Our results show that the great variety of *T. cruzi* myosins is also present in some closely related species and in *Bodo saltans*, a member of an early divergent branch of Kinetoplastida. Therefore, these myosins should no longer be considered “orphans”. We proposed the classification of a kinetoplastid-specific myosin group into a new class, XXXVI. Moreover, our phylogenetic data suggest that a great repertoire of myosin genes was present in the last common ancestor of trypanosomatids and *B. saltans*, mainly resulting from several gene duplications. These genes have since been predominantly maintained in synteny in some species, and secondary losses explain the current distribution. We also found two interesting genes that were clearly derived from myosin genes, demonstrating that possible redundant or useless genes, instead of simply being lost, can serve as raw material for the evolution of new genes and functions.

Myosins are important eukaryotic molecular motor proteins that bind actin filaments and are dependent of ATP hydrolysis¹. They are related to several molecular processes, such as muscle contraction in metazoans, cytokinesis, cell migration, intracellular transport of molecular cargoes and organelles, and host cell invasion of apicomplexan parasites^{2,3}.

Most myosin proteins consist of an N-terminal motor domain (myosin head domain - PF00063) responsible for actin binding and ATP hydrolysis, a neck region containing one or more IQ motifs that bind calmodulin or other members of the EF-hand family of proteins, and a C-terminal tail responsible for cargo binding and/or dimerization that may contain distinct domains^{4,5}.

In the last 20 years, a number of works have addressed the diversity, classification and evolution of myosin genes using different approaches and datasets, revealing that myosins are a large and diversified gene family^{3,5-10}. Due to this diversity, their classification is not trivial work and there is still not a consensus from the last works. Two most recent myosin superfamily works, Odronitz and Kollmar⁵ and Sebé-Pedrós *et al.*⁹, used the phylogenetic relationship of myosin motor domains for classification. However, mainly due to taxon sampling differences, the authors defined a distinct number of classes (35⁵ or 31 classes⁹, respectively). Most studies included several taxa, such as early branching eukaryotes of Kinetoplastida class from Excavata.

Kinetoplastids are a widespread and diverse group of flagellated protozoans that are free-living or parasites and comprise five orders: Trypanosomatida, Eubodonida, Parabodonida, Neobodonida and Prokinetoplastida¹¹. They contain a characteristic structure called kinetoplast, which is a complex mitochondrial DNA inside of a single mitochondrion¹². Within kinetoplastids, the most studied family is the Trypanosomatidae, which is composed mainly of monoxenous parasite species that infect invertebrates and of dixenous species that can be pathogenic to plants, animals and/or humans¹³. Among these, *Trypanosoma cruzi*, *T. brucei* and *Leishmania spp.* are the

¹Laboratório de Genômica Funcional, Instituto Carlos Chagas- ICC/Fiocruz-, PR, Curitiba, 81350-010, Brazil.

²Programa de Pós-graduação em Biociências e Biotecnologia - ICC/Fiocruz-, PR, Curitiba, 81350-010, Brazil.

³Instituto de Biologia Molecular do Paraná, Curitiba, 81350-010, Brazil. Correspondence and requests for materials should be addressed to M.A.K. (email: mkrieger@fiocruz.br) or A.L. (email: adriludwig@gmail.com)

Gene	Gene ID	Gene size (bp)	Protein size (aa)	Domains
Myo1	TCDM_07314	3,498	1,165	Myosin motor domain (cd01378/PF00063); IQ motif; Unconventional myosin tail domain (Myosin TH1 superfamily domain - cl26987); WW domain (cd00201); FYVE_like_SF (cd00065); coiled-coil region
Myo13	TCDM_05821	3,177	1,058	Myosin motor domain (MYSc_Myo13, cd14875); coiled-coil region; two tandem UBA superfamily domains (cl21463)
MyoA	TCDM_09957	3,012	1,003	Myosin motor domain (cd00124); IQ motif; two coiled-coil regions
MyoB	TCDM_07433	3,357	1,118	Myosin motor domain (cd00124); coiled-coil regions; IQ motif
MyoC	TCDM_02877	3,304	1,167	Myosin motor domain (cd00124); coiled-coil region
MyoD	TCDM_07686	3,678	1,225	Myosin motor domain (cd00124); IQ motif
MyoE	TCDM_07686	3,288	1,095	Myosin motor domain (cd00124); two IQ motifs
MyoF	TCDM_08875	4,446	1,481	Myosin motor domain (cd00124); IQ motif; two coiled-coil regions
MyoG	TCDM_02016	3,669	1,222	Myosin motor domain (cd00124); TPH (pfam13868)
MyoH-derived	TCDM_02145	3,594	1,197	Short region of class XIII myosin motor domain (cd14875); TPH (ERM superfamily - cl25742); Neuromodulin_N superfamily (cl26511); SMC_N superfamily (cl25732); two coiled-coil regions

Table 1. Information of *T. cruzi* Dm28c myosin genes used as query in the BlastP searches.

causative agents of important neglected illnesses: Chagas disease, African sleeping sickness and leishmaniasis, respectively¹⁴.

Odriontz and Kollmar⁵ used several trypanosomatid genomes (*T. cruzi*, *T. brucei*, *T. vivax*, *T. congolense*, *L. major*, *L. infantum* and *L. braziliensis*) in their myosin phylogenetic tree. All of these species presented a Myo1 gene, representative of class I that is found in almost all organisms, and a Myo13 gene, classified as class XIII, a Kinetoplastida-specific class. Notably, *T. cruzi* has seven additional myosin genes (named MyoA, MyoB, MyoC, MyoD, MyoE, MyoF and MyoG) that were not classified, being considered “orphan” myosins since no orthologues were found in other species⁵. In Seb e-Pedr os *et al.*⁹, authors only included *L. major* sequences, and the trypanosomatid myosin classification remained the same.

With the sequencing of some *T. cruzi*-related species, we observed that *T. cruzi* “orphan” myosins are more widely distributed than previously thought, leading to questions about the evolution and function of these genes. Here, we present a deep evolutionary analysis of the myosin gene family from Trypanosomatidae, with exciting new data about expansion, diversification, loss and neofunctionalization of these genes in Kinetoplastida. Although the role of these genes remains to be addressed, this work is a starting point for functional studies.

Results and Discussion

Evolutionary analyses of myosin genes. For all myosins, we used the *T. cruzi* Dm28c amino acid sequences as query on BlastP searches and the main features of these genes are available in Table 1. We focus our main text on the evolutionary aspects, while a detailed result of protein domain searches of all sequences found in this work can be seen in Supplementary Table S1 together with a Supplementary Discussion. Supplementary Table S2 contains the information about the genome assemblies used in work and we provided all BlastP and synteny analysis results in the Supplementary Tables S3 to S22.

To understand the evolution of myosin genes, we compared their phylogenies to the known trypanosomatid phylogenetic relationships (Fig. 1). Trypanosomatid phylogeny can be separated into two major clades, one comprising the *Trypanosoma* genus, and the other, which we named the *Leishmania* clade, that encompasses *Blechnomonas ayalai*, *Phytomonas* sp. and the subfamilies Strigomonadinae and Leishmaniinae. The separation of these two groups is estimated at 231–283 million years ago (mya)¹⁵. *Bodo saltans*, a non-trypanosomatid free-living kinetoplastid also studied here, is clearly an early branch. All evolutionary scenarios presented here were hypothesized according to the available data and based on the most parsimonious picture. However, alternative explanations cannot be discarded, and the addition of new kinetoplastid species into the phylogeny could give a better view of the evolution of the myosin family.

Myo1 protein is widely found in trypanosomatids. Using TritypDB BlastP searches, we were able to find Myo1 orthologous genes in all analysed species. In addition, NCBI BlastP was performed to search for trypanosomatid sequences that are not available at TritypDB, and we found Myo1 sequences in *T. theileri*, *Angomonas deanei*, *Strigomonas culicis* and in *B. saltans*. Some species (*T. grayi*, *T. rangeli*, *T. theileri* and *B. saltans*) presented several significant hits, indicating that these species have a wide repertoire of myosin genes, similar to *T. cruzi*. The wide distribution of Myo1 in Kinetoplastida is expected since this class has the widest taxonomic distribution (being absent only in Viridiplantae and Alveolata), probably by being the first myosin to have evolved⁵. Additionally, Myo1 genes were found in previous works that sampled trypanosomatid species^{3,5,8,9}. We found a conserved synteny (the same gene order in a chromosomal segment between species) for Myo1 genes in almost all species.

In general, Myo1 sequences presented conserved structure and motifs/domains composition, having a myosin motor domain in the N-terminal region, an IQ motif and an unconventional myosin tail domain (Myosin TH1 superfamily domain) with a WW and a FYVE_like_SF domain. This protein also contains a putative C-terminal coiled-coil forming region. *T. rangeli* presents a truncated protein; however, we cannot discard that it is an artefact of genome sequencing/assembly.

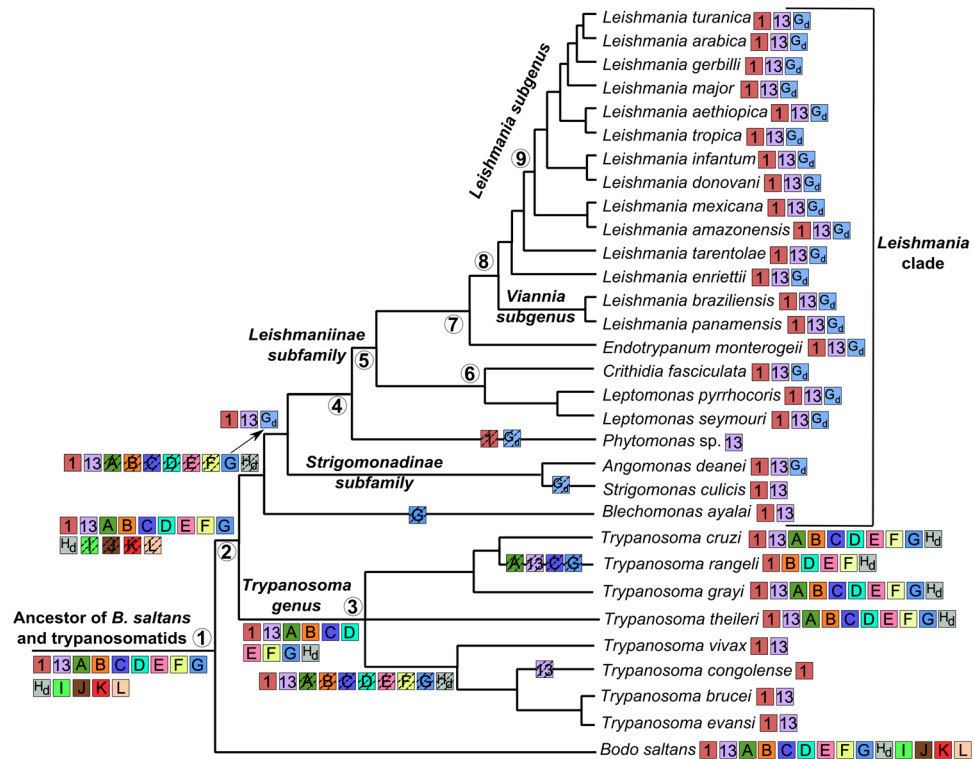


Figure 1. Schematic of phylogenetic relationships among different trypanosomatids employed in this study based on several works^{15,42–48}. *Bodo saltans* was represented as outgroup. Coloured boxes represent the presence of different myosin genes. Hashed boxes represent possible gene loss events. Gd – MyoG-derived gene. Hd – MyoH-derived gene. Numbers in circles near some nodes are references to the divergence time estimates in mya (millions years ago), as inferred by Lukes *et al.*¹⁵: 1 – 463–695 mya; 2 – 231–283 mya; 3 – 96–105 mya; 4 – 118–170 mya; 5 – 52–96 mya; 6 – 30–63 mya; 7 – 31–65 mya; 8 – 25–54 mya; 9 – 9–23 mya. The branching is not drawn to scale.

Interestingly, *Phytomonas* sp. (isolates Hart1 and EM1) did not present any hit that corresponded to a Myo1 orthologue. It is not clear if this absence is due to genome assembly issues. Nonetheless, the absence in two independent genome assemblies supports the idea of gene loss. This last hypothesis is corroborated by the whole-genome analysis that indicates a minimized gene repertoire in *Phytomonas* EM1 and HART1 genomes¹⁶.

Excluding the kinetoplastids from NCBI BlastP, the best hits corresponded to sequences from the *Phytophthora* genus (~87% extension, 41% identity). In the Odrionitz and Kollmar (2007) myosin phylogeny, trypanosomatid Myo1 sequences also grouped, with high support, with myosin from *Phytophthora* spp.⁵ Considering this close relation, we used two *Phytophthora* sequences to root the Kinetoplastida Myo1 tree (Fig. 2). Myo1 phylogeny presents low support for the basal nodes and some incongruities in relation to the species phylogeny. *B. saltans* is positioned closer to the *Trypanosoma* genus instead of being a basal branch. One hypothesis for this incongruity could be ancestral polymorphism followed by independent lineage sorting. However, the conserved synteny suggests the same copy gene was maintained from the ancestor. The clade containing *Bl. ayalai*, *S. culicis* and *A. deanei* is misplaced, but the bootstrap value is very low. The wrong positioning could be explained by long branch attraction, since the long branches of outgroups can frequently “attract” long branches of species to the base of the tree¹⁷. We can observe by the short branches among species that Myo1 genes from Leishmaniinae subfamily members are highly conserved; nevertheless, there was likely an accelerated evolutionary rate in the ancestor species that is represented by the long branch connecting this group.

In general, we can postulate that the Myo1 gene was present in the ancestor of kinetoplastids as a single copy gene and was conserved in almost all species, possibly due to an essential function, although little is known about the role of this gene in these organisms. In *T. brucei* bloodstream forms, Myo1 protein partially colocalizes with elements of the endocytic pathway and compartments containing internalized cargo, indicating its involvement in this pathway, being lethal in knockdown cells. However, in procyclic forms, the knockdown has no obvious effect on growth and morphology¹⁸. In contrast, in *L. donovani*, the protein Myo1 was not found by antibody recognition in either procyclic or amastigote forms, indicating that this gene is not expressed in this parasite¹⁹. Despite this finding, it is not reasonable, based on our phylogenetic analysis, that this myosin has remained conserved along the evolution without functional activity. It is possible that this gene is expressed in intermediate forms of the parasite’s cycle or expressed in such low amounts that it was not detected by the authors. In addition, we cannot presume that the function of Myo1 genes is the same in all trypanosomatids, mainly because some species demonstrate high divergence in the protein sequence, such as *A. deanei* and *S. culicis*. Moreover, the function of

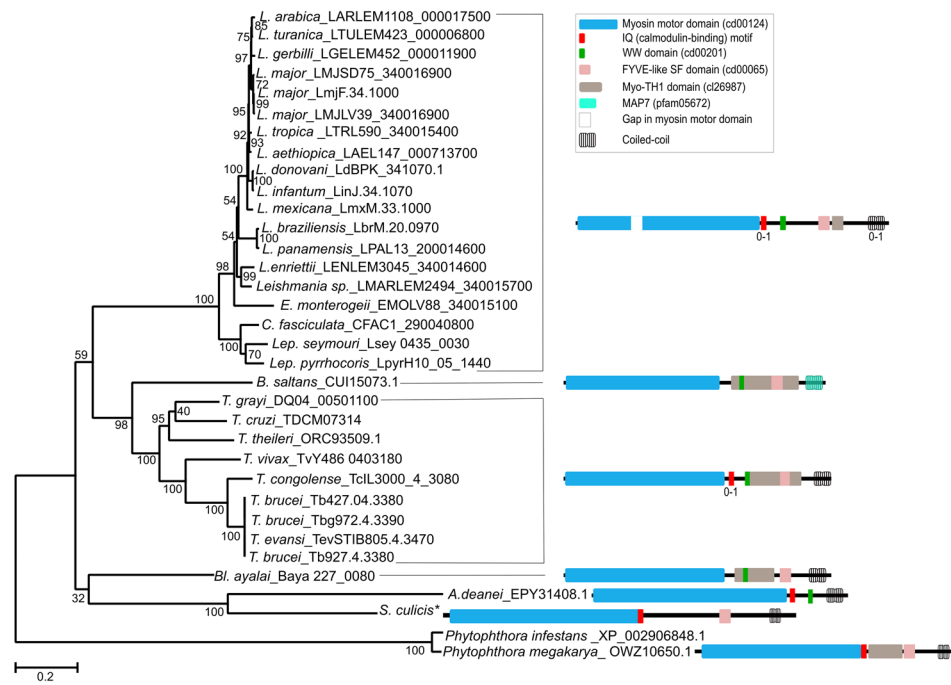


Figure 2. Phylogenetic tree of kinetoplastid Class I myosin proteins. The evolutionary history was inferred by using the maximum likelihood method based on the JTT + G (4 categories) model. Bootstrap values from 1,000 replicates are shown next to the branches. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 34 amino acid sequences. There were 1,185 positions in the final dataset. *Phytophthora* genus related myosins were used to root the tree. Sequences are identified by the name of species and the GenBank or TritypDB accession number. Asterisk represents the *S. culicis* Myo1 sequence obtained from the *S. culicis* TCC012E genome, contig coordinates AUXH01000328.1:350-4390. *T. rangeli* sequence (TRSC58_03135) was not included in the phylogeny because it is very short. Protein domain architectures of main sequences are shown.

myosin genes is expected to differ among species with reduced or expanded repertoires of myosin genes. A deeper discussion concerning Myo1 domains can be found in the Supplementary Discussion.

Myo13 is a widely distributed myosin in kinetoplastids, and *T. cruzi* myosins MyoA, MyoB, MyoC, MyoD, MyoE and MyoF can no longer be considered “orphans”. In the myosin global phylogeny presented by Odronitz and Kollmar (2007), it is clear that the *T. cruzi* orphan myosins MyoA, MyoB, MyoC, MyoD, MyoE and MyoF are related to Myo13 since they grouped together with high support⁵. We found orthologous sequences to these genes in some other species (Fig. 3).

The Myo13 gene has orthologues in almost all kinetoplastid species that were analysed and in conserved synteny. Sequences from most species are of similar size and domains composition, containing the myosin motor domain, followed by a coiled-coil region and two tandem UBA superfamily domains. For some species, additional less-significant domains were found. The Myo13 gene is absent in *T. congolense*, and no remnants of this gene were found in the syntenic region. It also appears absent in *T. rangeli* and the evaluation of the syntenic region for this species was not possible for Myo13 and all other genes, as the genome contains very small contigs.

MyoA gene was found in syntenic conserved regions relative to *T. cruzi* only in *T. grayi* and *T. theileri*. Unexpectedly, MyoA is also present in the distantly related species *B. saltans*, also in synteny. Most predicted proteins have a myosin motor domain, an IQ motif and two coiled-coil regions. No signal or remnants of the MyoA gene were found in the syntenic region from the other species analysed.

MyoB was also found in conserved synteny with the *T. cruzi* gene in *T. grayi*, *T. theileri* and *B. saltans*. The MyoB proteins contain the myosin head domain and can have a coiled-coil region and an IQ motif; other additional domains were also predicted for some species. *T. rangeli* has a truncated MyoB gene that could be due to poor genome assembly.

Interestingly, *B. saltans* presented four additional hits (CUG92185.1, CUG87167.1, CUG93737.1, CUG90281.1) in which reciprocal BlastP against *T. cruzi* showed MyoB as the best hit. Since these proteins do not cluster with other myosin clades in the phylogeny, they can be considered new myosin genes. The adopted nomenclature for these genes followed the nomenclature of kinetoplastid myosin genes (MyoI, CUG92185.1; MyoJ, CUG87167.1; MyoK, CUG93737.1 and MyoL, CUG90281.1; since MyoH was reserved for another gene that is discussed in a next section). These *B. saltans* proteins present the myosin head domain and two or three coiled-coil regions, and the MyoI protein has a predicted IQ motif. Overlapping the coiled-coil regions, some distinct and non-related domains were predicted with less significance.

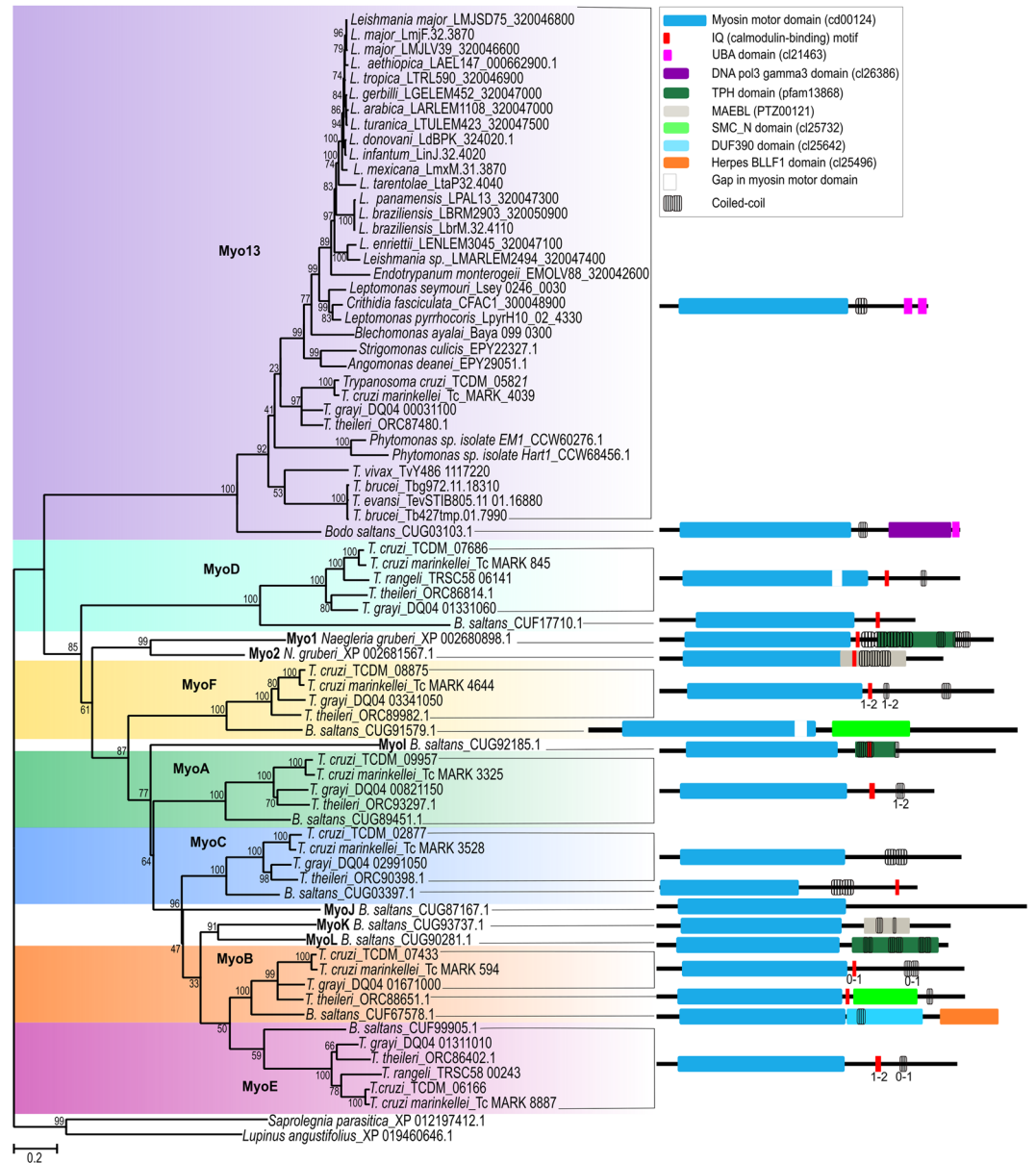


Figure 3. Phylogenetic tree of class XIII myosin proteins. The evolutionary history was inferred using the maximum likelihood method based on the LG + G (4 categories) + I model. Bootstrap values from 1,000 replicates are shown next to the branches. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 78 amino acid sequences. There were 621 positions of head domain in the final dataset. *Saprolegnia parasitica* (XP 012197412.1) and *Lupinus angustifolius* (XP 019460646.1) myosins were used to root the tree. Sequences are identified by the name of species and the GenBank or TritypDB accession number. Protein domain architectures of main sequences are shown.

The MyoC gene also has orthologues only in *T. grayi*, *T. theileri* and *B. saltans* in syntenic conserved regions compared to the *T. cruzi* gene. The predicted proteins have similar structures, containing a myosin motor domain and a coiled-coil region. The IQ motif was only predicted in *B. saltans* MyoC. No remnants of MyoC gene were found in the other species.

The MyoD gene was also found in *T. rangeli*, *T. grayi*, *T. theileri* and *B. saltans*. *T. grayi* and *T. theileri* MyoD genes are in syntenic conserved locations in relation to the *T. cruzi* gene, while the *B. saltans* gene location is not conserved. No remnants of MyoD were found in the syntenic regions from other species. *B. saltans* MyoD protein is shorter than the sequences from the other species, with a gap in the myosin head domain. IQ motifs and coiled-coil regions are predicted for some sequences.

T. cruzi MyoE orthologous genes were found in *T. rangeli*, *T. grayi*, *T. theileri* and *B. saltans*. Except for *T. rangeli*, the gene is in conserved synteny in all species. The predicted proteins have similar structure, containing a myosin motor domain and two IQ motifs. Additional coiled-coil regions are predicted in *T. rangeli* and *T. theileri*.

The MyoF gene also has orthologues in *T. grayi*, *T. theileri* and *B. saltans* in syntenic conserved regions compared to the *T. cruzi* gene. The proteins contain a myosin motor domain, one or two IQ motifs and two or three coiled-coil domains. A C-terminal truncated version of MyoF gene is also found in *T. rangeli*, encoding only part of the myosin head domain.

Excluding the kinetoplastids from NCBI BlastP searches and using all the queries (Myo13 and MyoA to MyoF), the two best hits correspond to sequences from the amoeba *Naegleria gruberi* (*Excavata, Heterolobosea class*) (named by us Myo1: XP_002680898.1 and Myo2: XP_002681567.1). In the myosin phylogeny presented by Odronitz and Kollmar⁵, these two *N. gruberi* genes were grouped inside of this kinetoplastid myosin clade. The following hits corresponded to myosin genes from different organisms, with alignments ranging from 40–70% of extension with approximately 35% identity. Due to the conservation of the myosin head domain and the huge number of myosin genes available at NCBI, numerous hits can be found with similar e-values. Some examples of hits in non-kinetoplastid species were cited in the Supplementary BlastP tables. However, except for the two *N. gruberi* myosins, the other sequences can be considered outgroup sequences since they do not locate inside the kinetoplastid myosin clade (Supplementary Figure S1).

To have a better view of evolution and relationship among all these genes, we inferred a phylogenetic tree based on the most conserved regions of the myosin head domain (621 positions), which is shown in Fig. 3. Two non-kinetoplastid sequences retrieved from BLAST in the other species were used as outgroup. The phylogeny shows significant support for most of the important clades, and the relationships among myosin groups are congruent to those presented by Odronitz and Kollmar⁵. Myo13 was the first clade to branch, followed by the MyoD clade, *Naegleria* myosins, MyoF clade, *B. saltans* MyoI, MyoA clade, MyoC clade, *B. saltans* MyoJ, MyoK and MyoL, and finally the MyoB and MyoE clades. This topology reflects the pattern of gene duplication that will be discussed later.

Analysing the relationships inside Myo groups, we observed some incongruities in the Myo13 clade in relation to the species tree (Fig. 1), such as wrong positioning of some *Leishmania* species and *Bl. ayalai*, the clustering of *Lep. pyrrocoris* with *C. fasciculata*, and in the base of the tree, *Trypanosoma* species were separated into two clades interspersed by *Phytomonas* sp. sequences. Different rates of evolution among sequences or long-branch attraction could explain these incongruities. The relationships found among sequences from MyoA to MyoF clades are in congruence with the species tree (except for the absence of sequences in some species), with *B. saltans* sequences at the base of clades, but the positioning of *T. theileri* changes in some groups. The positioning of this species in the *Trypanosoma* genus is not clear^{20–22}.

Class XIII myosins were described earlier as an exclusive kinetoplastid class comprising only Myo13, being the myosins MyoA to MyoF from *T. cruzi* considered “orphans”⁵. We found that MyoA, MyoB, MyoC, MyoD, MyoE and MyoF have orthologous genes in other kinetoplastid species that were not previously analysed, grouping together with Myo13 and, therefore, they were considered as Class XIII myosins (Fig. 3). In addition, we found that *B. saltans* has an even greater and unique repertoire of class XIII myosin genes, since the four extra genes (MyoI to MyoL) do not cluster with the other myosin clades and were positioned inside the phylogeny, before the last branching, suggesting that these sequences were also present in the ancestor species. Despite the clear relationship between MyoK and MyoL, their divergence suggests ancient duplication.

The positioning of *B. saltans* Myo13, MyoD, MyoF, MyoA, MyoC, Myo B and MyoE in the base of clades grouping with high support indicates that all these genes were present in the common ancestor of trypanosomatids and *B. saltans*, being transmitted vertically. Horizontal transfer among species could be discarded due to the synteny conservation. *Naegleria* myosin sequences do not group with any kinetoplastid myosin group, but they are clearly related, branching after MyoD separation. Independent secondary loss events should be postulated to explain the pattern of gene absence. Myo13 was lost in *T. congolense* and *T. rangeli*; the other class XIII myosin genes present very similar patterns of loss in the ancestor of the *T. vivax/T. congolense/T. brucei/T. evansi* lineage and in the ancestor of the *Leishmania* clade. The absence of any remnants of these genes in the syntenic regions supports the idea of long-term loss events. If we consider that the *T. rangeli* genome is complete, the MyoA and MyoC genes were also lost in this species.

The primary origin of Class XIII myosin is not clear since it lacks a significant phylogenetic relationship to other myosin genes. It could be a result of a very ancient horizontal transfer event from an unidentified source to an ancient ancestor species. Alternatively, the ancestor Class XIII myosin may have originated by mutations of a redundant duplicated myosin gene. The addition of myosin sequences from new species that sample both Kinetoplastida and Heterolobosea will certainly help to clarify this point.

Clues of class XIII myosin functions come mainly from *L. major* Myo13 studies, where it seems to be related to the assembly of flagellum^{19,23}. Myo13 protein is expressed both in promastigote and amastigote parasite forms, besides its increased expression in promastigotes, where it localizes preferentially at the proximal region of the flagellum and colocalizes with paraflagellar rod (PFR) proteins²³. In a later work, Katta *et al.*¹⁹ showed that *L. major* Myo13 associates with actin and is involved in intraflagellar transport (IFT), since knockout cells presented reduced flagellar length and absence of PFR. LmMyo13 gene appears to be essential for parasite survival whereas the double mutant generated Myo13 ploidy¹⁹. Additionally, the IQ motif predicted in Myo13 genes was analysed in *Leishmania* and proven to be functional, with the binding of calmodulin regulating dimerization, motility and lipid binding of this molecular motor^{24,25}. However, in *T. brucei* there is no indication of Myo13 function, since its knockdown in bloodstream forms had no effect on vesicle traffic or growth¹⁸.

MyoG protein has evolved to a new protein in Leishmaniinae subfamily. MyoG was considered an “orphan” myosin since it was previously found only in *T. cruzi*⁵. Using TritypDB and NCBI BlastP searches, we found orthologues in the closely related species *T. grayi* and *T. theileri* and in *B. saltans*. Analysis of *T. grayi* and *T.*

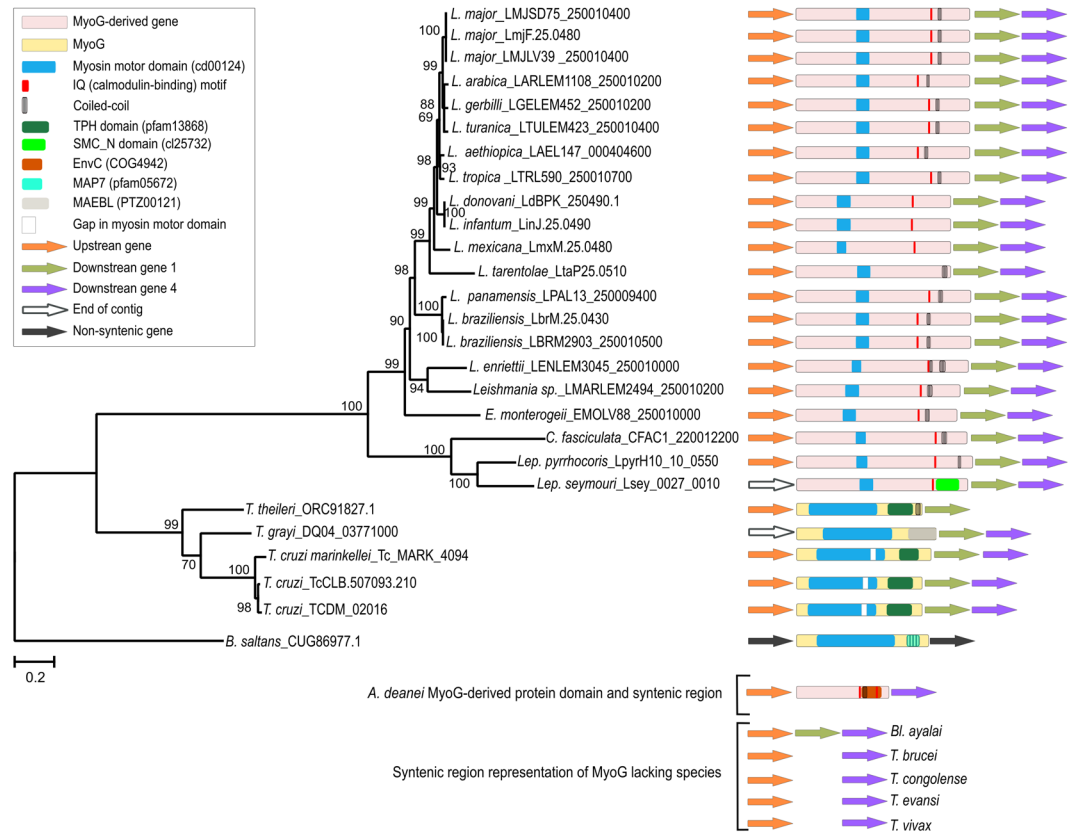


Figure 4. Phylogenetic tree of kinetoplastid MyoG and MyoG-derived proteins. The evolutionary history was inferred by using the maximum likelihood method based on the JTT + G (4 categories) + I model. Bootstrap values from 1,000 replicates are shown next to the branches. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 28 amino acid sequences. There were 568 positions in the final dataset. *B. saltans* MyoG was used to root the tree. Sequences are identified by the name of species and the GenBank or TritypDB accession number. Protein domain architectures of sequences are shown together with a representation of their genomic context. Due to the high divergence, the *A. deanei* MyoG-derived protein was not included in the phylogeny.

theileri genomes indicated that the gene is in a syntenic conserved location in relation to the *T. cruzi* MyoG. The products of these genes present similar patterns of domains as *T. cruzi* MyoG.

For *Leishmania* and *Leptomonas* species, *C. fasciculata*, *E. monterogei* and *A. deanei*, the two best BlastP hits represent Myo13 and Myo1 genes. Interestingly, a third hit with a higher e-value was also found in these species. These hits correspond to genes annotated as “conserved hypothetical protein”. We then identify that these genes are in syntenic locations in relation to *T. cruzi* MyoG in all these species (Fig. 4). Their protein products have similarity to MyoG at the C-terminal region (last 400 aa) and in a short region (only 130 aa) of the motor domain. For the *A. deanei* sequence, however, the similarity is limited to a very short region at the C-terminal portion (approximately 240 aa) and there is no remnants of the myosin motor domain. Clearly, these sequences represent a MyoG-derived gene that had accumulated numerous mutations, almost losing the motor domain. The low conservation of the myosin head domain in MyoG-derived proteins likely prevented the authors from finding it in *L. major* in the previous works^{3,5,8,9}. Additionally, the MyoG-derived proteins have a predicted coiled-coil region and an IQ motif.

Bl. ayalai, *T. brucei*, *T. evansi*, *T. vivax*, *S. culicis* and *Phytomonas sp.* only presented BlastP hits corresponding to Myo13 and Myo1, and *T. congolense* only to Myo1. To evaluate the possible loss event of MyoG in these species, we analysed the syntenic regions where the gene should be located and we found no remnants of MyoG in these species. Moreover, *T. brucei*, *T. evansi* and *T. congolense* also lost the downstream gene. *T. vivax* has a large region of missing data in this location; nonetheless, it is plausible that the loss of the genes occurred in the ancestor of these species. *T. rangeli* apparently lost the MyoG gene.

To better understand the evolution of MyoG and MyoG-derived genes, their conserved amino acid regions (568 positions) were used to infer a phylogeny (Fig. 4). The *A. deanei* MyoG-derived protein was not included due to the high divergence that made the alignment non-confident. The MyoG tree shows high supported nodes, and its topology largely reflects the relationships of species (Fig. 1). *B. saltans* was the first species to branch, followed by the split of the *Trypanosoma* genus and *Leishmania* clade.

Excluding the kinetoplastids in the NCBI BlastP searches, the best hits were from several distinct organisms, including fungi, insects, fishes and others, presenting e-values around 1E-60 (approximately 30% identity covering 50–70% of query). Reciprocal BlastP showed that other myosins are more similar to these sequences than MyoG. This finding is in accordance with the data of Odronitz and Kollmar⁵ since in the myosin phylogeny presented by these authors the *T. cruzi* MyoG does not group with good support to any other myosin. Thus, the kinetoplastid MyoG genes comprise a new class that we named Class XXXVI, a class that was not used in the recent works of myosin classification^{5,9}.

The origin of MyoG in Kinetoplastida is unknown and could be a result of a very ancient horizontal transfer event or originated by mutations of a redundant duplicated myosin gene. In this case, we need to suppose that the gene underwent extremely high rates of mutations that make it impossible to track the relationship with its precursor. After the primary origin, it is presumed that the MyoG gene was vertically transmitted throughout trypanosomatids evolution. Although there is a discontinuity in the gene presence among species and absence of a conserved synteny region in *B. saltans*, the divergence among sequences does not support a horizontal transfer hypothesis. Nevertheless, considering the picture that we have now, we need to suggest several events of MyoG loss during species evolution: in the ancestor of *T. brucei*, *T. evansi* and *T. congolense*, in *T. rangeli*, in *Bl. ayalai*, in *Phytomonas sp.* and in *S. culicis*. The function of MyoG and MyoG-derived genes remains to be studied.

MyoH-derived: a possible myosin-derived gene. In Myo13 BlastP searches, we also found a less significant hit (e-value > E-10) that caught our attention in *T. cruzi*, *T. rangeli* and *T. grayi*. In *T. cruzi*, this hit corresponded to the gene TCDM_02145, annotated as “hypothetical protein.” It has a very short conserved region (only 144 aa) of the myosin domain (MYSc_Myo13, cd14875). In the C-terminal portion of the protein, it contains a predicted TPH domain (pfam13868). As for MyoG, overlapping this domain there is a Neuromodulin_N superfamily (cl26511), a SMC_N superfamily (cl25732) and an ERM superfamily (cl25742) domains. Despite the similarity of domain composition between MyoG and this protein, their alignment on BlastP-2-sequences was not significant (e-value 0.087), suggesting they are not clearly related. However, the similarity between *T. cruzi* TCDM_02145 and Myo13 proteins was significant (e-value 3E-10 on BlastP-2-sequences), but restricted to the short region of myosin head domain (25% identity and 46% similarity over 145 aa). The similarity was more evident when we compared the *T. rangeli* protein (TRSC58_00453) with the Myo13 protein from *L. panamensis* (LPAL13_3200047300), since we observed a good alignment in the first 430 aa (24% identity, 39% similarity, e-value 3E-10 on BlastP-2-sequences).

Due to the presence of a remnant of a myosin head domain and the similarity with Myo13, we conclude that this gene had derived from an ancient class XIII myosin gene. However, unlike what occurred to MyoG-derived genes, we could not find any copy with an intact myosin head domain. Following the nomenclature of *T. cruzi* myosin genes, we assumed that the mother gene was an ancestral MyoH gene; thus, we called it MyoH-derived gene. We then conducted BlastP searches using the *T. cruzi* MyoH-derived protein as the query, and the gene was also found in *T. theileri* and *B. saltans* and synteny conservation was found for all MyoH-derived carrier species. When we excluded the kinetoplastids from NCBI BlastP searches, only one hit was found with low significance matching a non-myosin protein from *Branchiostoma belcher* (E = 0.89; XP_019640798.1). Since MyoH-derived proteins accumulated higher divergence in relation to Myo13 and the other myosins (that could be considered outgroups), the alignment of sequences were very poor with low confidence; thus, we did not construct a MyoH-derived phylogenetic tree. The conversion of MyoH gene to MyoH-derived gene must have occurred in the ancestor of *B. saltans* and trypanosomatids. Interestingly, the MyoH-derived gene has the same pattern of loss as other class XIII myosins.

Kinetoplastid myosins expansion, neofunctionalization and secondary losses. Several molecular mechanisms are known to be involved in the creation of new gene structures, such as exon shuffling, gene duplication, retroposition, horizontal transfer, domestication of mobile elements, gene fusion/fission and *de novo* origination²⁶. The class XIII myosins are clear results of gene duplications and at least 13 very ancient events need to be postulated to explain their diversification, showing that the expansion of this class had an impact on kinetoplastid evolution.

A possible scenario of gene duplications is postulated in Fig. 5A. The first duplication events (1 and 2) probably led to the MyoH and Myo13 genes, since they are related and Myo13 is the first clade to branch in the phylogeny. Those are followed by the duplications that generated the genes MyoD (3), the *Naegleria* myosin ancestor gene (4), MyoF (5), MyoI (6), MyoA (7), MyoC (8), MyoJ (9) and finally the duplication generating the ancestor of MyoK and MyoL and the ancestor of MyoE and MyoB (10). The order of duplications 11, 12 and 13 are hard to determine.

The positioning of *Naegleria* sequences in the phylogeny indicates that duplication events until this point must have occurred in the ancestor of Kinetoplastida and *Naegleria* species, as shown in Fig. 5B. In this scenario, Heterolobosea species have lost previously duplicated genes. Alternatively, to explain the phylogeny, all duplication events could have occurred in the ancestor of Kinetoplastida, but after its separation from the ancestor of Heterolobosea (Fig. 5C). Then, an ancient horizontal transfer event of a Class XIII myosin from a kinetoplastid to a *Naegleria* ancestor could explain this relationship.

Duplicated genes could have different fates, such as pseudogenization, conservation of function, subfunctionalization or neofunctionalization^{27,28}. As the gene duplication generates functional redundancy, one of the copies is free to change, and the accumulation of mutations may transform the gene in a pseudogene that can be deleted after a long time²⁸. We were not able to find any clear case of myosin pseudogene, corroborating the idea that the identified myosin gene deletions occurred long ago. Duplicated genes can otherwise be conserved with the same function, maintained by concert evolution or purifying selection, if there is an advantage in producing extra amounts of proteins or RNA products^{27,28}. This seems not be the case for kinetoplastid myosins, since they

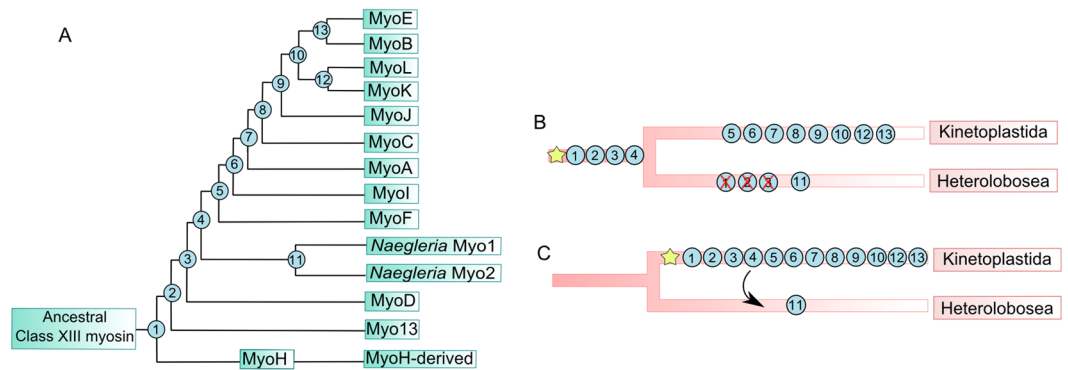


Figure 5. Possible evolutionary scenario of class XIII myosin genes duplications. **(A)** Representation of 13 putative duplication events (numbered circles) from an ancestral class XIII myosin gene with an unknown primary origin. The order of duplications from 2 to 10 is supported by the phylogeny, while the others are not clear. **(B)** Possible scenario of evolution where the origin of the first class XIII myosin (yellow star) and duplications 1, 2, 3 and 4 occurred before separation of Kinetoplastida and Heterolobosea classes. In this case, some genes were lost in Heterolobosea (circles filled with red “X”). **(C)** Possible scenario of evolution where the origin of the first class XIII myosin (yellow star) and almost all duplication events occurred after separation of the Kinetoplastida and Heterolobosea classes. The presence of two copies of this myosin class in *Naegleria* can be explained by an ancient horizontal transfer event (black arrow) from a kinetoplastid to a Heterolobosea member followed by an event of duplication.

present significant divergence in amino acid sequence and structure. Except for a few cases, the conservation of function is unlikely to be maintained.

The subfunctionalization occurs when each daughter gene adopts part of the functions of the parental gene^{28,29}. This could be the case for class XIII myosins. Due to the vast repertoire found in some species, it is possible to suggest that the original function of ancestral myosin was distributed to the derived genes. However, this hypothesis should be accompanied by a reversion of this scenario, since only one class XIII gene was maintained in several species.

Alternatively, it is possible that those myosin gene duplications gave rise to novel functions, a process called neofunctionalization, and in many cases, a related function evolves after gene duplication instead of a completely new function²⁸. The conservation of the myosin head domain in most cases suggests that the function was partially conserved, and the diversified C-terminal tails may indicate specializations for different cargos/localizations. In this context, the high divergence among class XIII myosin sequences is expected since the evolutionary rate after duplication is predicted to be accelerated because of positive selection and functional diversification³⁰.

On the other hand, there are several examples of entirely new function created by gene duplication²⁸. It is reasonable that MyoH-derived genes have a non-myosin function, since the key domain conferring the properties of actin binding and ATP hydrolysis was almost lost. Additionally, although the MyoG is not a direct product of gene duplication, the probable redundancy of functions provided by the other myosin genes allowed the MyoG gene to accumulate mutations and, therefore, MyoG-derived genes also likely play a non-myosin function.

Our data shows that the ancestor of *B. saltans* and trypanosomatids had all class XIII myosins along with Myo1, MyoG and the MyoH-derived gene (Fig. 1). Most genes were found in the free-living kinetoplastid *B. saltans* and in *T. cruzi* and related *Stercoraria* species indicating that these genes were preserved in these species for more than 500 million years. Additionally, almost all myosin genes were found in conserved synteny, contrary to what is generally found, since the conservation of gene order between *B. saltans* and trypanosomatids is low, making up approximately 9% of the genes³¹. This finding raises the question of what is the function of these genes and why they were conserved only in those species.

The pattern of gene loss also seen for the myosin genes during the trypanosomatids evolutions is in agreement to a recent work that shows that trypanosomatids present streamlined genomes consistent with loss of redundancy that could be associated with the alteration from an ancestral free-living state inhabiting diverse environments to an obligate parasite that explores relatively constant host environments³¹. The authors also observed several genes that are unique to one or more parasites, indicating that gene gain also has a significant role in the origin of parasitism, likely via the rapid evolution of multi-copy gene families. Thus, we can predict that other redundant or useless genes, such as MyoG and MyoH, could have evolved new functions, and, rather than simply being lost, they can serve as a raw material for the evolution of new genes.

Conclusions

In this work, we present the largest and most comprehensive evolutionary analysis of Kinetoplastida myosin genes, identifying interesting aspects of this family evolution such as its expansion by gene duplication events, the potential of myosin family as a source of new genes and the extensive gene loss of several members in a great number of trypanosomatid species. This work is an important beginning for future functional studies to comprehend the diverse roles of myosin proteins in Kinetoplastida species.

Methods

Myosin amino acid sequences from the *T. cruzi* Dm28c strain were used as query on BlastP searches³² against genomes available at TritypDB³³ up to July 2017. Information about genomes is available in Supplementary Table S2. BlastP searches against non-redundant protein sequences (nr) from GenBank^{32,34} were also conducted to determine group sequences and some kinetoplastid sequences that are not available at TritypDB. Since the motor domain is a well-conserved domain of a diversified protein family, BlastP searches provide several significant hits (e-value cutoff of E-10). Thus, sequences were considered orthologues when genes were in the same genomic context (conserved syntenic location) and/or if the query sequences appeared as the first hit on a reciprocal BlastP.

Syntenic conservation was evaluated using the Genome Browser tool for species available at TritypDB. For those available at NCBI (Supplementary Table S2), genome sequences were downloaded and submitted to local TBlastN³² using as query the myosin amino acid sequences or the amino acid sequences of upstream and downstream genes. The TcruziDm28cPB1 genome assembly (GCA_002219105.1) was also used for some specific gene searches. A Perl programming code was used to retrieve desired sequences. If the myosin genes were not present/annotated, the syntenic regions where the genes were supposed to be present were analysed by BlastX against the annotated proteins (e-value cutoff E-5) searching for non-annotated ORFs or remnants of the genes.

CD-search on NCBI Conserved Domain Database (CDD)³⁵, SMART (Single Modular Architecture Research Tool)³⁶ and ELM (The Eukaryotic Linear Motif resource)³⁷ were used to find protein domains and motifs using default parameters.

Amino acid alignments were obtained by PSI-Coffee³⁸. Alignments were visualized and edited using the Genedoc 2.7 package³⁹. To avoid loss of phylogenetic signal, phylogenies from the three groups were constructed separately. For Myo1 and MyoG, well-aligned regions from entire proteins were selected based on PSI-Coffee alignment scores. For class XIII myosins, well-aligned regions were also selected but restricted to the myosin head domain. Phylogenies for each protein group were constructed by maximum likelihood (ML) method, using the amino acid substitution model indicated by the Model Selection. Reliability of branches was accessed by bootstrap test with 1,000 replicates. Phylogenetic analyses were implemented in MEGA 7.0 software⁴⁰.

To understand the evolution of myosin genes, we compared their phylogenies to the known trypanosomatid phylogenetic relationships. Thus, Fig. 1 represents a consensus phylogenetic tree of trypanosomatids modified from Ludwig and Krieger⁴¹ to encompass the species used in this work and is based on well-supported trees from several studies. Positioning of *T. theileri* is inconclusive from works available up to now^{20–22} and is represented as a polytomy. For evolutionary inferences, we adopted the most parsimonious assumptions; however, we did not discard alternative hypotheses and/or further updates with studies on additional species.

References

- Vale, R. D. The molecular motor toolbox for intracellular transport. *Cell* **112**, 467–480 (2003).
- Woolner, S. & Bement, W. M. Unconventional myosins acting unconventionally. *Trends Cell Biol.* **19**, 245–252 (2009).
- Foth, B. J., Goedecke, M. C. & Soldati, D. New insights into myosin evolution and classification. *Proc. Natl. Acad. Sci. USA* **103**, 3681–3686 (2006).
- Krendel, M. & Mooseker, M. S. Myosins: tails (and heads) of functional diversity. *Physiology (Bethesda)*. **20**, 239–251 (2005).
- Odonitz, F. & Kollmar, M. Drawing the tree of eukaryotic life based on the analysis of 2,269 manually annotated myosins from 328 species. *Genome Biol.* **8**, R196 (2007).
- Sellers, J. R., Goodson, H. V. & Wang, F. A myosin family reunion. *J. Muscle Res. Cell Motil.* **17**, 7–22 (1996).
- Hodge, T. & Cope, M. J. A myosin family tree. *J. Cell Sci.* **113**, 3353–3354 (2000).
- Richards, T. A. & Cavalier-Smith, T. Myosin domain evolution and the primary divergence of eukaryotes. *Nature* **436**, 1113–1118 (2005).
- Sebé-Pedrós, A., Grau-Bové, X., Richards, T. A. & Ruiz-Trillo, I. Evolution and classification of myosins, a paneukaryotic whole-genome approach. *Genome Biol. Evol.* **6**, 290–305 (2014).
- Thompson, R. F. & Langford, G. M. Myosin superfamily evolutionary history. *Anat. Rec.* **268**, 276–289 (2002).
- Moreira, D., López-García, P. & Vickerman, K. An updated view of kinetoplastid phylogeny using environmental sequences and a closer outgroup: proposal for a new classification of the class Kinetoplastea. *Int. J. Syst. Evol. Microbiol.* **54**, 1861–1875 (2004).
- Vickerman, K. In *Biology of the Kinetoplastida* (eds. Lumsden, W. H. R. & Evans, D. A.) 1–34 (Academic Press, 1976).
- Kaufer, A., Ellis, J., Stark, D. & Barratt, J. The evolution of trypanosomatid taxonomy. *Parasit. Vectors* **10**, 287 (2017).
- Nussbaum, K., Honek, J., Cadmus, C. M. Cv. C. & Efferth, T. Trypanosomatid parasites causing neglected diseases. *Curr. Med. Chem.* **17**, 1594–1617 (2010).
- Lukes, J., Skalický, T., Týč, J., Votýpka, J. & Yurchenko, V. Evolution of parasitism in kinetoplastid flagellates. *Mol. Biochem. Parasitol.* **195**, 115–122 (2014).
- Porcel, B. M. *et al.* The streamlined genome of *Phytomonas* spp. relative to human pathogenic kinetoplastids reveals a parasite tailored for plants. *PLoS Genet.* **10**, e1004007 (2014).
- Telford, M. J. & Copley, R. R. Animal Phylogeny: Fatal Attraction. *Curr. Biol.* **15**, R296–R299 (2005).
- Spitznagel, D., O'Rourke, J. F., Leddy, N., Hanrahan, O. & Nolan, D. P. Identification and characterization of an unusual class I myosin involved in vesicle traffic in *Trypanosoma brucei*. *PLoS One* **5**, e12282 (2010).
- Katta, S. S., Tamma, T. V. S., Sahasrabudhe, A. A., Bajpai, V. K. & Gupta, C. M. Trafficking activity of myosin XXI is required in assembly of *Leishmania flagellum*. *J. Cell Sci.* **123**, 2035–2044 (2010).
- Lima, L. *et al.* *Trypanosoma livingstonei*: a new species from African bats supports the bat seeding hypothesis for the *Trypanosoma cruzi* clade. *Parasit. Vectors* **6**, 221 (2013).
- Fraga, J. *et al.* Phylogenetic analysis of the *Trypanosoma* genus based on the heat-shock protein 70 gene. *Infect. Genet. Evol.* **43**, 165–172 (2016).
- Hamilton, P. B., Gibson, W. C. & Stevens, J. R. Patterns of co-evolution between trypanosomes and their hosts deduced from ribosomal RNA and protein-coding gene phylogenies. *Mol. Phylogenet. Evol.* **44**, 15–25 (2007).
- Katta, S. S., Sahasrabudhe, A. A. & Gupta, C. M. Flagellar localization of a novel isoform of myosin, myosin XXI, in *Leishmania*. *Mol. Biochem. Parasitol.* **164**, 105–110 (2009).
- Batters, C., Woodall, K. A., Toseland, C. P., Hundscheil, C. & Veigel, C. Cloning, expression, and characterization of a novel molecular motor, *Leishmania myosin-XXI*. *J. Biol. Chem.* **287**, 27556–27566 (2012).

25. Batters, C. *et al.* Calmodulin regulates dimerization, motility, and lipid binding of Leishmania myosin XXI. *Proc. Natl. Acad. Sci. USA* **111**, E227–236 (2014).
26. Long, M., Betrán, E., Thornton, K. & Wang, W. The origin of new genes: glimpses from the young and old. *Nat. Rev. Genet.* **4**, 865–875 (2003).
27. Magadum, S., Banerjee, U., Murugan, P., Gangapur, D. & Ravikesavan, R. Gene duplication as a major force in evolution. *J. Genet.* **92**, 155–161 (2013).
28. Zhang, J. Evolution by gene duplication: an update. *Trends Ecol. Evol.* **18**, 292–298 (2003).
29. Nowak, M. A., Boerlijst, M. C., Cooke, J. & Smith, J. M. Evolution of genetic redundancy. *Nature* **388**, 167–171 (1997).
30. Pegueroles, C., Laurie, S. & Albà, M. M. Accelerated evolution after gene duplication: a time-dependent process affecting just one copy. *Mol. Biol. Evol.* **30**, 1830–1842 (2013).
31. Jackson, A. P. *et al.* Kinetoplastid Phylogenomics Reveals the Evolutionary Innovations Associated with the Origins of Parasitism. *Curr. Biol.* **26**, 161–172 (2016).
32. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
33. Aslett, M. *et al.* TriTrypDB: a functional genomic resource for the Trypanosomatidae. *Nucleic Acids Res.* **38**, D457–462 (2009).
34. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Wheeler, D. L. GenBank. *Nucleic Acids Res.* **33**, D34–8 (2005).
35. Marchler-Bauer, A. *et al.* CDD/SPARCLE: Functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res.* **45**, D200–D203 (2017).
36. Letunic, I., Doerks, T. & Bork, P. SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Res.* **40**, D302–305 (2012).
37. Dinkel, H. *et al.* ELM 2016—data update and new functionality of the eukaryotic linear motif resource. *Nucleic Acids Res.* **44**, D294–300 (2016).
38. Floden, E. W. *et al.* PSI/TM-Coffee: a web server for fast and accurate multiple sequence alignments of regular and transmembrane proteins using homology extension on reduced databases. *Nucleic Acids Res.* **44**, W339–343 (2016).
39. Nicholas, K. B., Nicholas, H. B. J. & Deerfield, D. W. I. Genedoc: a tool for editing and annotating multiple sequence alignments. *EMBNEW.NEWS* **4** (1997).
40. Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol. Biol. Evol.* **33**, 1870–1874 (2016).
41. Ludwig, A. & Krieger, M. A. Genomic and phylogenetic evidence of VIPER retrotransposon domestication in trypanosomatids. *Mem. Inst. Oswaldo Cruz* **111**, 765–769 (2016).
42. Flegontov, P. *et al.* Genome of *Leptomonas pyrrocoris*: a high-quality reference for monoxenous trypanosomatids and new insights into evolution of *Leishmania*. *Sci. Rep.* **6**, 23704 (2016).
43. Teixeira, M. M. G. *et al.* Phylogenetic Validation of the Genera *Angomonas* and *Strigomonas* of Trypanosomatids Harboring Bacterial Endosymbionts with the Description of New Species of Trypanosomatids and of Proteobacterial Symbionts. *Protist* **162**, 503–524 (2011).
44. Flegontov, P. *et al.* Paratrypanosoma is a novel early-branching trypanosomatid. *Curr. Biol.* **23**, 1787–1793 (2013).
45. Harkins, K. M., Schwartz, R. S., Cartwright, R. A. & Stone, A. C. Phylogenomic reconstruction supports supercontinent origins for *Leishmania*. *Infect. Genet. Evol.* **38**, 101–109 (2016).
46. Du, Y., Maslov, D. A. & Chang, K. P. Monophyletic origin of beta-division proteobacterial endosymbionts and their coevolution with insect trypanosomatid protozoa *Blastocrithidia culicis* and *Crithidia* spp. *Proc. Natl. Acad. Sci. USA* **91**, 8437–8441 (1994).
47. Kelly, S., Ivens, A., Manna, P. T., Gibson, W. & Field, M. C. A draft genome for the African crocodylian trypanosome *Trypanosoma grayi*. *Sci. Data* **1**, 140024 (2014).
48. Stoco, P. H. *et al.* Genome of the Avirulent Human-Infective Trypanosome—*Trypanosoma rangeli*. *PLoS Negl. Trop. Dis.* **8**, e3176 (2014).

Acknowledgements

We are very grateful to Maykol Adrian Noll for the support with Perl programming code and to Instituto Carlos Chagas (ICC/Fiocruz) and its graduation programme “Programa de Pós-Graduação em Biociências e Biotecnologia” for support. We also thank Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and Instituto de Biologia Molecular do Paraná (IBMP) for fellowships. MAK is a research fellow from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Author Contributions

D.A.S.S. and A.L. design the research, performed the Blast searches and phylogenetic analyses, analysed the results and wrote the manuscript; M.A.K. and D.P.P. assisted in analysis of the results and manuscript writing. All authors read and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at <https://doi.org/10.1038/s41598-017-18865-y>.

Competing Interests: The authors declare that they have no competing interests.

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) 2018