Marine gregarine genomes illuminate current understanding of apicomplexan glideosome

Author List : Julie Boisard[1,2], Evelyne Duvernois-Berthet[3], Linda Duval[1], Joseph Schrével[1], Laure Guillou[4], Amandine Labat[1], Sophie Le Panse[5], Gérard Prensier[6], Loïc Ponger[2] and Isabelle Florent[1].

Author Contact:

Julie Boisard: julie.boisard@edu.mnhn.fr

Isabelle Florent: isabelle.florent@mnhn.fr

Loïc Ponger: loic.ponger@mnhn.fr

Lead Contact:

Isabelle Florent

Author Affiliation:

1. Molécules de Communication et Adaptation des Microorganismes (MCAM, UMR 7245 CNRS), Département Adaptations du vivant (AVIV), Muséum National d'Histoire Naturelle, CNRS, CP 52, 57 rue Cuvier, 75231 Paris Cedex 05, France
2. Structure et instabilité des génomes (STRING UMR 7196 CNRS/INSERM U1154), Département Adaptations du vivant (AVIV), Muséum National d'Histoire Naturelle, CNRS, INSERM, CP 26, 57 rue Cuvier, 75231 Paris Cedex 05, France
3. Muséum National d'Histoire Naturelle, Centre National de la Recherche Scientifique, Laboratoire Physiologie Moléculaire et Adaptation (PhyMA), UMR7221 CNRS-MNHN, 75005, Paris, France
4. Sorbonne Université, CNRS, UMR7144 Adaptation et Diversité en Milieu Marin, Ecology of Marine Plankton (ECOMAP), Station Biologique de Roscoff SBR, 29680 Roscoff, France
5. Plateforme d'Imagerie Merimage, FR2424, Centre National de la Recherche Scientifique, Station Biologique de Roscoff, Roscoff, 29680, France
6. Cell biology and Electron Microscopy Laboratory, François Rabelais University, 10 Boulevard Tonnellé, BP 3223 Tours Cedex, France

SUMMARY

Apicomplexans, parasite protists of a very wide diversity of metazoan hosts, are mostly known from species infecting human. Absence or limited data for basal lineages prevents a comprehensive view of evolutionary history and adaptive capacities of Apicomplexa. Here, we characterized the genome of the marine eugregarine *Porospora gigantea*, remarkable for the gigantic size of its vegetative feeding forms (trophozoites) and their speed of gliding movement, the fastest so far recorded for an Apicomplexa. Not a single but two highly related genomes named A and B were assembled. Highly syntenic, of similar size (9 Mb) and coding capacities (~5300 genes), they display a 10.8% divergence at nucleotide level corresponding to 16-38 My divergent time. Orthogroups analyses across 25 (proto)Apicomplexa including *Gregarina niphandrodes* showed that A and B are highly divergent from all other known apicomplexan species, revealing an unexpected diversity. These two related species branch in phylogenetic studies at the base of Cephaloidophoroidea, forming a new family in these crustacean gregarines. Gliding proteins data mining found a strong conservation of actin-related proteins, as well as of regulatory factors, within apicomplexan. In contrast, the conservation of core glideosome proteins and adhesion proteins appears to be highly variable among apicomplexan lineages, especially in gregarines. These results confirm the importance of studying gregarines to widen our biological and evolutionary view of apicomplexan parasites, to better apprehend species diversity and revise our understanding of the molecular bases of some key functions such as observed for the glideosome.

Key words: Apicomplexa, marine gregarine, genome assembly, comparative genomics, gliding, phylogeny

INTRODUCTION

Apicomplexan are unicellular eukaryotic microorganisms that have evolved towards a strict parasitic lifestyle. Some species are extremely pathogenic such as *Plasmodium* spp., *Toxoplasma gondii* and *Cryptosporidium* spp., responsible for malaria, toxoplasmosis and cryptosporidiosis. While the genomic knowledge of apicomplexan parasites currently concerns a dozen of genera, more precisely those of highly pathogenic agents[1], it remains highly skewed towards intracellular parasites of vertebrates, notably Coccidia, Hemosporidia and *Cryptosporidium* (see references in Table S1). Yet, the Apicomplexa include about 350 genera[2] for 6,000 documented species. This includes gregarines, which represent early diverging lineages with low pathogenicity that have attracted less interest from biomedical research. However, gregarines include at least 1,770 species[3], their diversity being highly understated, as gregarines were identified as the most abundant and widely

reported apicomplexan in a recent environmental study[4]. As they remain non-cultivable organisms, their study in the laboratory is technically challenging, explaining why gregarines have hardly been explored at the -omic levels[5].

Yet, ignoring gregarines not only hides a part of the evolutionary history of Apicomplexa, but also deprives access to a diversity of possibly alternative molecular mechanisms, that sustain some specific adaptive traits of this group. For instance, gregarines are mostly extracellular, infecting a wide diversity of marine and terrestrial non-vertebrate hosts[6,7]. Still, available genomic data are very limited (partial data on *Ascogregarina taiwanensis,* intestinal parasite of *Aedes albopictus*[8], draft genome of *Gregarina niphandrodes,* intestinal parasite of *Tenebrio molitor* (unpublished, available in CryptoDB[9]), and concern only terrestrial gregarines. Transcriptomic studies on trophozoite feeding stages of terrestrial and marine gregarine species have recently provided important insights[10–13], especially about organellar genomes and metabolic pathways, but they cannot account for the whole genetic landscape of gregarines, as developmental stages are very distinctive, nor can they tell us about their genome structure.

To overcome this lack of data, we focused on the marine eugregarine *Porospora gigantea* (van Beneden, 1869) Schneider, 1875, an intestinal parasite of lobster *Homarus gammarus. Porospora gigantea* was described in 1869 by E. van Beneden, who called it *Gregarina gigantea*, by reference to the gigantic size taken by its trophozoite stages, up to 16,000 µm i.e. visible by the naked eye[14]. He reported that "cyst" forms of this parasite accumulated within the chitineous folds of the host's rectum, the "rectal ampulla". Schneider further showed that these cysts enclosed thousands of "gymnospores" or "heliospores", corresponding to spherical groups of very tiny zoites radiating from a central, optically void mass, and renamed the species *Porospora gigantea* (van Beneden, 1869) Schneider, 1875[15]. As biological material from non-cultivable microorganisms is particularly difficult to gather, we took advantage of the existence of these structures later confirmed by others[16–19]. Indeed, they provide a remarkable natural source of genomic DNA, each cyst containing several thousands of "gymnospores", themselves composed of hundreds "zoites". *Porospora gigantea* trophozoites, also known to glide at rates of up to 60µm/s[20], appears as prime candidates to study the mechanism of gliding, a specific form of motility which is a characteristic of Apicomplexa[21–25]. Currently about 40 proteins have been identified mainly in *T. gondii* and *P. falciparum,* and assembled in a commonly accepted structural model, the glideosome (see Frénal et al, 2017[26] for review).

In this study, we report the first draft genome of *P. gigantea.* Remarkably, not one but two related genomes have been assembled. We present their main characteristics and their associated proteomes in the context of the other available apicomplexan genomes. We also determined their position within the phylogeny of marine gregarines. One of the main objectives of this study was to provide an overview of the conservation of proteins involved in the gliding, at the apicomplexan level including gregarines.

RESULTS

Phenotypic characterization

Several specimens of *Homarus gammarus,* the *Porospora gigantea* type host species, were collected from two different sources (Figure 1, Table S2). All 35 lobsters (26 from Roscoff lobster tanks, 9 from Roscoff bay; France) were dissected (Figure 1, Figure S1). Globally, the infection levels were significantly much higher in the lobsters freshly captured from the Roscoff bay than that in the lobsters maintained in captivity in lobster tanks (Table S2), a result yet reported by Van Beneden (1869)[14]. Morphological measurements were performed on cysts, gymnospores, zoites and trophozoites (Figure 1, Tables S3, S4 and S5). Cysts, mostly spherical but sometimes ovoid, have diameters ranging from ~108µm to ~240µm (mean 151.1±45.3µm) and enclose thousands of gymnospores, that are also mostly spherical, with diameters from less than 5µm to almost 7µm (mean 5.63±0.08µm). These gymnospores are indeed composed of radially arranged zoites forming a monolayer with an optically void center. Observation of broken gymnospores (SEM analyzes) allowed measuring the length of their constitutive zoites (mean 1.04±0.16µm) as well as their apical width (mean 0.630±0.129µm). Trophozoite stages are very thin and long, up to 2585µm in our hands for a mean width of 41.8±10.4µm. As described by several authors, their posterior end is slightly thinner, ~30µm. The whole trophozoite surface is covered by longitudinal epicytic folds (Figure S1.B), that are reported necessary to allow eugregarine gliding movement[27]. The sum of these morphological observations are all in agreement with the species being *P. gigantea*, from the type host *H. gammarus*[6,14,15].
Dynamic recording of gliding, performed by isolated trophozoites, confirmed that they move uni-directionally, protomerite ahead, following straight or curved lines depending on the observed individuals, with the whole body (deutomerite) following the trace initiated by the apical protomerite (Film S1). The speed of trophozoites displacement has been calculated to be ~60µm/sec, as initially observed by King and Sleep (2005)[20] and up to more than 100µm/sec in some recordings (Table S6). These variations probably depend upon trophozoites fitness following their isolation from hosts. Syzygies were not clearly observed contrary to solitary encysting trophozoites, sustaining the observation by Leger and Duboscq (1909)[28] that the encysted gymnospores would correspond to schizogonic rather than gamogonic developmental phase, a still debated hypothesis concerning *Porospora*[6].

Two highly related genomes

Four biological samples were sequenced and analyzed independently, and then secondarily assembled together (Figure 2.A). The raw assembly produced 214,938

contigs (99.6Mb) among which 13,656 contigs had a length greater than 1kb (47.9Mb). The obtained scaffolds were cleaned by removing contaminants such as bacteria, fungi and host sequences (Figure 2.B), which created a raw assembly of 1719 contigs for 18Mb.

The analysis of contigs coverage by each individual library revealed a bimodal distribution suggesting a mixture of genomes with a proportion depending on the biological sample (Figure 2, Figure S3). More precisely, while only one set of scaffolds displayed a significant coverage for the lobster tank parasite sample (JS-470, peak around 250X), the three other parasite samples, from freshly captured hosts (JS-482, JS-488, JS-489) showed two distinct sets of scaffolds with different coverage values. The *in-silico* analysis of these two sets revealed an equivalent genome size of ~9Mb. This coverage difference was used to split the contigs into two sets that were given the names A (for the set present in all four samples) and B (for the set present only in three lobsters freshly captured in the wild) (Figure 2.C). The proportion of genomes A and B in each biological DNA sample has been estimated (Figure S3) as 100%-0% for JS-470, 63.2%-36.8% for JS-482, 70.5%-29.5% for JS-488 and 62.4%-37.6% for JS-489, based on medium coverage levels. The genome A contains 786 scaffolds for a total of 8.8Mb whereas the genome B contains 933 scaffolds for a total of 9.0Mb. The contigs from both genomes can be aligned over 7.7Mb, with a percentage of divergence around 10.8% at nucleotide level.

These two genomes have a similar size (~9Mb), are highly syntenic with approximatively 10% of divergence. These highly related genomes have been named A and B and are associated to the species name *P. cf. gigantea* (Figure 2).

Genome features

Two genomes with similar coding capacities. A total of 10,631 putative genes were predicted on the raw assembly, which could be splitted into two sets of similar size: 5270 genes in genome A vs. 5361 genes in genome B (Table 1, Figure 2). The completeness of both A and B genomes was addressed by using the BUSCO software[29] and the Apicomplexa geneset (n=446). Genomes A and B respectively showed a completeness score of 70% (n=312) and 67.7% (n=302) using this Apicomplexa geneset (all BUSCO's scores are shown on Figure S4).

Orthologues were searched between A and B. The proteins of *P. cf. gigantea* A and B were splitted into 5656 orthogroups including 4443 groups (88%) with at least one orthologous gene for both A and B. This percentage of common orthogroups between genomes A and B is higher than the percentage of common orthogroups observed between *P. falciparum* and *P. berghei* (70%), documented to have diverged around 33 Mya ago (TimeTree[30]) but similar to that observed between *P. falciparum* and *P. reichenowi* (86%, 3.3 – 7.7 Mya, TimeTree).

The percentages of shared orthogroups between *P. cf. gigantea* genomes and each of the reference apicomplexan species are similar despite the highly variable divergence (*C. parvum*, 18%; *G. niphandrodes*, 17%; *P. falciparum*, 14%; *T. gondii*, 14%) but it is higher than the percentages observed with chromerid species (*C. velia*,

8%; *V. brassicaformis*, 10%). We can underline from this result that the *P. cf. gigantea* genomes don't share significantly more orthogroups with *G. niphandrodes*, the only other available gregarine genome (Figure 3).

Two gene dense genomes with small introns. The proportion of coding sequences (84%) in A and B genomes is particularly high compared to other reference species (from 25% to 76%; Table 1). The genomic compaction of non-coding DNA in genomes A and B can be explained by the small size of most introns (Figure S5). We observed a specific class of introns with a length around 25-30bp (mode at 28bp) representing 71-72% of the introns. The donor and acceptor sites of these small introns display specific consensus patterns (Figure S5) which are different from the other *Porospora* introns. Especially, these introns exhibit a strongly conserved adenine located 6 bp upstream of the 3' of the acceptor site which could represents the intron branch point as observed for the small introns in *B. microti* (introns of 20bp)[31].

Evolutive history of both *P. cf. gigantea*

Genomes A and B diverged several million years ago. We estimated the putative divergence time of A and B genomes by using the divergence between *P. falciparum* and *P. reichenowi* as a calibration point. The synonymous divergence (dS) was calculated for 1003 quartets of orthologous genes. The mean dS value observed between *P. falciparum* and *P. reichenowi* orthologs was 0.0959, similar to that calculated by Neafsey et al[32] (0.068 substitutions per site) or Reid et al[33] (0.086-0.11 per site). We assumed that these *Plasmodium* species diverged between 3.3 – 7.7 Mya (TimeTree). The mean dS value observed between the same orthologs in both *P. cf. gigantea* genomes was about 0.4295 substitutions per site. Using the hypothesis of similar substitution rates in gregarines and *Plasmodium* species, we dated the split for genomes A and B between 15.5 Mya and 37.7 Mya. This order of magnitude is similar to the basal split estimation for the mammal *Plasmodium*[34] (12.8 Mya) or all *Plasmodium*[35] (21.0–29.3 Mya) but this dating remains significantly posterior to the emergence of Nephropidae, whose lobster is part of[36,37] (~180 Mya).
The 18S SSU rDNA, for which the largest taxonomic sampling for gregarines is available in databases, was used to investigate *P. cf. gigantea* position within apicomplexan and gregarines, especially crustacean gregarines. Using a combination of specific primers amplifications, initially based on Simdyanov et al (2015)[38] and Schrével et al (2016)[39] then in part redesigned (Figure S2, Table S7), and *in silico* clusterings, we were able to fully reconstruct complete ribosomal loci covering: 18S-ITS1-5.8S-ITS2-28S (5977bp), for both A and B genomes. Thirty polymorphic positions were found between A and B that were unevenly distributed i.e. only one position within the 18S sequence, and 29 within the 28S (Figure S2).

Expanded superfamily of crustacean gregarines. Two phylogenetic studies were performed, one excluding environmental sequences (Figure 4; detailed phylogeny

in Figure S6), the other one including them (Figure S7). The vast majority of environmental sequences are derived from marine sediments, from a wide range of habitats with only two sequences from the North Atlantic, i.e. the geographical area of European and American lobsters.
Both phylogenies assigned *P. cf. gigantea* A and B together to one clade, placed as a sister group to all other crustacean gregarines (*Cephaloidophora*, *Heliospora*, *Thiriotia*, and *Ganymedes* species), as established in Rueckert et al (2011)[40], with nevertheless shorter branch lengths, thus looking less derived than the other crustacean gregarine sequences. Five main clades constituting the superfamily Cephaloidophoroidea were retrieved, distributed as follows: the 4 clades previously outlined[40] (redenominated as Ganymedidae, Cephalodophoridae, Thiriotiidae - following the proposal in Desportes and Schrével (2013)[6], and Uradiophoridae), and, at their base, the clade Porosporidae, constituted of the two sequences of *P. cf. gigantea.* We noted the presence of a new putative clade formed by the 5 environmental sequences from a Slovenian karst spring published by Mulec and Summers Engel (2019)[41] (Figure S7). This very well supported clade is placed as a sister group to 4 of the crustacean gregarines families, while the family Porosporidae retains its position as a sister group to all these other clades.

A partially conserved glideosome machinery

We have conducted an inventory of the presence/absence for proteins involved in the glideosome machinery, grouped according to their function as established by Frénal et al (2017)[26] (Figure 5.A, all orthologs for *P. cf. gigantea* are detailed in Table S8). These *T. gondii* and *P. falciparum* references proteins have been searched for in both *Porospora cf. gigantea* genomes and a selection of representative species.

Actin and associated factors. Actin in apicomplexan is characterized by a globular monomeric form (G-actin) which polymerizes as needed into short unstable filaments (F-actin)[42] using various regulators such as profilin[43–45], ADF cofilin[46], formin[47–49] and Cpβ[50]. The inactivation of actin or its associated regulators compromises motility, invasion and egress, although motility may persist in an altered form for a few days, raising the question of alternative mechanisms sustaining parasite motility[27,51–53]. With the exception of profilin in *G. niphandrodes* and Cpβ in piroplasma, all the described proteins were found in all species.

Glideosome apicomplexan-specific proteins. The glideosome machinery mainly comprises specialized apicomplexan-specific proteins. The single-headed short heavy chain myosin class XIV, named myosin A (MyoA), acts as a motor generating the rearward traction required for motility, invasion and egress, as evidenced by various conditional depletion protocols[54–56]. The glideosome itself takes place between the plasma membrane and the apicomplexan-specific inner membrane complex (IMC), in which MyoA is associated with a light chain (myosin light chain 1 – MLC1 in *T. gondii*[57] and MyoA tail domain-interacting protein - MTIP in *P.*

falciparum[58]) as well as glideosome associated proteins (GAP): GAP40, GAP45, GAP50[59–61], GAP70 and GAP80 being only described in *T. gondii*[55]. GAP45 binds the glideosome to the plasma membrane by recruiting MyoA that acts as a bridge[61]; GAP40 and GAP50 are thought to help anchoring MyoA to the parasite cytoskeleton[62], whereas another set of glideosome-associated proteins with multiple-membrane spans (GAPM) are believed to interact with the alveolin and subpellicular microtubules network, suggesting an indirect interaction with the IMC[26,63]. Finally, the conoid-associated myosin H is necessary for initiating gliding motility in *T. gondii*[64].
Myosins ABCDE and its associated light chain were found in all species. Myosin H is also widely distributed, although it is missing in *P. falciparum* and *V. brassicaformis*. The situation is more complex for glideosome associated proteins. Only GAP40 was found in all species, including probable homologues in chromerids. Surprisingly, given the central role attributed to GAP45 in the glideosome model, no orthologue was found in gregarines, *Cryptosporidium* or chromerids. GAP50 was found in all species except the two *P. cf. gigantea* genomes. As expected, GAP70 and GAP80, only identified in *T. gondii*, were not found in other species, except an orthologue for GAP80 in the coccidia *H. hammondi*. Concerning GAPMs, we found orthologues of at least one of its variations (GAPM 1, 2 or 3) for all species. Finally, GAC was found in all species except in chromerids, confirming its occurrence in apicomplexan only.

Adhesins and TRAP-like candidates. The glideosome machinery, anchored in the parasite cytoskeleton, needs to interact with extracellular receptors of the host cell, in order to propel the parasite forward on its surface; this is made possible by the presence of extracellular adhesins secreted by the micronemes[65,66] and connected to the glideosome through the glideosome associated connector (GAC) protein[67]. One adhesin described in *Plasmodium* in particular is required for gliding: TRAP (Thrombospondin Adhesive Protein[68]) whose homologue in *T. gondii* is MIC2[69]. At the end of the gliding process, rhomboid protease 4 (ROM4) attaches to the adhesins, disengaging them from receptors and, for intracellular parasites, allowing them to enter the host cell[70–72]. TRAP-like proteins, while highly divergent from a species to another, constitute a family of functionally homologous proteins involved in parasite gliding motility and cell penetration[73–75]. TRAP-like or TRAP-related proteins have been described in various stages of *Plasmodium* (CTRP[76], MTRAP[77], TLP[78]) and have also been described *in silico* in *Cryptosporidium* (TRAPCs, CpTSPs[75,79,80]) as well as in several *Babesia* and *Theileria* species[81–84], in *Neospora caninum*[85] and in *Eimeria*[86,87].
We first looked for the TRAP proteins whose implication in gliding have been described by experimental studies (MIC2, TRAP, TPL, CTRP, MTRAP), as well as the ROM4 protein involved in adhesin cleavage. Unsurprisingly, the currently described TRAP proteins seem to be genus- or even species-specific. On the other hand, we found orthologues for ROM4 in all species, except for chromerids.
The TRAP proteins described to date have the following characteristics: an extracellular region containing one or more TSP1 domains and/or one or more vWA

domains[73–75]. They are also characterized by the presence of a single transmembrane domain, a signal peptide, as well as, in some cases, a juxtaposed rhomboid protease cleavage site, and a short and charged C-terminal cytoplasmic domain, together with aromatic residues. The presence of a YXXΦ tyrosine sorting signature has also been described[74] (X meaning any amino acid, Φ a hydrophobic amino acid - isoleucine, leucine, methionine, phenylalanine, or valine).
In order to evaluate the presence of TRAP-like proteins in *P. cf. gigantea* genomes, we inventoried all proteins containing at least one TSP1 domain (Table S8), and identified potential candidates displaying several structural characteristics of TRAP-like family (Figure 6). We identified a CpTSP2 orthologue within both *P. cf. gigantea* genomes, PgTSP2. Like CpTSP2, it is a large protein (~2800aa) composed of Notch, TSP1, and Sushi domains. It has an addressing signal, a transmembrane domain and a short and charged basic cytoplasmic tail. This protein also has orthologues in *G. niphandrodes,* in chromerids and coccidia.
We also demonstrated the presence of four other protein pairs present in both A and B genomes, most of which appear to be specific to *P. cf. gigantea.* PgTSP-1 has a TSP1 domain, a signal peptide, a transmembrane domain and a short acidic and charged cytoplasmic tail. PgTSP-2, very similar in structure to PgTSP-1 also has a TSP1 domain, a signal peptide, a transmembrane domain and a short, charged but basic cytoplasmic tail. PgTSP_EGF-1 has two TSP1 domains, a signal peptide, a transmembrane domain and a short acidic charged cytoplasmic tail, and several EGF or EGF-like domains in their extracellular portion, as also described in *C. parvum* (CpTSP7, CpTSP8 and CpTSP9[79]). We also identified another protein very similar in structure, PgTSP_EGF-2.

Moving junction associated proteins. In intracellular apicomplexan such as *T. gondii*, invasion occurs as the tachyzoite initiates a pivotal movement known as reorientation, while the mobile junction settles into the host cell membrane, allowing the parasite entry; gliding forces are also involved in this process[88], to which host cell also contributes[89]. A micronemal protein, AMA1, combines with rhoptries neck proteins (RON2, RON4, RON5 and RON8) to firmly maintain the parasite attached to the host cell. In *P. falciparum,* another AMA-like protein, merozoite apical erythrocyte-binding ligand (MAEBL) has an important role in invasion alongside with AMA1[90].
Unsurprisingly, we have not identified any orthologue to the moving junction proteins nor in gregarines neither in *Cryptosporodium*; indeed, these groups are known to remain mainly extra-cellular (gregarines) or epi-cellular (*Cryptosporium*). We also searched for proteins described in *Cryptosporidium* as implicated in adherence and invasion, such as GP15/40, GP900 and mucins, but found no equivalent in gregarines[91,92].

Regulatory factors and signaling pathways. The increase of parasite intracellular calcium, by activating calcium-dependent protein kinases (CDPK), is involved in the regulation of motility, microneme secretion, invasion and egress[93,94]. Other proteins

known in such signaling pathways include phosphatidylinositol phospholipase C (PI-PLC), diacylglycerol kinase 1 (DGK1), acylated pleckstrin homology domain-containing protein (APH) which are involved in micronemes secretion regulation [95,96], the C2 domains-containing protein DOC2.1 which mediates apical microneme exocytosis[97]; finally, the apical lysine methyltransferase (AKMT), which is involved in gliding motility, invasion and egress in *T. gondii*[98]. With the exception of the APH that we were unable to identify in *Porospora cf. gigantea* or chromerids, all the regulatory factors appeared to be largely conserved.

DISCUSSION

Molecular data support the presence of two species

We report here clear lobster coinfection by two distinct gregarines that we named *P. cf. gigantea A* and *P. cf. gigantea B*. At molecular level, these two organisms display highly similar genomes in terms of size, protein coding capacity, GC content and overall organization (86% synteny conservation). The delineation of species now requires integrative morpho-molecular approaches, especially in protistology. Currently, the only molecular tool available for species discrimination in gregarines is the nucleotide sequence of the 18S SSU rDNA. At this molecular marker level, *P. cf. gigantea A* and *P. cf. gigantea B* differ by only a single nucleotide, a divergence level classically considered as indicative of organisms belonging to the same species.

However, at the genomic level, both genomes show a nucleotide divergence of more than 10% which is incompatible with subspecies or strain definitions. By comparison, the same protocol applied at *P. falciparum* and *P. reichenowi* concluded at a divergence of only 3.2%. Moreover, a divergence of 3-5% has also been reported between the genomes of *Cryptosporidium parvum* and *C. hominis*[99]. This large overall divergence at genomic level indicates that *P. cf. gigantea A* and *P. cf. gigantea B* are probably not inter-fertile, and thus should be considered as different species.

Our morphological observations remain however insufficient to point to distinctive morphological features between *P. cf. gigantea* A and *P. cf. gigantea* B as they both corresponded to the published features of *P. gigantea* (type host infected, host-compartments infected, morphology of developmental stages, including speed of gliding). We believe that a specific study should be designed and performed, combining extensive imaging (SEM, TEM) and single cell –omics, using preferably lobsters directly caught from the wild to further explore their morphological specificities.

We thus maintain the proposal of *P. cf. gigantea A* and *P. cf. gigantea B* to name the two organisms we found in *H. gammarus*, pending a more integrated morpho-molecular definition of their taxonomy, as well as a better documentation of Cephaloidophoroidea species.

Two species with particularly compact genomes displaying highly specific common gene set

These two genomes, the first published genomes for marine gregarines, highlight several important findings. Highly similar in terms of size and gene content, both A and B genomes are also highly reduced, with a small genome size compared to other apicomplexan, and an especially high gene density (e.g., at similar genome size, *Cryptosporidium* spp. display a number of protein-coding genes of about 3900 only). This result could be partially explained by the absence of certain non-coding sequences into the assemblies such as centromeres, telomeres and repeated sequences which are particularly difficult to sequence and assemble, notably in de novo assembled genomes. This compaction is partially due to the presence of short introns. Small introns with similar consensus sequences have been described in *Babesia microti*[31]. We have not identified any organellar genome (mitochondrion and apicoplast), an absence that needs to be further investigated especially concerning the mitochondrial genome, that may be underrepresented in cysts stages.

BUSCO genome completeness scores of ~70% were found for the two *P. cf. gigantea* genomes, a value not unusual for non-model species[29], but that is lower than that found for the *G. niphandrodes* genome (83%) and the 24 other representative species we evaluated (from 76.9% for *C. suis* to 100% for *P. falciparum* (Figure S4)). This result also illustrates that the definition of "Apicomplexa core genome" is certainly currently highly biased towards notably *Plasmodium*. Thus, gregarines should be taken into consideration, as their divergence compared to other apicomplexan models is confirmed by the orthogroup analysis indicating a low percentage of genes conserved between A or B and other studied apicomplexa (<18%).

Even among gregarines a wide diversity may be pointed out as the vast majority of proteins shared by A and B are absent from the *G. niphandrodes* genome. Therefore, studying gregarines will allow a better understanding of the evolutionary history of apicomplexan species by highlighting an astonishing protein diversity and a complex differential inheritance from the common ancestor. Through comparative analyses, we will be able to understand how this inheritance has allowed such a wide range of documented adaptations to parasitism in apicomplexans, which have been able to establish themselves in most Metazoan lineages, vertebrate or not, marine or terrestrial, in one or more hosts, intra- or extracellularly.

The gregarine glideosome(s)

An incomplete but operational machinery. Our molecular analysis of the glideosome components shows that the currently known mechanistic model described from *T. gondii* and *P. falciparum* cannot fully account for gliding in all apicomplexans, as previously anticipated[26,62,66]. Some key molecular components such as canonical adhesins or GAP45 appear to be missing, implying an only

partially retrieved glideosome in gregarines as well as in *Cryptosporidium* species (Figure 5.B). Since we have observed the gliding movement of *P. cf. gigantea* trophozoites at an impressive rate, this raised the question of how they manage to perform this rapid movement in the absence of a complete dedicated machinery.

The classical machinery may be partially compensated by alternative proteins. The TRAP adhesin in *T. gondii,* named TgMIC2, has been demonstrated to be an important but non-essential protein to motility[100]. This suggests that TRAP proteins may not be the only proteins involved in host surface adhesion. As we have seen, in the genomes of *P. cf. gigantea*, as well as in other apicomplexans, there are proteins with a structure close to TRAPs, called TRAP-like, that could replace the canonical TRAP proteins. This is why understanding the evolution of TRAPs proteins involves experimental validation of predicted adhesion proteins in gregarines and *Cryptosporidium* - especially since the presence of these domains in Alveolata does not always correlate with gliding motility[75]. Similarly, the vWA domains, which are found in the canonical TRAPs, appear to be absent from the *Cryptosporidum* genomes; however, since gliding is observed in these species, it can be assumed that, if the TRAP-like proteins described in *Cryptosporidium* are indeed involved in gliding, then the vWA domains are not essential for this process; it is also possible that the TSP1 domain genes represent only one adhesion pathway among others, and that other adhesion domains could perform functions similar to TRAPs, such as the Apple and EGF-like domains in *Cryptosporidium* (Morahan et al, 2009; Deng et al, 2002). As to GAP45, it is thought to maintain the interaction between the IMC and the plasma membrane, and acts as an essential bridge between the two structures[101]. Likewise, the absence of GAP45 in gregarines and *Cryptosporidium* maybe compensated by other GAP-like proteins or even not be a problem at all; indeed, it has been proposed that a motor architecture could be organized in a much looser manner, in which actin-myosin motors push in a general backward direction, but without necessarily being guided by GAP proteins (Tardieux and Baum, 2016). Furthermore, while TgMLC1 binding to TgGAP45 is considered a key component of the parasite's force transduction mechanism, it has recently been shown that loss of TgMLC1 binding to TgGAP45 has little effect on their ability to initiate or maintain movement[102], questioning again the real role of GAP45 and suggesting our comprehension of the glideosome's proteins' intrication is still incomplete.

A completely different structure taking advantage of the other forms of motility known in gregarines? Gregarines have other means of motility, presumably governed by other molecular mechanisms. Yet questions have been raised about the relevance of the glideosome concept as applied to gregarines[27,103]. In particular, it is known that archigregarines use several modes of movement such as rolling and bending but not gliding[6,19].

For their part, coelomic and intestinal eugregarines like crustacean gregarines have longitudinal, drapery-like surface structures called epicytic folds that represent the most noticeable feature that differentiates eugregarine trophozoites and gamonts from other apicomplexans, and are considered to be involved in eugregarines'

gliding, by increasing the surface area and facilitating actomyosin-based gliding motility - reviewed in Valigurová et al (2013)[27]. Indeed, actin and myosins A, B and F have been localized in epicytic folds in *G. polymorpha*[104,105]. Epicytic folds, together with the mucus, which refers to the material often observed in the trace left by gliding eugregarines[6,27], are definitely key structures to investigate in order to understand their exact composition and thus be able to propose an alternative model to the glideosome one, suited to the motility of eugregarines. A particularly interesting study concerning the crustacean gregarine *Cephaloidophora cf. communis* reports on the specific structures of its attachment apparatus[106]. While actin in its polymerized form (F-actin) is observed all along the gregarine, myosin is confined to the cortical region of the cell, in connection with the longitudinal epicytic folds as described in Valigurová et al (2013)[27]. It has also a septum, a structure that separates the epimerite from the protomerite at the cell apex, consisting of tubulin-rich filamentous structures. Together with microneme-like structures, these features suggest a production of adhesion proteins which would be sent through the membrane by the numerous pores visible on the epimerite[106]. We were unable to identify alternative movements to gliding motility in *P. cf. gigantea* (like peristaltic movement described in other coelomic eugregarines[6,107]), and we believe that additional observations are needed to fully document the range of potential motilities in this species, especially since *C. cf. communis* is capable of jumping or jerking movements during discontinuous gliding[106]. The different structures described, or their absence must be evidenced as well; indeed, in eugregarines, subpellicular microtubules have never been observed, whereas they are supposed to be involved in gliding motility in other apicomplexan[27,106].
In light of these hypotheses, involving alternative proteins compensating for canonical glideosome machinery or suggesting the implication of other motility mechanisms altogether, it is likely that the molecular mechanisms leading to gliding motility in *P. cf. gigantea* reveal a unique molecular structure, consecutive to the specific evolutionary path of gregarines, and which differs from what is currently documented in other apicomplexan lineages.

ACKNOWLEDGEMENTS

Roscoff Marine Station and service mer for the hosts; Pauline Konga for the ribosomal amplifications; Geraldine Toutirais, Cyril Willig, Marc Gèze and the MNHN Platform (Plateau technique de Microscopie Électronique, Muséum National d'Histoire Naturelle MNHN, Paris, France, http://ptme.mnhn.fr/) for the SEM imagin ; the computational cluster of the MNHN for assemblies/phylogenies.

This work was supported by a grant from the French *Agence Nationale de la Recherche* [LabEx ANR-10-LABX-0003-BCDiv], in the program "*Investissements d'avenir*" [ANR-11-IDEX-0004-02], by several interdisciplinary Programs of the

MNHN (*ATM-Microorganismes*, *ATM-Génomique et Collections*, *ATM-Emergence*, AVIV department), the CNRS (Julie Boisard's PhD fellowship, 2018–2021).

This work also benefited from access to the Station Biologique de Roscoff, an EMBRC-France and EMBRC-ERIC Site. The present work has been funded in parts by the call EMBRC-France 2016 (Investments of the Future program, reference ANR-10-INSB-02, Agence Nationale de la Recherche).

AUTHOR'S CONTRIBUTION

JB, EDB, LP and IF designed the study. IF, JS and LG performed the biological sampling and IF extracted the nucleic acids. IF, JS and LG performed the photonic microscopy analyses, IF performed the SEM while SLP and GP performed the TEM. JB, EDB and LP realized the bioinformatics analyses. AL and LD sequenced and assembled the complete ribosomal loci. JB performed the phylogenetic analyses and the glideosome expert annotation.
JB, EDB, LP and IF wrote the manuscript with contributions from all authors. All authors have read and approved the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

MAIN-TEXT FIGURE/TABLE LEGENDS

Figure 1. **Morphological characterization of *Porospora cf gigantea*. A.** Trophozoite stage (Tropho#8, Lobster#12) (scale=100µm). **B.** Zoom on A, showing trophozoite epimerite (scale=10µm). **C**. Rectal ampulla showing cysts in folds (Lobster#4) (scale=1mm). **D.** Isolated cyst (Cyst#4, Lobster#12) (scale=50µm). **E.** Broken cyst allowing to visualize enclosed, packed gymnospores (Lobster#4) (scale=10µm). **F**. Section across a cyst illustrating radial arrangement of zoites in gymnospores (JS449=Lobster#35) (scale=2µm). **G., H.** Zoom on intact and broken gymnospores allowing visualizing zoites (Lobster#4) (scale=1µm). Scanning (A, B, C, D, E, G, H) and transmission (F) electronic microscopy. See also Figure S1, Tables S2, S3, S4, S5 and S6.

Figure 2. Assembly protocol of the two genomes. A. Overview of the full protocol. **B.** Identification of apicomplexan vs. contaminant contigs based on k-mer composition. **C.** Identification of contigs from genomes A and B based on coverage data for each individual library. See also Figures S2, S3, S4 and S5.

Figure 3. Shared apicomplexan proteins. Distribution of the orthogroups among *P. cf. gigantea* A and B and 4 species of apicomplexans: the gregarine *G. niphandrodes*, the cryptosporidian *C. parvum*, the coccidian *T. gondii* and the hematozoan *P. falciparum*. Only bars with more than 20 orthogroups are shown. See also Table S1.

Figure 4. Gregarines/apicomplexan phylogeny. Phylogenetic tree built using 100 18S SSU rDNA sequences 1614 sites in order to situate the two *P. cf. gigantea* sequences among other known gregarines and apicomplexan clades. Chromerid sequences were used as outgroup, as they are considered the sister group of all other apicomplexans[108]. Evolutionary history was inferred by maximum likelihood and bayesian inference using a GTR+G+I model. Topologies were identical according to both methods. Black spots indicate 100/1 supports. Supports <70/0.7 are not shown. Families and associated literature are indicated. Sub-trees have been collapsed at the family level in order to make the phylogeny more legible. See also Figure S6 (extended phylogeny) and Figure S7 (including environmental sequences).

Figure 5. Comparative analysis of glideosome components. A. Table of presence/absence of glideosome proteins, distributed into functional groups. Glideosome components have been described mainly in *T. gondii* and *P. falciparum*. Proteins were searched for in both *Porospora* genomes as well as in a selection of representative species. Green indicates the presence, while white indicates the absence of a protein. Light green refers to the cases where one-to-one orthologous relationships have not been conclusively identified in *C. velia* and *V. brassicaformis*, but related protein expansions have been observed[108]. All *P. cf. gigantea* orthologous proteins are detailed in Table S8. **B.** Schematic comparison of the canonical model of the glideosome and the elements found in *P. cf. gigantea* A and B. Missing proteins are shown in dotted line.

Figure 6. Structures and molecular domains of candidate TRAP-like proteins in *P. cf. gigantea* A and B. See also Table S8.

Table 1. Metrics of the genomes of *P. cf. gigantea* and a selection of 6 reference species. See also Figure S1 and S5.

MATERIAL & METHODS

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Isabelle Florent (isabelle.florent@mnhn.fr).

Materials availability

This study did not generate new unique reagents.

Data and code availability

DNA and RNA reads are available into the NCBI database (Bioproject PRJNA734792).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Several specimens (n=35) of the lobster species *Homarus gammarus* were collected in the British Channel at Roscoff (Britain, France) between July 2015 and October 2017 (Table S2), either directly from the field (Roscoff bay) or through lobster tanks facilities, in which crustaceans are maintained in captivity several weeks to months before their commercialization. The intestinal tract was carefully dissected from each freshly killed host specimen, and was transferred to large Petri dishes filled with 0.22µm filtered and autoclaved sea water, supplemented with the antibiotics penicillin (100 U/mL), streptomycin (100 µg/mL) (Gibco, Life Technologies, USA) and gentamycin (50 µg/mL) (Interchim, Montluçon, France). Trophozoite stages, freely moving in the upper intestine lumen and cyst stages, loosely attached within the chitineous folds of the hosts' rectal ampulla (Figure S1), were individually collected using elongated Pasteur pipettes, under a classic binocular microscope. For the recording of gliding movement, trophozoites were kept in non-treated sea water. For all other applications, trophozoites, cysts and host tissues were carefully washed several times in 0.22µm filtered and autoclaved sea water, supplemented with the antibiotics indicated above. Trophozoites and cysts were collected for photonic live imaging, scanning electronic microscopy and transmission electronic microscopy, as well as for subsequent -omic studies (i.e. DNA and RNA sequencing). According to UICN Red list, *Homarus gammarus* is not an endangered species[109].

METHOD DETAILS

Electronic microscopy

For the Scanning Electron Microscopy (SEM) studies, isolated trophozoites and cysts, hosts intestines and rectal ampullas opened along their longitudinal axis, were washed as indicated above before fixation in 2.5% (v/v) glutaraldehyde in 0.1M sodium cacodylate (pH 7.2), at 4◦C, for 6 to 12 hours. After two washing steps in 0.1M sodium cacodylate (pH 7.2), biological specimens were transferred to Microporous Specimen Capsules (30µm porosity, 12mm diameter, 11mm high, ref

#70187-20, Electron Microscopy Science) and dehydrated in a graded series of ethanol in ddH_2O (50, 70, 90, and 100%). Biological specimens were then critical point-dried in liquid CO_2 (Emitech K850, Quorum Technologies) in the Microporous Specimen Capsules, then transferred to adhesive carbon coated holders, and coated with 20nm of gold (JEOL Fine Coater JFC-1200). The biological specimens were then examined with a Hitachi Scanning Electron Microscope SU3500 Premium.
For the Transmission Electron Microscopy (TEM) studies, the samples were fixed for 2h in 0.2M sodium cacocylate buffer with 4% glutaraldehyde, 0.25M sucrose in 0.2M sodium cacodylate buffer pH 7.4. Cells were then washed three times in sodium cacodylate buffer containing decreasing concentrations of sucrose (0.25M, 0.12M, 0M) for 15 min each time, followed by postfixation for 1h at 4 °C in 2% osmium tetroxide in 0.1M sodium cacodylate buffer. After three rinses in 0.2M sodium cacodylate buffer, samples were dehydrated by successive transfers through a graded ethanol series (25%, 50%, 70%, 90%, 3 × 100%), then embedded in Spurr's resin. Sections were cut using a diamond knife on a Leica Ultracut UCT ultramicrotome (Leica, Wetzlar, Germany) and after staining with saturated uranyl acetate for 15 min and Reynolds' lead citrate for 3min, grids were examined with a Jeol 1400 transmission electron microscope (Jeol, Tokyo, Japan).

DNA/RNA isolations

Genomic DNA (gDNA) was isolated from 4 biological samples, all composed of cysts' pools taken from 3 specimens of the host *Homarus gammarus*: sample JS-470 from Lobster #7 (~70 cysts), sample JS-482 from Lobster #11 (~50 cysts), samples JS-488 and 489 from Lobster #12 (~100 cysts each). Lobster #7 was provided by Roscoff lobster tank facility while Lobster#11 and Lobster #12 were caught from the field in Roscoff bay. DNA was extracted from the cyst pools using Macherey Nagel Tissue and Cells isolation kit (ref 740952.50) with a yield of respectively: 4.1µg (JS-470), 2µg (JS-482), 4.5µg (JS-488) and 6.7µg (JS-489) of total DNA per sample, as measured by Nanodrop quantification. The protocol was used as recommended by Macherey Nagel, except that the initial lysing step (at 56°C), was extended beyond the recommended to 1-3 hours with frequent microscopic (binocular) inspection to follow complete cysts digestion.
RNA was also isolated from 2 additional biological samples, both composed of cysts' pools taken from the rectal ampulla of their respective host specimens: JS-555 (~35 cysts, Lobster#26, Roscoff bay) and JS-575c (~40 cysts Lobster#34, Roscoff Lobster tank facility). Two distinct protocols were used to isolate total RNA from these two biological samples. For sample JS-555, we used Macherey Nagel basic RNA Isolation kit (ref 740955.10) which yielded ~155ng of total RNA in 55µl as assessed by Qbit quantification. For sample JS-575c, we used Macherey Nagel Nucleozol-based RNA Isolation kit (refs 74040.200 and 740406.10) which yielded ~50ng of total RNA in 55µl as assessed by Qbit quantification.

DNA/RNA sequencing and assembly

The gDNA extracted for the 4 biological samples (JS-470, JS-482, JS-488 and JS-489) were sequenced individually by using Illumina NextSeq technology (2*151bp; NextSeq 500 Mid Output Kit v2; Institut du Cerveau et de la Moelle Epinière - CHU Pitié-Salpêtrière - Paris). We obtained 2*50 M to 2*70 M reads which were checked using FastQC[110] (version 0.11.5). Reads were cleaned by using Trim Galore[111] (version 0.4.4) removing remnant Nextera adaptors, clipping 15 bp in 5'-end and 1 bp in 3'-end and trimming low-quality ends (phred score < 30). The assembly was carried out by using SPAdes[112] (version 3.9.1; options: careful mode, automatic k-mers) with the pooled libraries (Figure 2.A).

The RNA was extracted from both samples (JS-555 and JS-575c) and treated with RNAse-free DNase. Library preparations (Institut du Cerveau et de la Moelle – CHU Pitié Salpétrière - Paris) were realized following manufacturer's recommendations (SMART-Seq v4 Ultra Low Input RNA Kit from Takara). Then, final samples were sequenced on NextSeq 500 Illumina device with MidOutPut cartridge to generate a total of 2*87 M reads of 75bp. The read quality was checked by using FastQC[110] and cleaned by using Trim Galore[111] to remove remnant Nextera adaptors, clipping 15 bp in 5'-end and 1 bp in 3'-end and trimming low-quality ends (phred score < 30). The sequence reads of both samples were merged into one library which was assembled using Trinity[113,114].

All genomic contigs longer than 1kb were analyzed by using a principal component analysis (PCA) based on their 5-mer composition, which allowed classifying them into 6 groups by using a hierarchical clustering method (HCA) based on the Ward criterion (Figure 2.B).

For all contigs, the putative protein coding genes were then predicted by using Augustus[115] (version 3.3) and the only gene model implemented for an Apicomplexa: *T. gondii*. All the predicted proteins were thus compared with the NCBI non-redondant protein database by using BLAST[116]. The analysis of the taxonomic groups associated to the corresponding best hits, enabled us to identify five clusters as putative bacterial contaminants whereas the sixth cluster which included 1745 contigs (18.0Mb), was identified as organisms closely related to Apicomplexa, referred as "apicomplexa" cluster later (Figure 2.B).

Identification of genomes A and B

The contigs of the "apicomplexa" cluster were splitted into genomes A and B by using the difference of coverage observed for each of the 4 gDNA libraries (figure 2.C). Each gDNA library (JS-470, JS-482, JS-488 and JS-489) was individually mapped on the contigs by using Bowtie2[117] and the coverages' medians were calculated for each contig and each library by using the Samtools[118] and the Bedtools[119] libraries. This coverage information was processed with a principal component analysis and a k-means algorithm which allowed classifying the contigs into 2 clusters. Then, a linear discriminant model was trained with the coverage information and the result of this first classification before to be applied to all the contigs in order to improve

the classification. The linear discriminant method (training and classification) was iteratively repeated 3 times until convergence. A similar analysis was carried out with 1kb non-overlapping windows (instead of full length contigs) to identify some putative hybrid contigs. Then, contigs classified to different genomes depending on the windows were divided into sub-contigs which were re-assigned to their respective genomes.
The nucleic divergence between genome A and genome B was estimated from the alignment of contigs built with Mummer3.0[120]. All alignments of the syntenic regions were parsed to compute the divergence by using a home-made script. Assembly metrics were assessed by using respectively QUAST[121] (version 5.0).

Prediction and annotation

All *de novo* assembled transcripts were aligned against the "apicomplexa" cluster contigs with GMAP[122] within the PASA program[123]. Then, two *ab initio* gene prediction tools, SNAP[124] (version 2017-11-15) and Augustus[115] were trained using a subset of the PASA transcriptome assemblies. A specific gene model was trained with Augustus, including meta-parameter optimization and construction of hints for introns (allowing small introns length >10bp) using our "apicomplexan cluster" repeat-masked genome assembly as reference (RepeatMasker[125], version 4.0.8). Gene predictions were then performed allowing the prediction of alternative transcripts and noncanonical intron bounds. An alternative model was also trained with SNAP (default protocol) and used for gene predictions. The Augustus and SNAP outputs having sometimes predicted genes slightly differently, the predictions were then parsed by a home-made script in order to keep, for each prediction made, as many alternative genes and transcripts as possible. The completeness of the gene prediction was assessed by using BUSCO[29] (version 4.0.6).
The predicted proteins, have been automatically annotated by using i) the best hit of a BLASTP search against VEupathdb (version 2019-20-01), ii) the results of KoFamScam against the KEGG pathway database[126] (version 2019-05-11) and iii) the signature domains obtained with Interproscan[127] (version 5.39-77.0).
The ortholog groups were identified by using orthoMCL[128] (default parameters, version 2.0.9) applied to the proteome of a selection of representative organisms available on VEuPathDB (Table S1).
The divergence time of genome A and genome B was estimated from the divergence time of *P. falciparum* and *P. reichenowi* as estimated in TimeTree[30]. Then, the coding sequences of the orthologous groups/quartets including one and only one gene for the genome A, the genome B, *P. falciparum* and *P. reichenowi* were aligned by using MacSE[129]. For each alignment, the number of synonymous substitutions per site (dS) between genomes A/B and between *P. falciparum/reichenowi* were computed with the maximum likelihood method of Yang and Nielsen (2000)[130] implemented in PAML4[131].
The Infernal software[132] (version 1.3.3) and the Rfam database[133] (version 14.2) were used together to search for transfer RNAs, spliceosomal RNAs and ribosomal RNAs. The snoReport software[134] (version 2) was used to search C/D and H/ACA small

nucleolar RNAs.

Removal of contaminant sequences

Host. All the "apicomplexa" cluster contigs were screened by using the short reads available for the sequencing project of the genome of the closely related *Homarus americanus* species (PRJNA486050) in order to identify host contaminants. This dataset was supposed to be free of sequences from apicomplexan species, since it has been obtained from DNA extracted from the non-intestinal tissues (the tail, the leg or the pleiopod appendices). The mapping was carried out with Bowtie2[135] and the coverages were calculated by using the Samtools[118] library. The contigs thus identified, that were covered over more than 60% of their length by *Homarus's* short reads, were considered as host contaminants and were removed.

Prokaryote/Fungi. In parallel, predicted genes on the "apicomplexa" cluster contigs were also deeply analyzed for contamination by bacteria and fungi sequences. On scaffolds of this cluster containing at least one predicted protein, a BLASTP against NCBI NR database was launched. For contigs displaying a hit with a e-value lower than 1e-30 and more than 30% of length covered by Prokaryotes/Fungi hits on these contigs, an additional BLASTN against NCBI NR/NT was performed. For the remaining scaffolds without predicted protein, a BLASTN vs nr/nt was directly performed. At the end of this procedure, the contigs with Prokaryotes/Fungi hits covering more of 70% of length were labeled as contaminants and were removed from the genome assembly.

Experimental reconstruction of 18S/28S loci

First, a partial SSU rDNA locus was amplified by using JS-470 gDNA (including only the genome A) as DNA template and WL1 and EukP3 primers (Table S7) in a classical PCR reaction; the amplified bands were systematically cloned and sequenced as previously described[39]. This partial SSU rDNA sequence was further extended experimentally in the 3' direction still using JS-470 gDNA as DNA template and novel primers designed or re-designed based on the molecular data published for *Cephaloidophora cf. communis* and *Heliospora cf. longissima*[38] (Figure S2.A). The resulting sequence (>4 kb) was then used as anchor to reconstruct a complete ribosomal locus with the program iSeGWalker[136]. By clustering reads from JS-470 on this anchor, a 7322-bp theoretical sequence that corresponded to [partial 28S – 18S – ITS1 – 5.8S – ITS2 – partial 28S] including a perfect 1352-bp overlap between the 5' and 3' [partial 28S] segments was obtained. This allowed reconstructing a complete ribosomal locus [18S – ITS1 – 5.8S – ITS2 –28S] of 5977-pb for genome A, that was then experimentally validated by PCR amplification, cloning and sequencing (Figure S2.B). A similar clustering approach, using all reads for JS-482, JS-488 and JS-489, allowed reconstructing *in silico* the complete ribosomal locus for genome B, which was of same length, but with 30 polymorphic positions compared to the complete ribosomal locus for genome A (Figure S2.C). 60% of the complete ribosomal locus for genome B was then confirmed experimentally by PCR amplification, cloning and

sequencing (positions 1187 to 4220). This second clustering also allowed quantifying the respective distributions of genomes A and B, present in these three biological samples, at the full risobomal locus level. The validated sequence of 18S/28S has been manually added to the genome assemblies of genomes A and B, respectively. Schematic representation of rRNA loci was done using BioRender (biorender.com).

Phylogeny

Gregarines phylogeny. The 100 sequences phylogeny was built from the 18S SSU rDNA sequences of the 2 genotypes of *Porospora cf. gigantea*, which were aligned with 84 sequences from a diversity of gregarines, either marine or terrestrial, as well as 12 other apicomplexans sequences. Two chromerids sequences were used as outgroup[108]. A total of 1614 sites were conserved after a selection of conserved blocks as defined by Gblocks[137] (version 0.91b) (Parameters used: Minimum Number Of Sequences For A Conserved Position: 51; Minimum Number Of Sequences For A Flanking Position: 51; Maximum Number Of Contiguous Nonconserved Positions: 8; Minimum Length Of A Block: 3; Allowed Gap Positions: All). A General Time Reversible (GTR) substitution model with gamma-distributed rate variation across sites and a proportion of invariant sites was suggested as the best-fit model according to the Bayesian Information Criterion (BIC) and the Akaike Information Criterion (AIC) calculated by MEGA X[138]. Maximum likelihood analyses were performed using RAxML[139] (version 8.2.12); bootstraps were estimated from 1,000 replicates. Bayesian phylogenetic tree was constructed with MrBayes[140] (version 3.2.3) using the following parameters: lset nst = 6 rates = invgamma; mcmc ngen=10000000 relburnin=yes burninfrac=0.25 samplefreq=1000 printfreq=10000 nchains=4 nruns=2 savebrlens=yes; sump burnin=2500000; sumt burnin=2500000 contype = allcompat. Trees were visualized and edited using FigTree[141] (version 1.4.4) and Inkscape (www.inkscape.org).

Environmental phylogeny focused on crustacean gregarines. The 189 sequences phylogeny was built from the 18S SSU rDNA sequences of the genomes *A* and *B*, which were aligned with 14 sequences from crustaceans' gregarines, as well as 154 environmental sequences from several projects described in Rueckert et al (2011)[40] or gathered from NCBI Genbank. The sequences from Gregarinoidae clade (n=19) were used as outgroup, as this group has been placed as a sister group to the crustacean gregarines clade in recent literature[10–12]. A total of 1135 sites were conserved after a selection of conserved blocks as defined by Gblocks[137] (Parameters used: Minimum Number Of Sequences For A Conserved Position: 95; Minimum Number Of Sequences For A Flanking Position: 95; Maximum Number Of Contiguous Non conserved Positions: 8; Minimum Length Of A Block: 3; Allowed Gap Positions: All). Maximum likelihood and bayesian analyses were performed following the same protocol and parameters as in the previous phylogeny.

Expert annotation for glideosome proteins

A reference apicomplexan glideosome proteins dataset was elaborated based on glideosome protein repertoires described in the literature mainly for *T. gondii* and *P. falciparum*[26,66,67]. This reference dataset was used as a seed for parsing the orthogroups made for 25 reference proteomes (Table S1) and the proteomes of the two *P. cf. gigantea* genomes. The selection of these reference proteomes was made by considering the most recent data and associated publications, in order to have the most complete panorama of apicomplexan proteins and key functions/structures currently documented.

For each orthogroup containing at least one of the reference proteins, the list of proteins was extracted and the protein sequences were recovered, as well as their respective coding nucleic sequences for both *P. cf. gigantea* genomes. A BLASTP was performed for extracted proteins against the proteomes of *P. cf. gigantea*, as well as a BLASTP of the candidate proteins for each *P. cf. gigantea* genome against the 25 species reference proteomes. A BLASTN was performed against NCBI NR for the coding sequences of the candidate proteins of both *P. cf. gigantea* genomes. The sequences thus collected for each described protein were aligned with mafft[142]. Maximum likelihood molecular phylogeny was performed on each alignment using RAxML software[139]. Analyses were performed using the LG model; bootstraps were estimated from 1,000 replicates. Annotations of the conserved molecular domains were searched for in the automatic annotation and structure analyzed with SMART[143]. For each protein, all the performed analyses were then manually examined to validate the candidate proteins within the proteomes of the two *P. cf. gigantea* genomes. Presence/absence table of glideosome proteins was visualized using R using the tidyverse package[144]. The identification of TRAP-like proteins was done by searching for the TSP1 molecular domain (IPR000884) within the two *P. cf. gigantea* genomes. The structure of each candidate protein was then carefully studied. If necessary, partially predicted proteins were re-edited with Genewise[145]. Schematic representation of TRAP-like proteins was done using BioRender (biorender.com).

LEGEND FOR SUPPLEMENTAL VIDEO

Video. Photonic recording of trophozoites gliding (Ref 531004, .avi format, 31/05/2016).

REFERENCES

1. Swapna, L.S., and Parkinson, J. (2017). Genomics of apicomplexan parasites. Critical Reviews in Biochemistry and Molecular Biology *52*, 254–273. 10.1080/10409238.2017.1290043.
2. Adl, S.M., Bass, D., Lane, C.E., Lukeš, J., Schoch, C.L., Smirnov, A., Agatha, S., Berney, C., Brown, M.W., Burki, F., et al. (2019). Revisions to the Classification, Nomenclature, and Diversity of Eukaryotes. J. Eukaryot. Microbiol. *66*, 4–119. 10.1111/jeu.12691.
3. Portman, N., and Šlapeta, J. (2014). The flagellar contribution to the apical complex: a new tool for the eukaryotic Swiss Army knife? Trends in Parasitology *30*, 58–64. 10.1016/j.pt.2013.12.006.
4. del Campo, J., Pons, M.J., Herranz, M., Wakeman, K.C., del Valle, J., Vermeij, M.J.A., Leander, B.S., and Keeling, P.J. (2019). Validation of a universal set of primers to study animal-associated microeukaryotic communities. Environ Microbiol *21*, 3855–3861. 10.1111/1462-2920.14733.
5. Boisard, J., and Florent, I. (2020). Why the –omic future of Apicomplexa should include gregarines. Biol. Cell *112*, 173–185. 10.1111/boc.202000006.
6. Desportes, I., and Schrével, J. eds. (2013). Treatise on zoology--anatomy, taxonomy, biology: The Gregarines. The early branching Apicomplexa. (Brill).
7. Rueckert, S., Betts, E.L., and Tsaousis, A.D. (2019). The Symbiotic Spectrum: Where Do the Gregarines Fit? Trends in Parasitology *35*, 687–694. 10.1016/j.pt.2019.06.013.
8. Templeton, T.J., Enomoto, S., Chen, W.-J., Huang, C.-G., Lancto, C.A., Abrahamsen, M.S., and Zhu, G. (2010). A Genome-Sequence Survey for *Ascogregarina taiwanensis* Supports Evolutionary Affiliation but Metabolic Diversity between a Gregarine and *Cryptosporidium*. Molecular Biology and Evolution *27*, 235–248. 10.1093/molbev/msp226.
9. Aurrecoechea, C., Barreto, A., Basenko, E.Y., Brestelli, J., Brunk, B.P., Cade, S., Crouch, K., Doherty, R., Falke, D., Fischer, S., et al. (2017). EuPathDB the eukaryotic pathogen genomics database resource. Nucleic Acids Res *45*, D581–D591. 10.1093/nar/gkw1105.
10. Janouškovec, J., Paskerova, G.G., Miroliubova, T.S., Mikhailov, K.V., Birley, T., Aleoshin, V.V., and Simdyanov, T.G. (2019). Apicomplexan-like parasites are polyphyletic and widely but selectively dependent on cryptic plastid organelles. eLife *8*, e49662. 10.7554/eLife.49662.
11. Mathur, V., Kolísko, M., Hehenberger, E., Irwin, N.A.T., Leander, B.S., Kristmundsson, Á., Freeman, M.A., and Keeling, P.J. (2019). Multiple Independent Origins of Apicomplexan-Like Parasites. Current Biology *29*, 2936-2941.e5. 10.1016/j.cub.2019.07.019.
12. Mathur, V., Kwong, W.K., Husnik, F., Irwin, N.A.T., Kristmundsson, Á., Gestal, C., Freeman, M., and Keeling, P.J. (2021). Phylogenomics Identifies a New Major

Subgroup of Apicomplexans, Marosporida *class nov.* , with Extreme Apicoplast Genome Reduction. Genome Biology and Evolution *13*, evaa244. 10.1093/gbe/evaa244.
13. Salomaki, E.D., Terpis, K.X., Rueckert, S., Kotyk, M., Varadínová, Z.K., Čepička, I., Lane, C.E., and Kolisko, M. (2021). Gregarine single-cell transcriptomics reveals differential mitochondrial remodeling and adaptation in apicomplexans. BMC Biol *19*, 77. 10.1186/s12915-021-01007-2.
14. Van Beneden (1869). Sur une nouvelle espèce de Grégarine désignée sous le nom de *Gregarina gigantea*. Bulletins de l'Académie Royale de Belgique *28*, 444–456.
15. Schneider, A. (1875). Contribution à l' histoire des Grégarines des Invertébrés de Paris et de Roscoff. Arch Zool Exp Gen *4*, 493–604.
16. De Bauchamp, P. (1910). Sur une grégarine nouvelle du genre *Porospora*. C R Acad Sci Paris *151*, 997–999.
17. Hatt, P. (1931). L'évolution des Porosporides chez les mollusques. Archives de zoologie expérimentale et générale *72*, 341–415.
18. Desportes, I., I., and Theodorides, J. (1965). Ultrastructure of the Gymnospore of *Porospora* (Eugregarina, Porosporidae). C R Acad Sci Paris *260*, 1761–2.
19. Schrével, J., and Desportes, I. (2015). Gregarines. In Encyclopedia of Parasitology, H. Mehlhorn, ed. (Springer Berlin Heidelberg), pp. 1–47. 10.1007/978-3-642-27769-6_1335-2.
20. King, C., and Sleep, J. (2005). Modelling the mechanism of gregarine gliding using bead translocation. J Eukaryotic Microbiology *52*, 7S-27S. 10.1111/j.1550-7408.2005.05202003_1_41.x.
21. Russell, D.G. (1983). Host cell invasion by Apicomplexa: an expression of the parasite's contractile system? Parasitology *87*, 199–209. 10.1017/S0031182000052562.
22. King, C.A. (1988). Cell motility of sporozoan protozoa. Parasitology Today *4*, 315–319. 10.1016/0169-4758(88)90113-5.
23. Sibley, L.D., Håkansson, S., and Carruthers, V.B. (1998). Gliding motility: An efficient mechanism for cell penetration. Current Biology *8*, R12–R14. 10.1016/S0960-9822(98)70008-9.
24. Opitz, C., and Soldati, D. (2002). 'The glideosome': a dynamic complex powering gliding motion and host cell invasion by *Toxoplasma gondii*: Mechanism of host cell invasion by the Apicomplexa. Molecular Microbiology *45*, 597–604. 10.1046/j.1365-2958.2002.03056.x.
25. Keeley, A., and Soldati, D. (2004). The glideosome: a molecular machine powering motility and host-cell invasion by Apicomplexa. Trends in Cell Biology *14*, 528–532. 10.1016/j.tcb.2004.08.002.
26. Frénal, K., Dubremetz, J.-F., Lebrun, M., and Soldati-Favre, D. (2017). Gliding motility powers invasion and egress in Apicomplexa. Nature Reviews Microbiology *15*, 645–660. 10.1038/nrmicro.2017.86.
27. Valigurová, A., Vaškovicová, N., Musilová, N., and Schrével, J. (2013). The enigma of eugregarine epicytic folds: where gliding motility originates? Front Zool *10*, 57. 10.1186/1742-9994-10-57.
28. Léger, L., and Duboscq, O. (1909). Etude sur la sexualité des Grégarines. Arch.

Protistenk. *17*, 19–134.
29. Seppey, M., Manni, M., and Zdobnov, E.M. (2019). BUSCO: Assessing Genome Assembly and Annotation Completeness. In Gene Prediction Methods in Molecular Biology. (Springer New York), pp. 227–245. 10.1007/978-1-4939-9173-0_14.
30. Kumar, S., Stecher, G., Suleski, M., and Hedges, S.B. (2017). TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Molecular Biology and Evolution *34*, 1812–1819. 10.1093/molbev/msx116.
31. Cornillot, E., Hadj-Kaddour, K., Dassouli, A., Noel, B., Ranwez, V., Vacherie, B., Augagneur, Y., Brès, V., Duclos, A., Randazzo, S., et al. (2012). Sequencing of the smallest Apicomplexan genome from the human pathogen *Babesia microti*. Nucleic Acids Research *40*, 9102–9114. 10.1093/nar/gks700.
32. Neafsey, D.E., Hartl, D.L., and Berriman, M. (2005). Evolution of Noncoding and Silent Coding Sites in the *Plasmodium falciparum* and *Plasmodium reichenowi* Genomes. Molecular Biology and Evolution *22*, 1621–1626. 10.1093/molbev/msi154.
33. Reid, A.J., Vermont, S.J., Cotton, J.A., Harris, D., Hill-Cawthorne, G.A., Könen-Waisman, S., Latham, S.M., Mourier, T., Norton, R., Quail, M.A., et al. (2012). Comparative Genomics of the Apicomplexan Parasites *Toxoplasma gondii* and *Neospora caninum*: Coccidia Differing in Host Range and Transmission Strategy. PLoS Pathogens *8*, e1002567. 10.1371/journal.ppat.1002567.
34. Ricklefs, R.E., and Outlaw, D.C. (2010). A Molecular Clock for Malaria Parasites. Science *329*, 226–229. 10.1126/science.1188954.
35. Hayakawa, T., Tachibana, S.-I., Hikosaka, K., Arisue, N., Matsui, A., Horii, T., and Tanabe, K. (2012). Age of the last common ancestor of extant *Plasmodium* parasite lineages. Gene *502*, 36–39. 10.1016/j.gene.2012.04.037.
36. Crandall, K.A., Pérez-Losada, M., and Porter, M.L. (2009). Crabs, shrimps, and lobsters (Decapoda). In The Timetree Of Life (Oxford University Press), p. 551.
37. Bracken-Grissom, H.D., Ahyong, S.T., Wilkinson, R.D., Feldmann, R.M., Schweitzer, C.E., Breinholt, J.W., Bendall, M., Palero, F., Chan, T.-Y., Felder, D.L., et al. (2014). The Emergence of Lobsters: Phylogenetic Relationships, Morphological Evolution and Divergence Time Comparisons of an Ancient Group (Decapoda: Achelata, Astacidea, Glypheidea, Polychelida). Systematic Biology *63*, 457–479. 10.1093/sysbio/syu008.
38. Simdyanov, T.G., Diakin, A.Y., and Aleoshin, V.V. (2015). Ultrastructure and 28S rDNA Phylogeny of Two Gregarines: *Cephaloidophora cf. communis* and *Heliospora cf. longissima* with Remarks on Gregarine Morphology and Phylogenetic Analysis. Acta Protozoologica *54*, 241–262. 10.4467/16890027AP.15.020.3217.
39. Schrével, J., Valigurová, A., Prensier, G., Chambouvet, A., Florent, I., and Guillou, L. (2016). Ultrastructure of *Selenidium pendula*, the Type Species of Archigregarines, and Phylogenetic Relations to Other Marine Apicomplexa. Protist *167*, 339–368. 10.1016/j.protis.2016.06.001.
40. Rueckert, S., Simdyanov, T.G., Aleoshin, V.V., and Leander, B.S. (2011). Identification of a Divergent Environmental DNA Sequence Clade Using the Phylogeny of Gregarine Parasites (Apicomplexa) from Crustacean Hosts. PLoS ONE *6*, e18163. 10.1371/journal.pone.0018163.

41. Mulec, J., and Summers Engel, A. (2019). Karst spring microbial mat microeukaryotic diversity differs across an oxygen-sulphide ecocline and reveals potential for novel taxa discovery. AC *48*. 10.3986/ac.v48i1.4949.
42. Skillman, K.M., Diraviyam, K., Khan, A., Tang, K., Sept, D., and Sibley, L.D. (2011). Evolutionarily Divergent, Unstable Filamentous Actin Is Essential for Gliding Motility in Apicomplexan Parasites. PLoS Pathog *7*, e1002280. 10.1371/journal.ppat.1002280.
43. Plattner, F., Yarovinsky, F., Romero, S., Didry, D., Carlier, M.-F., Sher, A., and Soldati-Favre, D. (2008). *Toxoplasma* Profilin Is Essential for Host Cell Invasion and TLR11-Dependent Induction of an Interleukin-12 Response. Cell Host & Microbe *3*, 77–87. 10.1016/j.chom.2008.01.001.
44. Pino, P., Sebastian, S., Kim, E.A., Bush, E., Brochet, M., Volkmann, K., Kozlowski, E., Llinás, M., Billker, O., and Soldati-Favre, D. (2012). A Tetracycline-Repressible Transactivator System to Study Essential Genes in Malaria Parasites. Cell Host & Microbe *12*, 824–834. 10.1016/j.chom.2012.10.016.
45. Skillman, K.M., Daher, W., Ma, C.I., Soldati-Favre, D., and Sibley, L.D. (2012). *Toxoplasma gondii* Profilin Acts Primarily To Sequester G-Actin While Formins Efficiently Nucleate Actin Filament Formation *in Vitro*. Biochemistry *51*, 2486–2495. 10.1021/bi201704y.
46. Mehta, S., and Sibley, L.D. (2011). Actin depolymerizing factor controls actin turnover and gliding motility in *Toxoplasma gondii*. MBoC *22*, 1290–1299. 10.1091/mbc.e10-12-0939.
47. Tosetti, N., Pacheco, N.D.S., Soldati-Favre, D., and Jacot, D. (2019). Three F-actin assembly centers regulate organelle inheritance, cell-cell communication and motility in. 32.
48. Daher, W., Plattner, F., Carlier, M.-F., and Soldati-Favre, D. (2010). Concerted Action of Two Formins in Gliding Motility and Host Cell Invasion by *Toxoplasma gondii*. PLoS Pathog *6*, e1001132. 10.1371/journal.ppat.1001132.
49. Baum, J., Tonkin, C.J., Paul, A.S., Rug, M., Smith, B.J., Gould, S.B., Richard, D., Pollard, T.D., and Cowman, A.F. (2008). A Malaria Parasite Formin Regulates Actin Polymerization and Localizes to the Parasite-Erythrocyte Moving Junction during Invasion. Cell Host & Microbe *3*, 188–198. 10.1016/j.chom.2008.02.006.
50. Ganter, M., SchÃ¼ler, H., and Matuschewski, K. (2009). Vital role for the *Plasmodium* actin capping protein (CP) beta-subunit in motility of malaria sporozoites: Plasmodium actin capping protein. Molecular Microbiology *74*, 1356–1367. 10.1111/j.1365-2958.2009.06828.x.
51. Drewry, L.L., and Sibley, L.D. (2015). *Toxoplasma* Actin Is Required for Efficient Host Cell Invasion. mBio *6*, e00557-15. 10.1128/mBio.00557-15.
52. Egarter, S., Andenmatten, N., Jackson, A.J., Whitelaw, J.A., Pall, G., Black, J.A., Ferguson, D.J.P., Tardieux, I., Mogilner, A., and Meissner, M. (2014). The *Toxoplasma* Acto-MyoA Motor Complex Is Important but Not Essential for Gliding Motility and Host Cell Invasion. PLoS ONE *9*, e91819. 10.1371/journal.pone.0091819.
53. Whitelaw, J.A., Latorre-Barragan, F., Gras, S., Pall, G.S., Leung, J.M., Heaslip, A., Egarter, S., Andenmatten, N., Nelson, S.R., Warshaw, D.M., et al. (2017). Surface

attachment, promoted by the actomyosin system of *Toxoplasma gondii* is important for efficient gliding motility and invasion. BMC Biology *15*. 10.1186/s12915-016-0343-5.
54. Meissner, M., Schluter, D., and Soldati, D. (2002). Role of *Toxoplasma gondii* Myosin A in Powering Parasite Gliding and Host Cell Invasion. Science *298*, 837–840. 10.1126/science.1074553.
55. Frénal, K., Marq, J.-B., Jacot, D., Polonais, V., and Soldati-Favre, D. (2014). Plasticity between MyoC- and MyoA-Glideosomes: An Example of Functional Compensation in *Toxoplasma gondii* Invasion. PLoS Pathogens *10*, e1004504. 10.1371/journal.ppat.1004504.
56. Siden-Kiamos, I., Ganter, M., Kunze, A., Hliscs, M., Steinbüchel, M., Mendoza, J., Sinden, R.E., Louis, C., and Matuschewski, K. (2011). Stage-specific depletion of myosin A supports an essential role in motility of malarial ookinetes: Promoter swap to study *Plasmodium* myosin A function. Cellular Microbiology *13*, 1996–2006. 10.1111/j.1462-5822.2011.01686.x.
57. Herm-Gotz, A. (2002). *Toxoplasma gondii* myosin A and its light chain: a fast, single-headed, plus-end-directed motor. The EMBO Journal *21*, 2149–2158. 10.1093/emboj/21.9.2149.
58. Bergman, L.W. (2003). Myosin A tail domain interacting protein (MTIP) localizes to the inner membrane complex of *Plasmodium* sporozoites. Journal of Cell Science *116*, 39–49. 10.1242/jcs.00194.
59. Gaskins, E., Gilk, S., DeVore, N., Mann, T., Ward, G., and Beckers, C. (2004). Identification of the membrane receptor of a class XIV myosin in *Toxoplasma gondii* . Journal of Cell Biology *165*, 383–393. 10.1083/jcb.200311137.
60. Baum, J., Papenfuss, A.T., Baum, B., Speed, T.P., and Cowman, A.F. (2006). Regulation of apicomplexan actin-based motility. Nature Reviews Microbiology *4*, 621–628. 10.1038/nrmicro1465.
61. Frénal, K., Polonais, V., Marq, J.-B., Stratmann, R., Limenitakis, J., and Soldati-Favre, D. (2010). Functional Dissection of the Apicomplexan Glideosome Molecular Architecture. Cell Host & Microbe *8*, 343–357. 10.1016/j.chom.2010.09.002.
62. Tardieux, I., and Baum, J. (2016). Reassessing the mechanics of parasite motility and host-cell invasion. Journal of Cell Biology *214*, 507–515. 10.1083/jcb.201605100.
63. Bullen, H.E., Tonkin, C.J., O'Donnell, R.A., Tham, W.-H., Papenfuss, A.T., Gould, S., Cowman, A.F., Crabb, B.S., and Gilson, P.R. (2009). A Novel Family of Apicomplexan Glideosome-associated Proteins with an Inner Membrane-anchoring Role. J. Biol. Chem. *284*, 25353–25363. 10.1074/jbc.M109.036772.
64. Graindorge, A., Frénal, K., Jacot, D., Salamun, J., Marq, J.B., and Soldati-Favre, D. (2016). The Conoid Associated Motor MyoH Is Indispensable for *Toxoplasma gondii* Entry and Exit from Host Cells. PLOS Pathogens *12*, e1005388. 10.1371/journal.ppat.1005388.
65. Paing, M.M., and Tolia, N.H. (2014). Multimeric Assembly of Host-Pathogen Adhesion Complexes Involved in Apicomplexan Invasion. PLoS Pathog *10*, e1004120. 10.1371/journal.ppat.1004120.

66. Boucher, L.E., and Bosch, J. (2015). The apicomplexan glideosome and adhesins – Structures and function. Journal of Structural Biology *190*, 93–114. 10.1016/j.jsb.2015.02.008.
67. Jacot, D., Waller, R.F., Soldati-Favre, D., MacPherson, D.A., and MacRae, J.I. (2016). Apicomplexan Energy Metabolism: Carbon Source Promiscuity and the Quiescence Hyperbole. Trends in Parasitology *32*, 56–70. 10.1016/j.pt.2015.09.001.
68. Sultan, A.A., Thathy, V., Frevert, U., Robson, K.J.H., Crisanti, A., Nussenzweig, V., Nussenzweig, R.S., and Ménard, R. (1997). TRAP Is Necessary for Gliding Motility and Infectivity of *Plasmodium* Sporozoites. Cell *90*, 511–522. 10.1016/S0092-8674(00)80511-5.
69. Huynh, M.-H., and Carruthers, V.B. (2006). *Toxoplasma* MIC2 Is a Major Determinant of Invasion and Virulence. PLoS Pathog *2*, e84. 10.1371/journal.ppat.0020084.
70. Buguliskis, J.S., Brossier, F., Shuman, J., and Sibley, L.D. (2010). Rhomboid 4 (ROM4) Affects the Processing of Surface Adhesins and Facilitates Host Cell Invasion by *Toxoplasma gondii*. PLoS Pathog *6*. 10.1371/journal.ppat.1000858.
71. Shen, B., Buguliskis, J.S., Lee, T.D., and Sibley, L.D. (2014). Functional Analysis of Rhomboid Proteases during *Toxoplasma* Invasion. mBio *5*, e01795-14. 10.1128/mBio.01795-14.
72. Rugarabamu, G., Marq, J.-B., Guérin, A., Lebrun, M., and Soldati-Favre, D. (2015). Distinct contribution of *Toxoplasma gondii* rhomboid proteases 4 and 5 to micronemal protein protease 1 activity during invasion: ROM4 and ROM5 contribute to MPP1 activity. Molecular Microbiology *97*, 244–262. 10.1111/mmi.13021.
73. Kappe, S., Bruderer, T., Gantt, S., Fujioka, H., Nussenzweig, V., and Ménard, R. (1999). Conservation of a Gliding Motility and Cell Invasion Machinery in Apicomplexan Parasites. Journal of Cell Biology *147*, 937–944. 10.1083/jcb.147.5.937.
74. Morahan, B.J., Wang, L., and Coppel, R.L. (2009). No TRAP, no invasion. Trends in Parasitology *25*, 77–84. 10.1016/j.pt.2008.11.004.
75. Templeton, T.J., and Pain, A. (2016). Diversity of extracellular proteins during the transition from the 'proto-apicomplexan' alveolates to the apicomplexan obligate parasites. Parasitology *143*, 1–17. 10.1017/S0031182015001213.
76. Dessens, J.T., Beetsma, A.L., Dimopoulos, G., Wengelnik, K., Crisanti, A., Kafatos, F.C., and Sinden, R.E. (1999). CTRP is essential for mosquito infection by malaria ookinetes. EMBO J *18*, 6221–6227. 10.1093/emboj/18.22.6221.
77. Bargieri, D.Y. (2016). *Plasmodium* Merozoite TRAP Family Protein Is Essential for Vacuole Membrane Disruption and Gamete Egress from Erythrocytes. Cell Host Microbe *20*, 618–630. 10.1016/j.chom.2016.10.015.
78. Lacroix, C., and Ménard, R. (2008). TRAP-like protein of *Plasmodium* sporozoites: linking gliding motility to host-cell traversal. Trends in Parasitology *24*, 431–434. 10.1016/j.pt.2008.07.003.
79. Deng, M., Templeton, T.J., London, N.R., Bauer, C., Schroeder, A.A., and Abrahamsen, M.S. (2002). *Cryptosporidium parvum* Genes Containing Thrombospondin Type 1 Domains. IAI *70*, 6987–6995. 10.1128/IAI.70.12.6987-

6995.2002.
80. Putignani, L., Possenti, A., Cherchi, S., Pozio, E., Crisanti, A., and Spano, F. (2008). The thrombospondin-related protein CpMIC1 (CpTSP8) belongs to the repertoire of micronemal proteins of *Cryptosporidium parvum*. Molecular and Biochemical Parasitology *157*, 98–101. 10.1016/j.molbiopara.2007.09.004.
81. Gaffar, F.R., Yatsuda, A.P., Franssen, F.F.J., and Vries, E. de (2004). A Babesia bovis merozoite protein with a domain architecture highly similar to the thrombospondin-related anonymous protein (TRAP) present in *Plasmodium* sporozoites. Molecular and Biochemical Parasitology *136*, 25–34. 10.1016/j.molbiopara.2004.02.006.
82. Zhou, J., Fukumoto, S., Jia, H., Yokoyama, N., Zhang, G., Fujisaki, K., Lin, J., and Xuan, X. (2006). Characterization of the *Babesia gibsoni* P18 as a homologue of thrombospondin related adhesive protein. Molecular and Biochemical Parasitology *148*, 190–198. 10.1016/j.molbiopara.2006.03.015.
83. Yu, L., Liu, Q., Zhan, X., Huang, Y., Sun, Y., Nie, Z., Zhao, Y., An, X., Li, M., Wang, S., et al. (2018). Identification and molecular characterization of a novel *Babesia orientalis* thrombospondin-related anonymous protein (BoTRAP1). Parasites Vectors *11*, 667. 10.1186/s13071-018-3245-2.
84. Montenegro, V.N., Paoletta, M.S., Jaramillo Ortiz, J.M., Suarez, C.E., and Wilkowsky, S.E. (2020). Identification and characterization of a *Babesia bigemina* thrombospondin-related superfamily member, TRAP-1: a novel antigen containing neutralizing epitopes involved in merozoite invasion. Parasites Vectors *13*, 602. 10.1186/s13071-020-04469-5.
85. Lovett, J. (2000). Molecular characterization of a thrombospondin-related anonymous protein homologue in *Neospora caninum*. Molecular and Biochemical Parasitology *107*, 33–43. 10.1016/S0166-6851(99)00228-5.
86. Clarke, L.E., Tomley, F.M., Wisher, M.H., Foulds, I.J., and Boursnell, M.E. (1990). Regions of an *Eimeria tenella* antigen contain sequences which are conserved in circumsporozoite proteins from *Plasmodium* spp. and which are related to the thrombospondin gene family. Mol Biochem Parasitol *41*, 269–279. 10.1016/0166-6851(90)90190-w.
87. Witcombe, D.M., Belli, S.I., Wallach, M.G., and Smith, N.C. (2003). Molecular characterisation of EmTFP250: a novel member of the TRAP protein family in *Eimeria maxima*. International Journal for Parasitology *33*, 691–702. 10.1016/S0020-7519(03)00086-9.
88. Bichet, M., Joly, C., Hadj Henni, A., Guilbert, T., Xémard, M., Tafani, V., Lagal, V., Charras, G., and Tardieux, I. (2014). The toxoplasma-host cell junction is anchored to the cell cortex to sustain parasite invasive force. BMC Biol *12*, 773. 10.1186/s12915-014-0108-y.
89. Portes, J., Barrias, E., Travassos, R., Attias, M., and de Souza, W. (2020). *Toxoplasma gondii* Mechanisms of Entry Into Host Cells. Front. Cell. Infect. Microbiol. *10*, 294. 10.3389/fcimb.2020.00294.
90. Yang, A.S.P., Lopaticki, S., O'Neill, M.T., Erickson, S.M., Douglas, D.N., Kneteman, N.M., and Boddey, J.A. (2017). AMA1 and MAEBL are important for *Plasmodium*

falciparum sporozoite infection of the liver. Cellular Microbiology *19*, e12745. 10.1111/cmi.12745.
91. O'Hara, S.P., and Chen, X.-M. (2011). The cell biology of *Cryptosporidium* infection. Microbes Infect *13*, 721–730. 10.1016/j.micinf.2011.03.008.
92. Singh, P., Mirdha, B.R., Srinivasan, A., Rukmangadachar, L.A., Singh, S., Sharma, P., Hariprasad, G., Gururao, H., Gururao, H., and Luthra, K. (2015). Identification of invasion proteins of *Cryptosporidium parvum*. World J Microbiol Biotechnol *31*, 1923–1934. 10.1007/s11274-015-1936-9.
93. Lourido, S., and Moreno, S.N.J. (2015). The calcium signaling toolkit of the Apicomplexan parasites *Toxoplasma gondii* and *Plasmodium* spp. Cell Calcium *57*, 186–193. 10.1016/j.ceca.2014.12.010.
94. Ghartey-Kwansah, G., Yin, Q., Li, Z., Gumpper, K., Sun, Y., Yang, R., Wang, D., Jones, O., Zhou, X., Wang, L., et al. (2020). Calcium-dependent Protein Kinases in Malaria Parasite Development and Infection. Cell Transplant *29*, 096368971988488. 10.1177/0963689719884888.
95. Bullen, H.E., Jia, Y., Yamaryo-Botté, Y., Bisio, H., Zhang, O., Jemelin, N.K., Marq, J.-B., Carruthers, V., Botté, C.Y., and Soldati-Favre, D. (2016). Phosphatidic Acid-Mediated Signaling Regulates Microneme Secretion in *Toxoplasma*. Cell Host & Microbe *19*, 349–360. 10.1016/j.chom.2016.02.006.
96. Darvill, N., Dubois, D.J., Rouse, S.L., Hammoudi, P.-M., Blake, T., Benjamin, S., Liu, B., Soldati-Favre, D., and Matthews, S. (2018). Structural Basis of Phosphatidic Acid Sensing by APH in Apicomplexan Parasites. Structure *26*, 1059–1071. 10.1016/j.str.2018.05.001.
97. Farrell, A., Thirugnanam, S., Lorestani, A., Dvorin, J.D., Eidell, K.P., Ferguson, D.J.P., Anderson-White, B.R., Duraisingh, M.T., Marth, G.T., and Gubbels, M.-J. (2012). A DOC2 Protein Identified by Mutational Profiling Is Essential for Apicomplexan Parasite Exocytosis. Science *335*, 218–221. 10.1126/science.1210829.
98. Heaslip, A.T., Nishi, M., Stein, B., and Hu, K. (2011). The Motility of a Human Parasite, *Toxoplasma gondii*, Is Regulated by a Novel Lysine Methyltransferase. PLoS Pathog *7*, e1002201. 10.1371/journal.ppat.1002201.
99. Guo, Y., Tang, K., Rowe, L.A., Li, N., Roellig, D.M., Knipe, K., Frace, M., Yang, C., Feng, Y., and Xiao, L. (2015). Comparative genomic analysis reveals occurrence of genetic recombination in virulent *Cryptosporidium hominis* subtypes and telomeric gene duplications in *Cryptosporidium parvum*. BMC Genomics *16*. 10.1186/s12864-015-1517-1.
100. Gras, S., Jackson, A., Woods, S., Pall, G., Whitelaw, J., Leung, J.M., Ward, G.E., Roberts, C.W., and Meissner, M. (2017). Parasites lacking the micronemal protein MIC2 are deficient in surface attachment and host cell egress, but remain virulent in vivo. Wellcome Open Res *2*, 32. 10.12688/wellcomeopenres.11594.2.
101. Harding, C.R., Gow, M., Kang, J.H., Shortt, E., Manalis, S.R., Meissner, M., and Lourido, S. (2019). Alveolar proteins stabilize cortical microtubules in *Toxoplasma gondii*. Nature Communications *10*. 10.1038/s41467-019-08318-7.
102. Rompikuntal, P.K., Foe, I.T., Deng, B., Bogyo, M., and Ward, G.E. (2020). Blocking palmitoylation of *Toxoplasma gondii* myosin light chain 1 disrupts glideosome

composition but has little impact on parasite motility (Microbiology) 10.1101/2020.08.13.250399.
103. Valigurová, A., Vaškovicová, N., Diakin, A., Paskerova, G.G., Simdyanov, T.G., and Kováčiková, M. (2017). Motility in blastogregarines (Apicomplexa): Native and drug-induced organisation of *Siedleckia nematoides* cytoskeletal elements. PLOS ONE *12*, e0179709. 10.1371/journal.pone.0179709.
104. Heintzelman, M.B. (2004). Actin and myosin in *Gregarina polymorpha*. Cell Motility and the Cytoskeleton *58*, 83–95. 10.1002/cm.10178.
105. Heintzelman, M.B., and Mateer, M.J. (2008). GpMyoF, a WD40 Repeat-Containing Myosin Associated with the Myonemes of *Gregarina polymorpha*. Journal of Parasitology *94*, 158–168. 10.1645/GE-1339.1.
106. Kováčiková, M., Simdyanov, T.G., Diakin, A., and Valigurová, A. (2017). Structures related to attachment and motility in the marine eugregarine *Cephaloidophora cf. communis* (Apicomplexa). European Journal of Protistology *59*, 1–13. 10.1016/j.ejop.2017.02.006.
107. Diakin, A., Wakeman, K.C., and Valigurová, A. (2017). Description of *Ganymedes yurii* sp. n. (Ganymedidae), a New Gregarine Species from the Antarctic Amphipod *Gondogeneia* sp. (Crustacea). Journal of Eukaryotic Microbiology *64*, 56–66. 10.1111/jeu.12336.
108. Woo, Y.H., Ansari, H., Otto, T.D., Klinger, C.M., Kolisko, M., Michálek, J., Saxena, A., Shanmugam, D., Tayyrov, A., Veluchamy, A., et al. (2015). Chromerid genomes reveal the evolutionary path from photosynthetic algae to obligate intracellular parasites. eLife *4*. 10.7554/eLife.06974.
109. Butler, M., Cockcroft, A., MacDiarmid, A., and Wahle, R. (2011). *Homarus gammarus*. The IUCN Red List of Threatened Species. e.T169955A69905303. 10.2305/IUCN.UK.2011-1.RLTS.T169955A69905303.en.
110. Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc.
111. Krueger, Felix (2015). Trim galore. A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files. https://github.com/FelixKrueger/TrimGalore.
112. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., et al. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology *19*, 455–477. 10.1089/cmb.2012.0021.
113. Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol *29*, 644–652. 10.1038/nbt.1883.
114. Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., et al. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc *8*, 1494–1512. 10.1038/nprot.2013.084.
115. Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B.

(2006). AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Research *34*, W435–W439. 10.1093/nar/gkl200.
116. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. Journal of Molecular Biology *215*, 403–410. 10.1016/S0022-2836(05)80360-2.
117. Langmead, B., Wilks, C., Antonescu, V., and Charles, R. (2019). Scaling read aligners to hundreds of threads on general-purpose processors. Bioinformatics *35*, 421–432. 10.1093/bioinformatics/bty648.
118. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. 10.1093/bioinformatics/btp352.
119. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. 10.1093/bioinformatics/btq033.
120. Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S.L. (2004). Versatile and open software for comparing large genomes. Genome Biol *5*, R12. 10.1186/gb-2004-5-2-r12.
121. Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics *29*, 1072–1075. 10.1093/bioinformatics/btt086.
122. Wu, T.D., and Watanabe, C.K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics *21*, 1859–1875. 10.1093/bioinformatics/bti310.
123. Haas, B.J. (2003). Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. Nucleic Acids Research *31*, 5654–5666. 10.1093/nar/gkg770.
124. Korf, I. (2004). Gene finding in novel genomes. BMC Bioinformatics *5*, 59. 10.1186/1471-2105-5-59.
125. Smit, AFA. and Hubley, R. (2015). RepeatMasker Open-4.0. http://www.repeatmasker.org.
126. Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S., and Ogata, H. (2020). KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. Bioinformatics *36*, 2251–2252. 10.1093/bioinformatics/btz859.
127. Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics *30*, 1236–1240. 10.1093/bioinformatics/btu031.
128. Li, L., Stoeckert, CJ., and Roos, D. (2003). OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Research *13*, 2178–2189. 10.1101/gr.1224503.
129. Ranwez, V., Harispe, S., Delsuc, F., and Douzery, E.J.P. (2011). MACSE: Multiple Alignment of Coding SEquences Accounting for Frameshifts and Stop Codons. PLoS ONE *6*, e22594. 10.1371/journal.pone.0022594.

130. Yang, Z., and Nielsen, R. (2000). Estimating Synonymous and Nonsynonymous Substitution Rates Under Realistic Evolutionary Models. Molecular Biology and Evolution *17*, 32–43. 10.1093/oxfordjournals.molbev.a026236.
131. Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. Molecular Biology and Evolution *24*, 1586–1591. 10.1093/molbev/msm088.
132. Nawrocki, E.P., and Eddy, S.R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics *29*, 2933–2935. 10.1093/bioinformatics/btt509.
133. Kalvari, I., Nawrocki, E.P., Argasinska, J., Quinones-Olvera, N., Finn, R.D., Bateman, A., and Petrov, A.I. (2018). Non-Coding RNA Analysis Using the Rfam Database. Current Protocols in Bioinformatics *62*, e51. 10.1002/cpbi.51.
134. de Araujo Oliveira, J.V., Costa, F., Backofen, R., Stadler, P.F., Machado Telles Walter, M.E., and Hertel, J. (2016). SnoReport 2.0: new features and a refined Support Vector Machine to improve snoRNA identification. BMC Bioinformatics *17*, 464. 10.1186/s12859-016-1345-6.
135. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357–359. 10.1038/nmeth.1923.
136. Karadjian, G., Hassanin, A., Saintpierre, B., Gembu Tungaluna, G.-C., Ariey, F., Ayala, F.J., Landau, I., and Duval, L. (2016). Highly rearranged mitochondrial genome in *Nycteria* parasites (Haemosporidia) from bats. Proceedings of the National Academy of Sciences *113*, 9834–9839. 10.1073/pnas.1610643113.
137. Castresana, J. (2000). Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. Molecular Biology and Evolution *17*, 540–552. 10.1093/oxfordjournals.molbev.a026334.
138. Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Molecular Biology and Evolution *35*, 1547–1549. 10.1093/molbev/msy096.
139. Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313. 10.1093/bioinformatics/btu033.
140. Ronquist, F., and Huelsenbeck, J.P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics *19*, 1572–1574. 10.1093/bioinformatics/btg180.
141. Rambaut (2018). FigTree. tree.bio.ed.ac.uk/software/figtree/.
142. Katoh, K., and Standley, D.M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution *30*, 772–780. 10.1093/molbev/mst010.
143. Letunic, I., Khedkar, S., and Bork, P. (2021). SMART: recent updates, new developments and status in 2020. Nucleic Acids Res *49*, D458–D460. 10.1093/nar/gkaa937.
144. Wickham, H. (2009). Ggplot2: elegant graphics for data analysis (Springer).
145. Birney, E. (2004). GeneWise and Genomewise. Genome Research *14*, 988–995. 10.1101/gr.1865504.

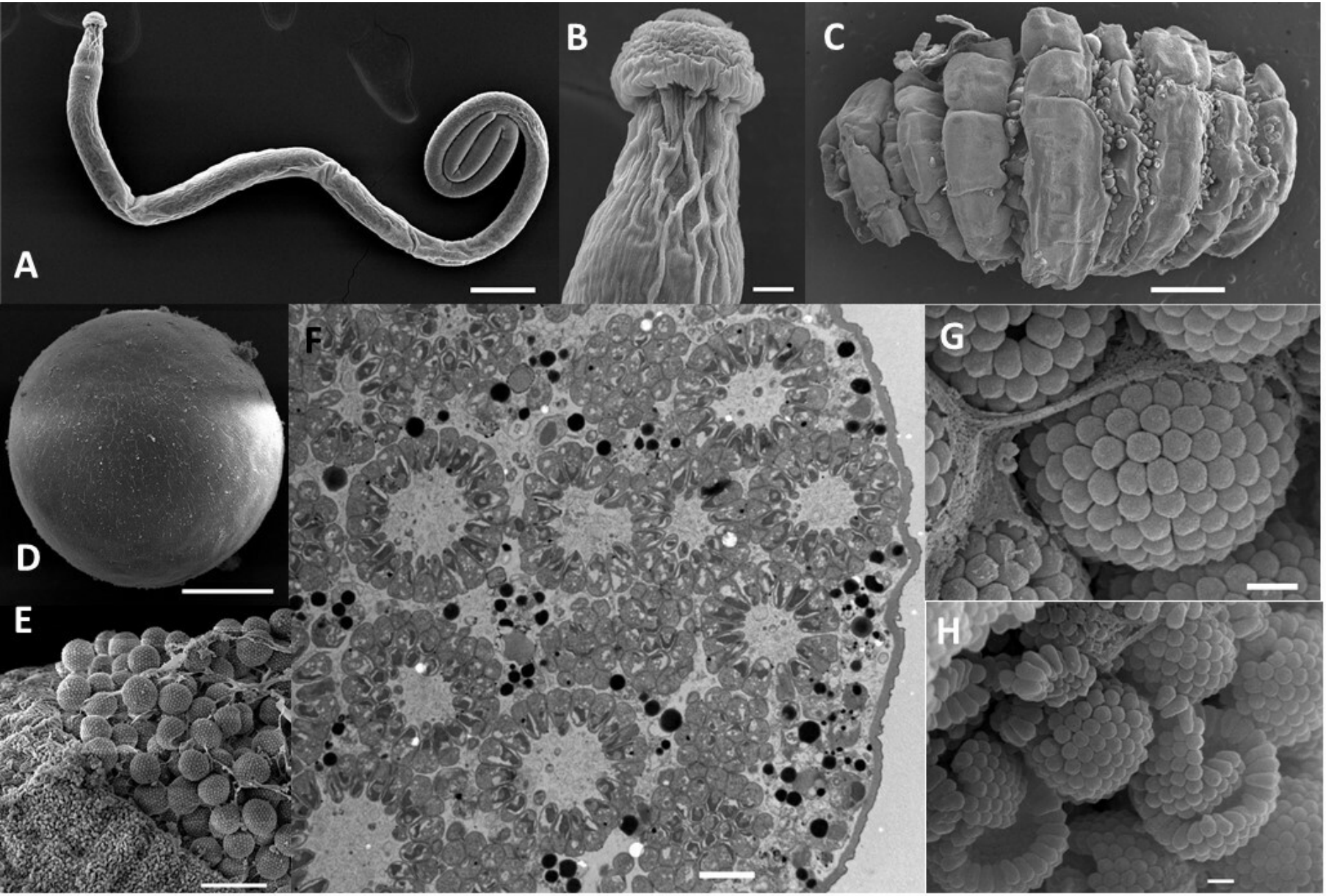


Figure 1

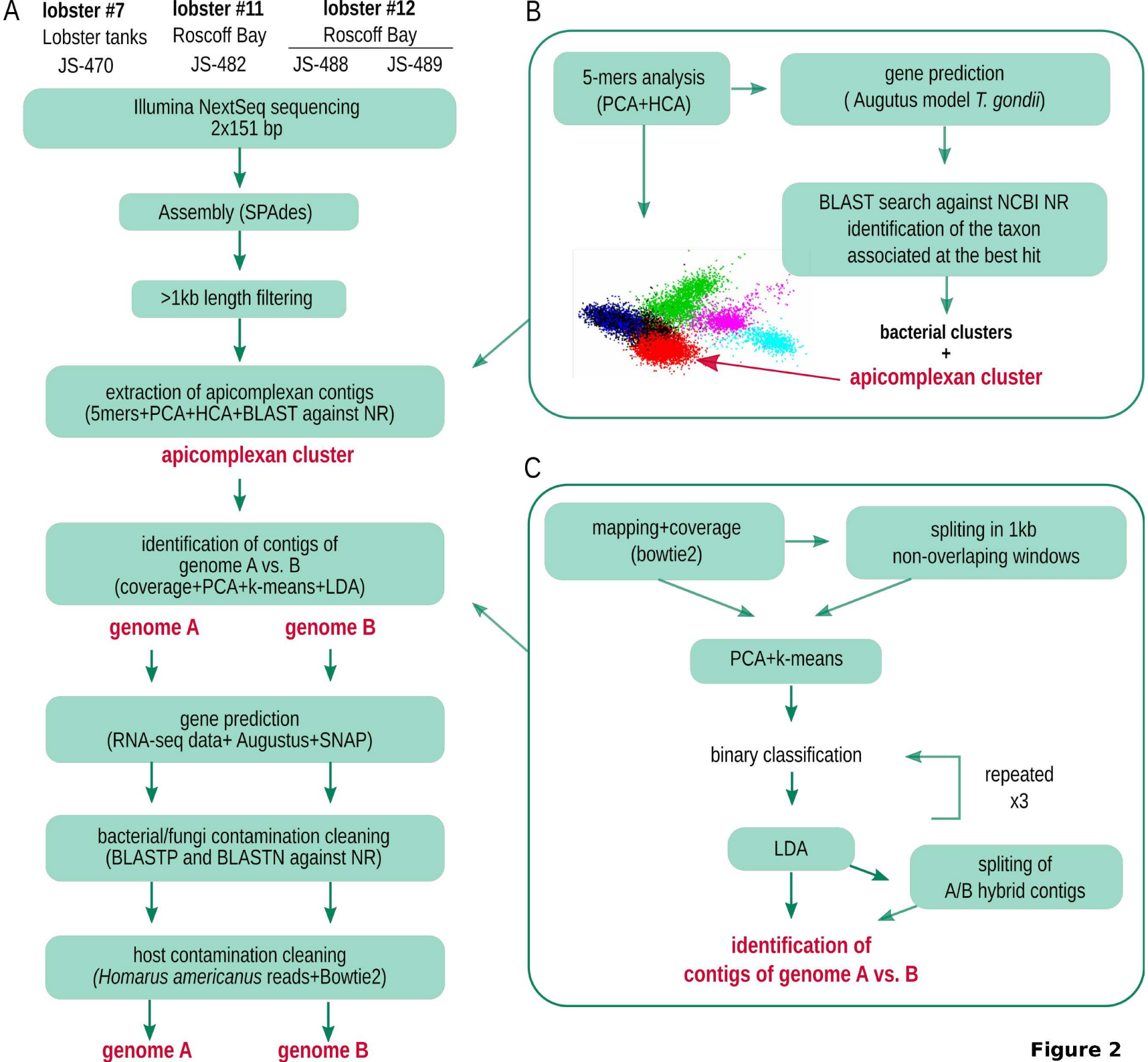

A
lobster #7
Lobster tanks
JS-470
lobster #11
Roscoff Bay
JS-482
lobster #12
Roscoff Bay
JS-488
JS-489
Illumina NextSeq sequencing
2x151 bp
Assembly (SPAdes)
>1kb length filtering
extraction of apicomplexan contigs
(5mers+PCA+HCA+BLAST against NR)
apicomplexan cluster
identification of contigs of
genome A vs. B
(coverage+PCA+k-means+LDA)
genome A
genome B
gene prediction
(RNA-seq data+ Augustus+SNAP)
bacterial/fungi contamination cleaning
(BLASTP and BLASTN against NR)
host contamination cleaning
(Homarus americanus reads+Bowtie2)
genome A
genome B
B
5-mers analysis
(PCA+HCA)
gene prediction
(Augustus model T. gondii)
BLAST search against NCBI NR
identification of the taxon
associated at the best hit
bacterial clusters
+
apicomplexan cluster
C
mapping+coverage
(bowtie2)
spliting in 1kb
non-overlaping windows
PCA+k-means
binary classification
repeated
x3
LDA
spliting of
A/B hybrid contigs
identification of
contigs of genome A vs. B

Figure 2

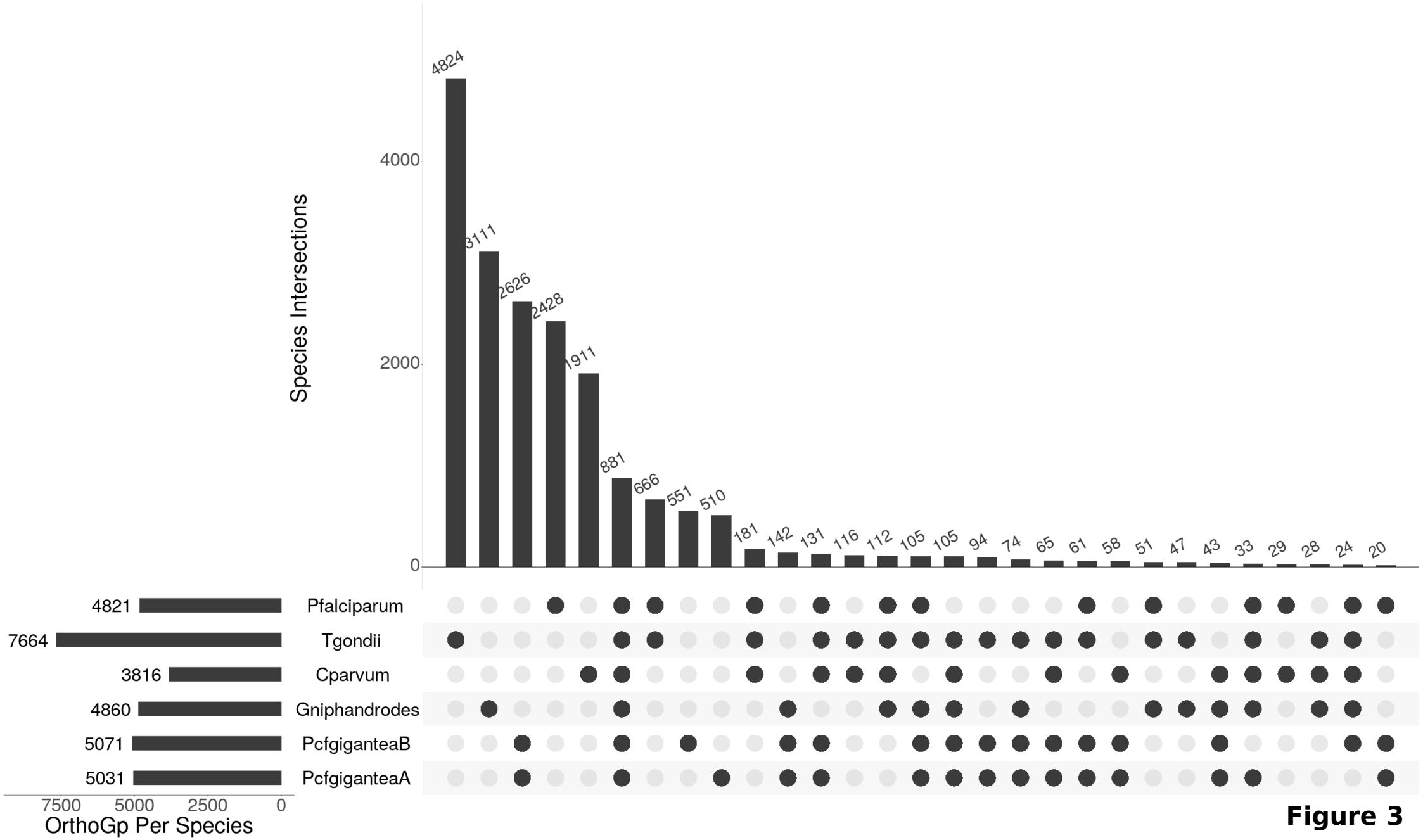

Species Intersections
4000
2000
0
4824
3111
2626
2428
1911
881
666
551
510
181
142
131
116
112
105
105
94
74
65
61
58
51
47
43
33
29
28
24
20
4821
7664
3816
4860
5071
5031
Pfalciparum
Tgondii
Cparvum
Gniphandrodes
PcfgiganteaB
PcfgiganteaA
7500
5000
2500
0
OrthoGp Per Species

Figure 3

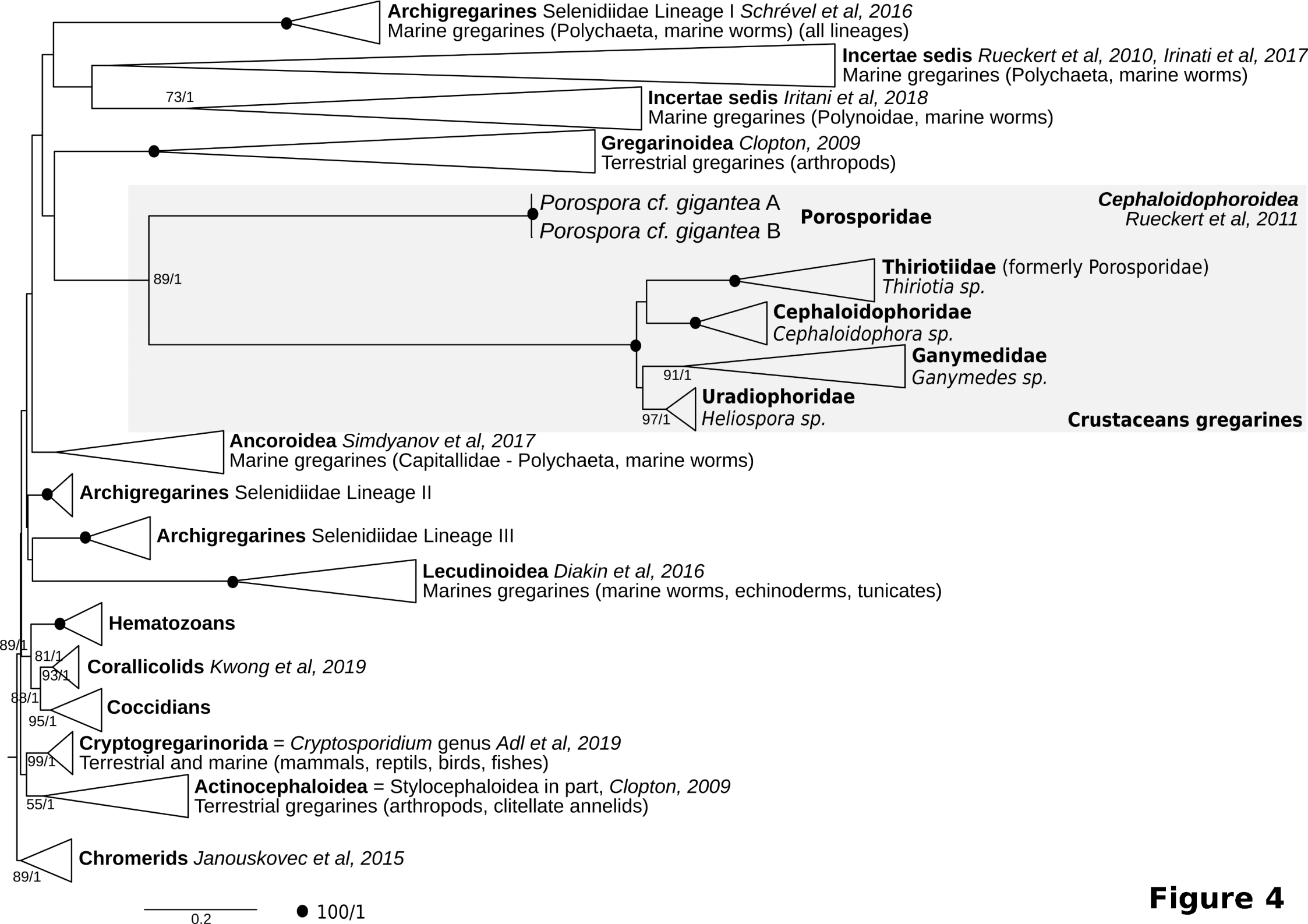
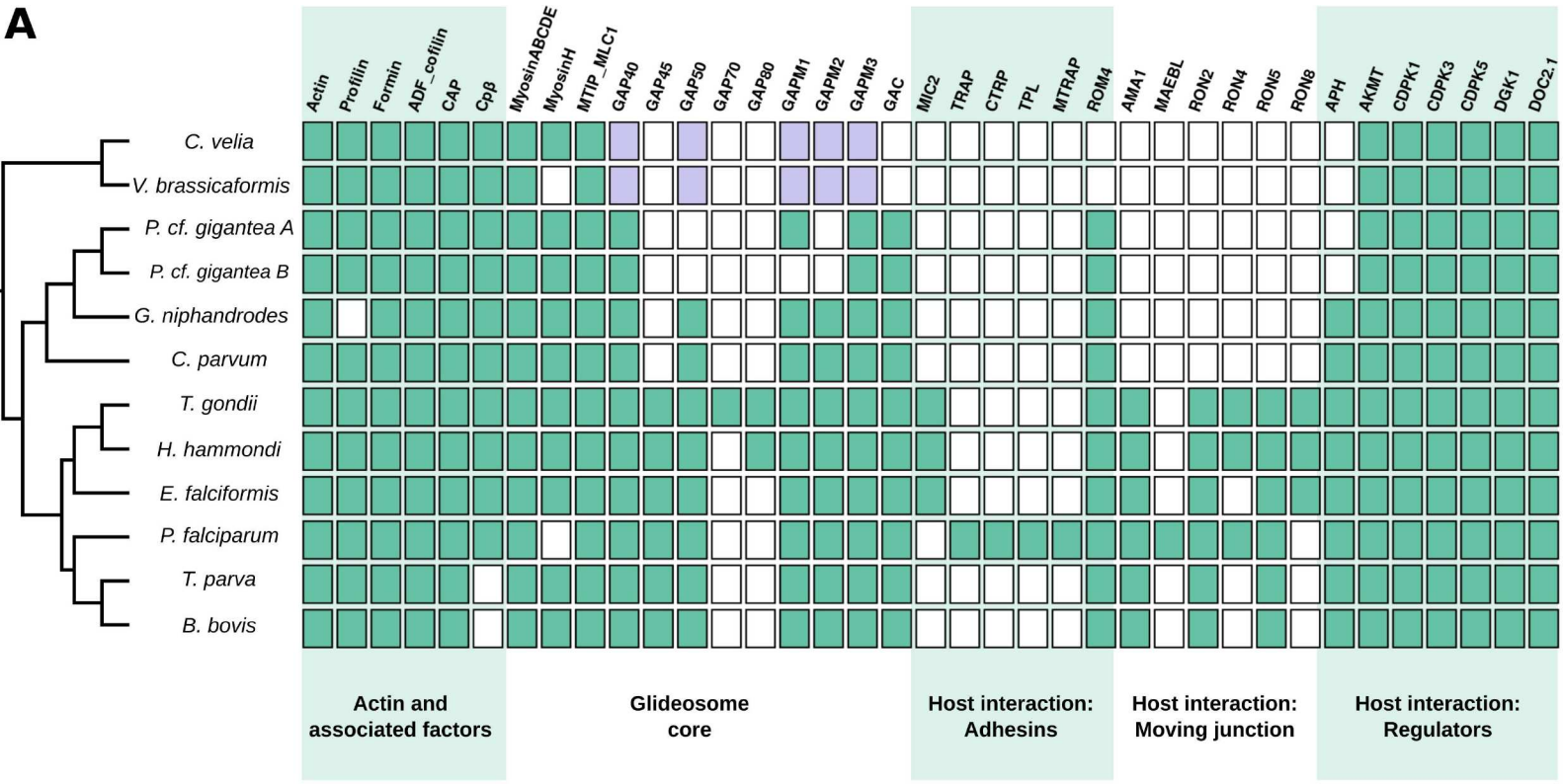

Archigregarines Selenidiidae Lineage I Schrével et al, 2016
Marine gregarines (Polychaeta, marine worms) (all lineages)
Incertae sedis Rueckert et al, 2010, Irinati et al, 2017
Marine gregarines (Polychaeta, marine worms)
73/1
Incertae sedis Iritani et al, 2018
Marine gregarines (Polynoidae, marine worms)
Gregarinoidea Clopton, 2009
Terrestrial gregarines (arthropods)
Porospora cf. gigantea A
Porospora cf. gigantea B
Porosporidae
Cephaloidophoroidea
Rueckert et al, 2011
89/1
Thiriotiidae (formerly Porosporidae)
Thiriotia sp.
Cephaloidophoridae
Cephaloidophora sp.
Ganymedidae
Ganymedes sp.
91/1
Uradiophoridae
Heliospora sp.
97/1
Crustaceans gregarines
Ancoroidea Simdyanov et al, 2017
Marine gregarines (Capitallidae - Polychaeta, marine worms)
Archigregarines Selenidiidae Lineage II
Archigregarines Selenidiidae Lineage III
Lecudinoidea Diakin et al, 2016
Marines gregarines (marine worms, echinoderms, tunicates)
Hematozoans
89/1
81/1
93/1
Corallicolids Kwong et al, 2019
88/1
Coccidians
95/1
Cryptogregarinorida = Cryptosporidium genus Adl et al, 2019
Terrestrial and marine (mammals, reptils, birds, fishes)
99/1
Actinocephaloidea = Stylocephaloidea in part, Clopton, 2009
Terrestrial gregarines (arthropods, clitellate annelids)
55/1
Chromerids Janouskovec et al, 2015
89/1
0.2
100/1

Figure 4


A
Actin
Profilin
Formin
ADF_cofilin
CAP
Cpβ
MyosinABCDE
MyosinH
MTIP_MLC1
GAP40
GAP45
GAP50
GAP70
GAP80
GAPM1
GAPM2
GAPM3
GAC
MIC2
TRAP
CTRP
TPL
MTRAP
ROM4
AMA1
MAEBL
RON2
RON4
RON5
RON8
APH
AKMT
CDPK1
CDPK3
CDPK5
DGK1
DOC2.1
C. velia
V. brassicaformis
P. cf. gigantea A
P. cf. gigantea B
G. niphandrodes
C. parvum
T. gondii
H. hammondi
E. falciformis
P. falciparum
T. parva
B. bovis
Actin and associated factors
Glideosome core
Host interaction: Adhesins
Host interaction: Moving junction
Host interaction: Regulators

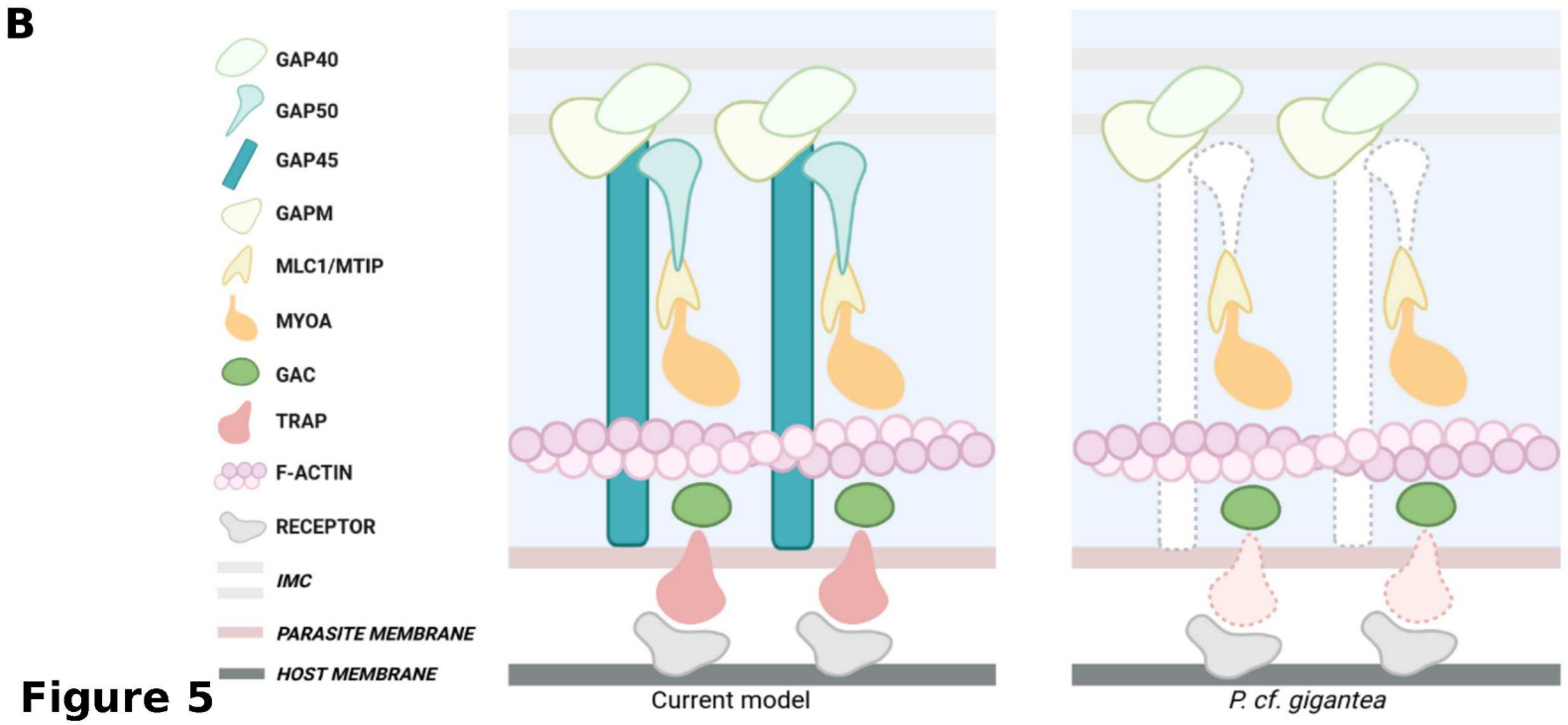

B
GAP40
GAP50
GAP45
GAPM
MLC1/MTIP
MYOA
GAC
TRAP
F-ACTIN
RECEPTOR
IMC
PARASITE MEMBRANE
HOST MEMBRANE
Current model
P. cf. gigantea

Figure 5

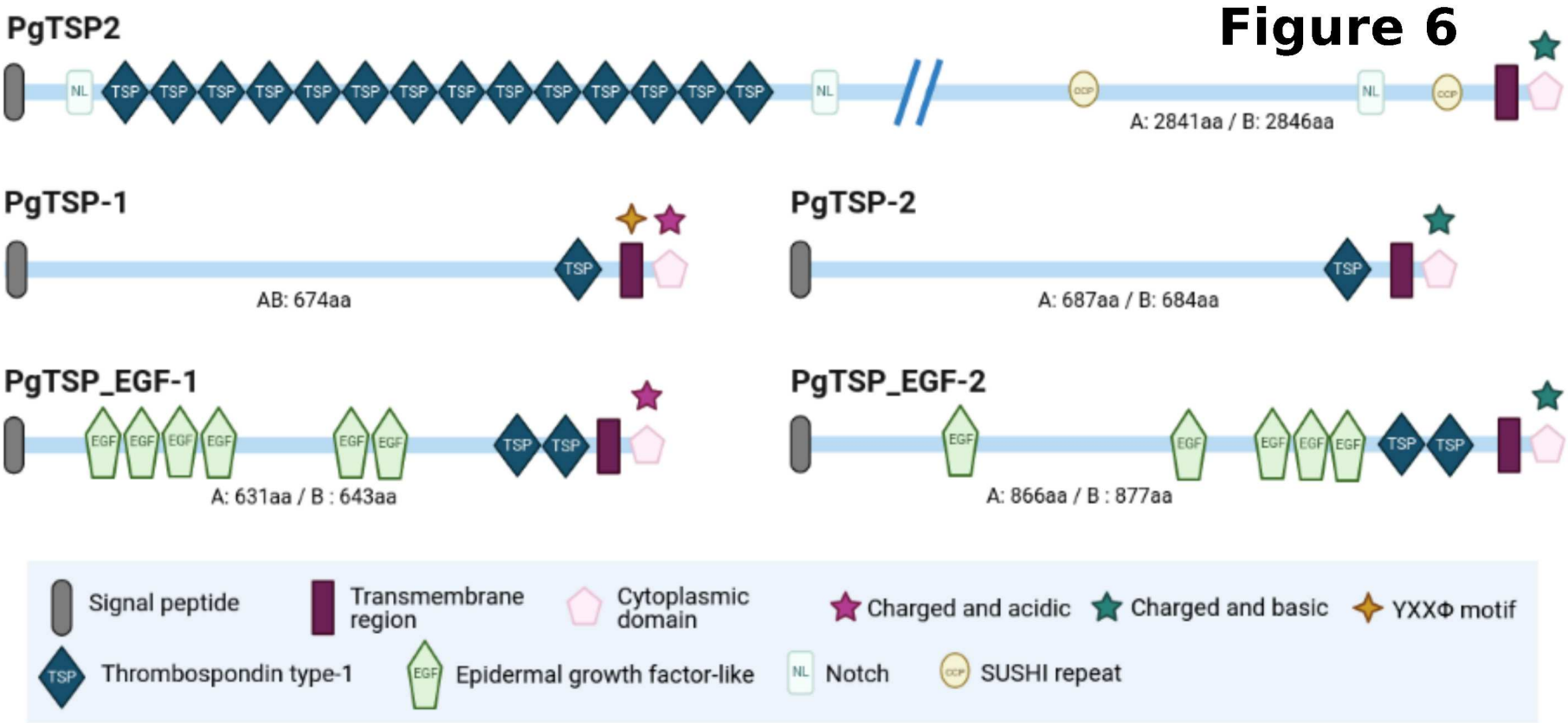


Figure 6
PgTSP2
NL
TSP
NL
CCP
NL
CCP
A: 2841aa / B: 2846aa
PgTSP-1
TSP
AB: 674aa
PgTSP-2
TSP
A: 687aa / B: 684aa
PgTSP_EGF-1
EGF
TSP
A: 631aa / B : 643aa
PgTSP_EGF-2
EGF
TSP
A: 866aa / B : 877aa
Signal peptide
Transmembrane region
Cytoplasmic domain
Charged and acidic
Charged and basic
YXXΦ motif
Thrombospondin type-1
Epidermal growth factor-like
Notch
SUSHI repeat

species	*P. cf. gigantea*		*G. niphandrodes*	*C. parvum*	*T. gondii*	*P. falciparum*	*C. velia*	*V. brassicaformis*
strain	**A**	**B**	***na***	**Iowaii**	**ME49**	**3D7**	**CCMP2878**	**CCMP3155**
nb of contigs/chromosomes	787	934	355	8	435	14	5470	1006
total length of assembly (bp)	8806768	9049943	13873624	9102324	63472444	23292622	192006978	72475329
mean length contigs/chromosomes (bp)	11190.3	9689.45	39080.63	1137790.5	145913.66	1663758.71	35101.82	72043.07
GC content (%)	54.3	54.3	53.8	30.2	52.4	19.3	49.1	58.1
nb of protein coding genes	5270	5361	6606	4020	8862	5602	30604	23412
mean length of coding genes (bp)	1438.2	1450.3	1392.6	1865.0	5602.9	2488.6	4507.6	2704.7
nb of tRNA	14	14	231	45	150	45	0	0
nb of rRNA	27	25	0	5	420	28	0	0
nb of gene with intron(s)	2957	2981	2390	575	6801	3010	21895	22163
median length of the introns (bp)	28 [27-30]	28 [27-30]	95 [56-145]	65 [51-91]	467 [322-632]	140 [110-184]	372 [273-520]	81 [70-98]
mode of intron length (bp)	28	28	37	44	55	121	320	74
mean nb of introns per gene*	1.8	1.8	1.4	1.8	5.9	2.9	5.4	7.9
non-coding DNA (%)	16	16	37	24	68	47	74	50

* by considering only genes with intron(s)

Table 1