Marine gregarine genomes reveal the breadth of apicomplexan diversity and provide new insights on gliding motility

Author List: Julie Boisard[1,2,3], Evelyne Duvernois-Berthet[4], Linda Duval[1], Joseph Schrével[1], Laure Guillou[5], Amandine Labat[1], Sophie Le Panse[6], Gérard Prensier[7], Loïc Ponger[2] and Isabelle Florent[1].

Author Contact:

Julie Boisard: julie.boisard@biol.lu.se

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. Department of Biology, Lund University, Sölvegatan 35, 223 62 Lund, Sweden
4. Muséum National d'Histoire Naturelle, Centre National de la Recherche Scientifique, Laboratoire Physiologie Moléculaire et Adaptation (PhyMA), UMR7221 CNRS-MNHN, 75005, Paris, France
5. Sorbonne Université, CNRS, UMR7144 Adaptation et Diversité en Milieu Marin, Ecology of Marine Plankton (ECOMAP), Station Biologique de Roscoff SBR, 29680 Roscoff, France
6. Plateforme d'Imagerie Merimage, FR2424, Centre National de la Recherche Scientifique, Station Biologique de Roscoff, Roscoff, 29680, France
7. Cell biology and Electron Microscopy Laboratory, François Rabelais University, 10 Boulevard Tonnellé, BP 3223 Tours Cedex, France

ABSTRACT

Our current view of the evolutionary history, coding and adaptive capacities of Apicomplexa, protozoan parasites of a wide range of metazoan, is currently strongly biased toward species infecting humans, as data on early diverging apicomplexan lineages infecting invertebrates is extremely limited. Here, we characterized the genome of the marine eugregarine *Porospora gigantea*, intestinal parasite of Lobsters, remarkable for the macroscopic size of its vegetative feeding forms (trophozoites) and its gliding speed, the fastest so far recorded for Apicomplexa. Two highly syntenic genomes named A and B were assembled. Similar in size (~9 Mb) and coding capacity (~5300 genes), A and B genomes are 10.8% divergent at the nucleotide level, corresponding to 16-38 My in divergent time. Orthogroup analysis across 25 (proto)Apicomplexa species, including *Gregarina niphandrodes*, showed that A and B are highly divergent from all other known apicomplexan species, revealing an unexpected breadth of diversity. Phylogenetically these two species branch sister to Cephaloidophoroidea, and thus expand the known crustacean gregarine superfamily. The genomes were mined for genes encoding proteins necessary for gliding, a key feature of apicomplexans parasites, currently studied through the molecular model called glideosome. Sequence analysis shows that actin-related proteins and regulatory factors are strongly conserved within apicomplexans. In contrast, the predicted protein sequences of core glideosome proteins and adhesion proteins are 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 species diversity, and to deepen our understanding of the molecular bases of key functions enabling parasitism, such as the glideosome.

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

BACKGROUND

Apicomplexans are unicellular eukaryotic microorganisms that have evolved towards endobiotic symbionts or parasites. The Apicomplexa include about 350 genera[1] for 6,000 documented species. Some species are extremely pathogenic such as *Plasmodium* spp., *Toxoplasma gondii* and *Cryptosporidium* spp., responsible for malaria, toxoplasmosis and cryptosporidiosis, respectively. Current knowledge of apicomplexan genomes is based on sequence data from a dozen genera, and more precisely, the genera which include highly pathogenic species[2]. Consequently, our view of the Apicomplexa genome is highly skewed towards intracellular parasites of vertebrates, notably Coccidia, Hemosporidia and *Cryptosporidium* (see references in Table S1). By comparison, the gregarines, of which there are at least 1,770 species[3], have hardly been explored at an omic level[4]. Gregarines were identified as the most abundant and widely reported apicomplexan in a recent environmental study[5]. However, as they have low pathogenicity and are non-cultivable in the laboratory, they have attracted less interest.

Overlooking the gregarines risks leaving part of the evolutionary history of Apicomplexa unexplored, because they represent early diverging lineages as well as displaying a diversity of specific adaptive traits. For instance, gregarines are mostly extracellular, infecting a wide diversity of marine and terrestrial non-vertebrate hosts[6,7]. At this time, available genomic data are very limited to terrestrial gregarines, such as partial data on *Ascogregarina taiwanensis,* an intestinal parasite of the tiger mosquito *Aedes albopictus*[8], and the draft genome of *Gregarina niphandrodes,* an intestinal parasite of the mealworm *Tenebrio molitor* (unpublished, available in CryptoDB[9]). 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. These developmental stage-dependent data, however, do not provide a complete picture of the genetic landscape of gregarines, nor can they provide information on their genome structure.

To study the gregarine genome, we focused on the marine eugregarine *Porospora gigantea* (Van Beneden, 1869) Schneider, 1875, which is an intestinal parasite of the lobster *Homarus gammarus.* First described in 1869, E. Van Beneden named the organism *Gregarina gigantea* in reference to the “gigantic” size (up to 16,000 µm) of the trophozoite stages, being visible to the naked eye[14]. Van Beneden reported that “cyst” forms of this parasite accumulated within the chitinous folds of the lobster rectum, the “rectal ampulla”. Schneider went on to show 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]. Biological material for genomic studies is particularly difficult to gather from non-cultivable microorganisms, so we took advantage of the existence of these well-described structures[16–19], knowing that each cyst contains several thousand “gymnospores”, each composed of hundreds of zoites, involving the natural amplification of its genomic material. Cysts indeed proved to be a remarkable natural source of genomic DNA. Gliding is a characteristic apicomplexan movement that also happens to be essential for the invasion and egress of host cells, and thus for the intracellular parasitic lifestyle[20–24]. *P. gigantea* trophozoites are known to glide at rates of up to 60µm/s[25], so are prime candidates in which to study the mechanism of gliding motility. Currently about 40 proteins, identified mainly in *T. gondii* and *Plasmodium falciparum*, compose the glideosome, a commonly accepted structural model of the apicomplexan motor complex (see Frénal et al, 2017[26] for review).

In this study, we report the assembly of the first two draft genomes of *P. gigantea.* We present their main features and predicted proteomes and compare them to other available apicomplexan genomes, revealing an unexpected diversity. We investigated their position within Apicomplexa and among the major subgroups of gregarines through a phylogenomic analysis. We also examined their position within the crustacean gregarines according to 18S ribosomal gene sequences. Finally, a comparative study was performed to gain insight into the conservation of gliding proteins for these gregarines, the currently fastest moving extracellular Apicomplexa.

RESULTS

Phenotypic characterization

Specimens of the lobster *Homarus gammarus*, the type host species for *Porospora gigantea*, were collected either from the sea in Roscoff bay (France) or from commercial lobster tanks in Roscoff (Figure 1, Table S2). A total of 35 lobsters (9 from the wild and 26 from captivity) were dissected and infection with *P. gigantea* was quantified (Figure 1, Figure S1). Overall, infection levels were significantly higher in lobsters freshly caught from the sea (prevalence of 100%, high parasitic loads) than in lobsters that had been held in captivity in lobster tanks (prevalence < 62%, low parasitic loads, see Table S2), a similar result to that reported by Van Beneden (1869)[14]. The morphology of cysts, gymnospores, zoites and trophozoites was imaged and measured (Figure 1, Tables S3, S4 and S5). Cysts were mostly spherical but some were ovoid, with diameters ranging from ~108µm to ~240µm (mean ± standard deviation, 151.1 ± 45.3µm, n = 97), and they enclosed thousands of gymnospores, that were also mostly spherical, with diameters from less than 5µm to almost 7µm (5.63 ± 0.69µm, n = 265). These gymnospores were indeed composed of radially arranged zoites forming a monolayer with an optically void center. Observation of broken gymnospores by scanning electron microscopy made it possible to measure the length of the constituent zoites (1.04 ± 0.16µm, n = 105) and their apical width (0.630 ± 0.129µm, n = 176). Trophozoites were very thin and long, up to 2585µm for a mean width of 41.8 ± 10.4µm (n = 104). As previously described, the posterior of the trophozoite was slightly thinner, ~30µm. The whole trophozoite surface was covered by longitudinal epicytic folds (Figure S1.B) that are thought to be necessary for eugregarine gliding[27]. The sum of these morphological observations all accord with the species being *P. gigantea* from the type host *H. gammarus*[6,14,15].

Gliding of isolated trophozoites was filmed. The dynamic recordings confirm that trophozoites moved uni-directionally, with the protomerite forwards, in either straight or curved lines depending on the individuals observed, with the whole body (deutomerite) following the same path as the apical protomerite (Film S1). The speed of trophozoite displacement was estimated to be ~60µm/sec, as initially observed by King and Sleep (2005)[25], but was faster than 100µm/sec in some recordings (Table S6). No syzygy was observed. Some solitary encysted trophozoites were observed, supporting the observation of Léger and Duboscq (1909)[28], who considered that encysted gymnospores correspond to a schizogonic rather than a gamogonic phase of *Porospora* development, a hypothesis that is still debated[6].

Two highly related genomes

Four biological samples were sequenced and analyzed independently, and then assembled together (Figure 2.A). The raw assembly produced 214,938 contigs (99.6 Mb) among which were 13,656 contigs longer than 1 kb (47.9 Mb). The scaffolds obtained were cleaned by removing contaminants such as bacterial, fungal and host sequences (Figure 2.B), resulting in a raw assembly of 1719 contigs covering 18 Mb.

The analysis of contig coverage for each individual library revealed a bimodal distribution suggesting a mixture of genomes in differing proportions depending on the biological sample (Figure S2). More precisely, while only one set of contigs displayed a significant coverage for the lobster tank parasite sample (JS-470, peak around 250×), the three other parasite samples from freshly captured hosts (JS-482, JS-488, JS-489) showed two distinct sets of scaffolds with similar size (~9Mb) and different coverage values. The difference in coverage was used to split the whole assembled contigs into two sets that were named A for the set of contigs present in all four samples, and B for the set present only in the three lobsters freshly captured in the wild (Figure 2.C). The percentages of genomes A and B in each biological DNA sample was estimated (Figure S2) as 100% A for JS-470, 63.2% A and 36.8% B for JS-482, 70.5% A and 29.5% B for JS-488, and 62.4% A and 37.6% B for JS-489, based on medium coverage levels. Genome A maps to 787 contigs for a total of 8.8 Mb, whereas genome B maps to 933 contigs for a total of 9.0 Mb. Contigs from the two genomes can be aligned with each other over 7.7Mb, with a percentage of divergence around 10.8% at the nucleotide level.

To summarize, the A and B genomes associated to the species named *P. cf. gigantea* are similar in size (~9Mb) and are syntenic but divergent (Figure 2).

Genome features

Two genomes with similar coding capacities. A total of 10,631 putative genes were predicted from the raw assembly (17,930 including alternative splicing), which were split into two sets of similar size: 5270 genes in genome A (8835 transcripts) and 5361 genes (9035 transcripts) in genome B (Table 1, Figure 2). The completeness of both A and B genomes was assessed by using BUSCO software[29] on the Apicomplexa geneset (n = 446). Genomes A and B respectively showed completeness scores of 70% (n = 312) and 67.7% (n = 302) (Figure S3).

The number of A and B orthologues was investigated. The predicted proteins of *P. cf. gigantea* A and B were split into 5656 orthogroups including 4443 groups (88%)

which had at least one orthologous gene for both A and B. This percentage of common orthogroups between genomes A and B is higher than that observed between *Plasmodium falciparum* and *Plasmodium berghei* (70%), thought to have diverged around 33 Mya ago (TimeTree[30]), but similar to that observed between *P. falciparum* and *Plasmodium 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 (*Cryptosporidium parvum*, 18%; *G. niphandrodes*, 17%; *P. falciparum*, 14%; *T. gondii*, 14%) despite the differences in divergence, but it is higher than the percentages observed with chromerid species (*Chromera velia*, 8%; *Vitrella brassicaformis*, 10%). We can deduce from these results that the *P. cf. gigantea* genomes do not share significantly more orthogroups with *G. niphandrodes*, the only other available gregarine genome, than with any other apicomplexan (Figure 3).

Two gene-dense genomes with small introns. The proportion of coding sequences in A and B genomes is 84%, which is particularly high compared to other reference species (with values ranging from 25% to 76%; Table 1). The genomic compaction of non-coding DNA in genomes A and B can be explained by the shortness of most introns (Figure S4). A specific class of introns with lengths around 25-30 bp (mode at 28 bp) represents 71-72% of the introns. The donor and acceptor sites of these small introns have specific consensus patterns (Figure S4) which are different from other *Porospora* introns. Specifically, these introns exhibit a strongly conserved adenine located 6 bp upstream of the 3' acceptor site which could represent the intron branch point, as observed for the small introns (20bp) in *B. microti*[31].

Loss of organellar genomes. Recent studies suggest that organellar genomes are lost in most gregarines[10,32]. A precise protocol was set up to identify putative contigs associated with organellar genomes in *P. gigantea*. All the assembled contigs (assigned to *P. gigantea* or not) were searched for regions similar to known organellar genomes. A sensitive protocol based on TBLASTX identified 108 putative regions that were aligned to the NCBI NR library. 102 regions were discarded as bacterial contamination. The 4 contigs corresponding to the remaining 6 regions with at least one significant hit against an eukaryotic sequence were manually curated. Two contigs were assigned to host-derived contaminants whereas the two other long contigs (L=24892 and L=33594) corresponded to *P. gigantea* nuclear genome. Thus, our analyses did not reveal any putative contigs compatible with mitochondrial or apicoplastic genomes.

Evolutionary histories of *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* orthologues was 0.0959, similar to that calculated by Neafsey et al[33] (0.068 substitutions per site) or Reid et al[34] (0.086-0.11 per site). We assumed that these *Plasmodium* species diverged between 3.3 and 7.7 Mya (TimeTree). The mean dS value observed between the same orthologues in both *P. cf. gigantea* genomes was about 0.4295 substitutions per site. Assuming similar substitution rates in gregarines and *Plasmodium* species, we dated the split between genomes A and B to have occurred between 15.5 Mya and 37.7 Mya. This order of magnitude is similar to the estimation of when the basal splits of the mammal *Plasmodium*[35] (12.8 Mya) or all *Plasmodium*[36] (21.0–29.3 Mya) occurred, but is significantly later than the emergence of Nephropidae (lobster family) around 180 Mya[37,38].

Expanded superfamily of crustacean gregarines. To assess the position of *P. cf. gigantea* A and B within Apicomplexa, we constructed a genome-wide phylogeny based on 312 concatenated proteins from the datasets published by Salomaki et al, 2021[13] and all recently published transcriptomic data from gregarines[10,11,13] (Figure 4). This phylogeny grouped *P. cf. gigantea* A and B into one clade, placed as a sister group of other crustacean gregarines (*Cephaloidophora communis*, *Heliospora caprellae*), although having shorter branch lengths. In agreement to Salomaki et al (2021)[13] *Cryptosporidium* species remain at the base of A+G (Apicomplexa + gregarines), using a LG+C60+G+F model in maximum likelihood phylogenomic analyses. However, the bayesian analysis using classical partitioned model LG+G+F is in favor of a A+C topology (Apicomplexans + *Cryptosporidium*) (average standard deviation of split frequencies = 0.020977). More sampling of *Cryptosporidium* relatives is required to address the apicomplexan topology issue.

The sequences of 18S small subunit ribosomal DNA, for which the largest taxonomic sampling for gregarines is available in databases, was also used to position *P. cf. gigantea* within the crustacean gregarines. Using a combination of amplifications with specific primers (initially based on Simdyanov et al. (2015)[39] and Schrével et al. (2016)[40] then partly redesigned (Figure S5, Table S7)) and *in silico* clustering, 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, only one within the 18S sequence, and 29 within the 28S sequence (Figure S5). Two phylogenetic studies were performed, one excluding environmental sequences (Figure S6), the other including them (Figure S7). Most environmental sequences are derived from marine sediments from a wide range of habitats but only two sequences are from the North Atlantic where European and American lobsters live.

Congruent with the concatenated phylogeny (Figure 4), both 18S phylogenies assigned *P. cf. gigantea* A and B to their own clade, placed as a sister group to all other crustacean gregarines (*Cephaloidophora*, *Heliospora*, *Thiriotia*, and *Ganymedes* species), as established in Rueckert et al (2011)[41] (Figures S6 and S7). Five main clades constituting the superfamily Cephaloidophoroidea were retrieved. The four clades previously outlined[41], redenominated as Ganymedidae, Cephalodophoridae, Thiriotiidae (as proposed by Desportes and Schrével (2013)[6]), and Uradiophoridae, had at their base the clade Porosporidae. Historically defined as the family gathering *Porospora* and *Nematopsis* genera[6], this clade is constituted of the two sequences of *P. cf. gigantea.* A new putative clade was formed by the five sequences from a Slovenian karst spring published by Mulec and Summers Engel (2019)[42] (Figure S7), and it is very well supported to be a sister group to four of the crustacean gregarine families, while the family Porosporidae retains its position as a sister group to all these other clades.

Partially conserved glideosome machinery

We conducted an inventory of the presence or absence of genes encoding proteins involved in the gliding motility based on the molecular description of the so-called glideosome machinery, grouped according to their function as established by Frénal et al (2017)[26] (Figure 5.A, all orthologues for *P. cf. gigantea* are detailed in Table S8). Genes for these *T. gondii* and *P. falciparum* reference proteins were searched for in both *P. cf. gigantea* genomes and in the genomes of a selection of representative species, as well as the recently published gregarine transcriptomes[10,11,13].

Actin and associated factors. Actin in apicomplexans is characterized by a globular monomeric form (G-actin) which polymerizes as needed into short unstable filaments (F-actin)[43] using various regulators such as profilin[44–46], ADF cofilin[47], formin[48–50], cyclase-associated proteins (CAP)[51] and F-actin capping protein Cpβ[52]. The inactivation of actin or its associated regulators compromises motility and host cell invasion and egress, although motility may persist in an altered form for a few days, perhaps through alternative mechanisms[26,53–55]. Overall, these proteins are well conserved among Apicomplexa. However, profilin appears to be absent in insect-infecting Gregarinoridea; CAP and Cpβ also seem to be poorly conserved in gregarine transcriptomes but present in both *P. cf. gigantea*.

Apicomplexan-specific glideosome proteins. The core glideosome machinery mainly comprises specialized proteins found only in apicomplexans. The single-headed short heavy chain myosin class XIV, named myosin A (MyoA), acts as a motor generating the rearward traction required for gliding motility, invasion and egress, as evidenced by various conditional depletion experiments[56–58]. The glideosome itself is situated between the plasma membrane and the apicomplexan-specific inner membrane complex (IMC). In the IMC, MyoA is associated with a light chain, myosin light chain 1 (MLC1) in *T. gondii* or MyoA tail domain-interacting protein (MTIP) in *P. falciparum*[59], as well as several glideosome associated proteins (GAP), GAP40, GAP45, GAP50[60–62], GAP70 and GAP80 as yet only described in *T. gondii*[57]. GAP45 is thought to anchor the glideosome to the plasma membrane by recruiting MyoA as a bridge[62], whereas GAP40 and GAP50 are predicted to help anchor MyoA to the parasite cytoskeleton[63]. 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[26,64]. Finally, the conoid-associated myosin H is necessary for initiating gliding motility in *T. gondii*[65].

Genes encoding myosins A, B, C, D and E and associated light chains were found in all species. Myosin H is also widely conserved in intracellular apicomplexans. However, among the gregarines Myosin H is only present in a few species. For glideosome associated proteins, only GAP40 was found in all species, although the sequences from gregarine transcripts and chromerids were less well conserved. Surprisingly, given the central role attributed to GAP45 in the glideosome model, no ortholog was found in gregarines except for two poorly conserved sequences in *Lankesteria abotti*, *Lecudina tuzetae*, *Cryptosporidium* and chromerids. However, we identified a short conserved 3' domain (<50aa) in *L. tuzetae*, *Pterospora schizosoma* and *Siedleckia nematoides*. A similar domain is found in *P. cf. gigantea* A and B. It is however not sufficient to conclude whether it is an orthologous protein. GAP50 seems to be more conserved among apicomplexans, but is absent or only partially conserved in most of the gregarines. As expected, GAP70 and GAP80, only identified so far in *T. gondii*, were not found in other species, except for an orthologue of GAP80 in the coccidia *Hammondia hammondi*. Concerning GAPMs, we found orthologues of at least one of its variants (GAPM 1, 2 or 3) in most species. However, GAPMs seem to be totally absent in at least 7 species of gregarines (*Ancora sagittata, Protomagalhaensia* sp. *Gyna, Protomagalhaensia wolfi, Gregarina* sp. *Pseudo, Pterospora schizosoma, Selenidium pygospionis, Siedleckia nematoides*). Finally, GAC is overall well conserved in apicomplexans but absent from chromerids, supporting its apicomplexan-specific status. However, we were not able to identify GAC in several gregarine transcriptomes (*P.* sp. *Gyna, P. wolfi, G.* sp. *Pseudo, H. capreallae, L. abotti, L. tuzetae, P. schizosoma*) (Figure 4).

Adhesins and TRAP-like candidates. The glideosome machinery, anchored in the parasite cytoskeleton, needs to interact with extracellular receptors of the host cell to propel the parasite forward over the host surface. This is made possible by the presence of extracellular adhesins secreted by the micronemes[66,67] and connected to the glideosome through the glideosome associated connector (GAC) protein[68]. Thrombospondin adhesive protein (TRAP)[69] is a *Plasmodium* adhesin required for gliding, whose homologue in *T. gondii* is MIC2[70]. At the end of the gliding process, rhomboid protease 4 (ROM4) cleaves the adhesins, disengaging them from receptors and, for intracellular parasites, allowing them to enter the host cell[71–73]. TRAP-like proteins, while highly divergent from one species to another, constitute a family of functionally homologous proteins sharing adhesive domain types, involved in parasite motility and cell penetration[74–76]. TRAP-like or TRAP-related proteins have been detected in various stages of *Plasmodium* (CTRP[77], MTRAP[78], TLP[79]) and have also been found *in silico* in *Cryptosporidium* (TRAPCs, CpTSPs[76,80,81]) as well as in several *Babesia* and *Theileria* species[82–85], in *Neospora caninum*[86] and in *Eimeria*[87,88]. We first looked for the TRAP proteins which have been implicated in gliding through 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 all have an extracellular region containing one or more TSP1 domains and/or one or more vWA domains[74–76]. They are also characterized by the presence of a single transmembrane domain, a signal peptide, and, in some cases, a juxtaposed rhomboid protease cleavage site, and a short, charged C-terminal cytoplasmic domain with aromatic residues. The presence of a YXXΦ tyrosine sorting signature has also been described[75] (where X signifies any amino acid, and Φ any hydrophobic amino acid).

To evaluate the presence of TRAP-like proteins in *P. cf. gigantea* genomes, we inventoried all predicted proteins containing at least one TSP1 domain (Table S8), then identified potential candidates with several TRAP-like structural characteristics (Figure 6). We identified a CpTSP2 orthologue within both *P. cf. gigantea* genomes, designated PgTSP2. Like CpTSP2, it is a large protein (~2800 aa) composed of Notch, TSP1, and Sushi domains. PgTSP2 has a localization signal, a transmembrane domain and a short, charged, basic cytoplasmic tail. This protein also has orthologues in *G. niphandrodes,* in chromerids and coccidia.

We demonstrated the presence of genes encoding four other related protein pairs 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, charged, acidic 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, charged, acidic cytoplasmic tail, plus several extracellular EGF or EGF-like domains, as also described in *C. parvum* CpTSP7, CpTSP8 and CpTSP9[80]. We identified another protein, PgTSP_EGF-2, very similar in structure.

Moving-junction associated proteins. In apicomplexans with intracellular stages such as *T. gondii*, invasion occurs when the tachyzoite initiates a pivotal movement known as reorientation, and a mobile junction settles into the host cell membrane allowing the parasite to enter. Gliding forces are also involved in this process[89], to which the host cell contributes[90]. A micronemal protein, AMA1, combines with rhoptry neck proteins (RON2, RON4, RON5 and RON8) to firmly secure the parasite to the host cell. In *P. falciparum,* another AMA-like protein, merozoite apical erythrocyte-binding ligand (MAEBL) has an important role in invasion alongside AMA1[91].

Gregarines remain extracellular during their entire life cycle and *Cryptosporodium* display an intracellular but extra-cytoplasmic stage, so it was not surprising that we did not identify any orthologue of the moving-junction proteins of either these groups. We also searched for predicted proteins implicated in adherence and invasion in *Cryptosporidium*, such as GP15/40, GP900 and mucins, but found no equivalent in either *P. cf. gigantea*[92,93].

Regulatory factors and signaling pathways. Increases in parasite intracellular calcium activate calcium-dependent protein kinases (CDPK) that regulate motility, microneme secretion, invasion and egress[94,95]. Other proteins acting in such signaling pathways include diacylglycerol kinase 1 (DGK1) and acylated pleckstrin homology domain-containing protein (APH), which are involved in microneme secretion regulation[96,97]; the C2 domain-containing protein DOC2.1 which mediates apical microneme exocytosis[98]; and the apical lysine methyltransferase (AKMT), which is involved in gliding motility, invasion and egress in *T. gondii*[99]. We were unable to identify APH in most gregarines and chromerids, and DOC2.1 could not be identified in several transcriptomes. All other regulatory factors appeared to be largely conserved.

DISCUSSION

Molecular data support the presence of two species

We report here clear lobster coinfection by two gregarines believed to be distinct that we have named *Porospora cf. gigantea A* and *Porospora cf. gigantea B*. At the molecular level, these two organisms have very similar genomes in terms of size, protein coding capacity, GC content and overall organization with 86% synteny conservation. The delineation of species now requires more precise integrative morpho-molecular approaches, combining extensive imaging (SEM, TEM) and single cell –omics, to find specific traits. 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 a single nucleotide, a non-significant divergence for discriminating species.

However, at the genomic level, the genomes show a nucleotide divergence of more than 10% which is incompatible with subspecies or strain definitions. By comparison, applying the same protocol to *P. falciparum* and *P. reichenowi* genomes concluded that the divergence between the two *Plasmodium* species is only 3.2%. Similarly, a divergence of 3-5% between the genomes of *C. parvum* and *Cryptosporidium hominis* has been reported[100]. The large overall genomic divergence between *P. cf. gigantea A* and *P. cf. gigantea B* indicates that they are probably not interfertile, and thus should be considered as different species.

Pending a more integrated morpho-molecular definition of their taxonomy, and better documentation of Cephaloidophoroidea species in general (Figure 4), we propose that *P. cf. gigantea A* and *P. cf. gigantea B* are two distinct organisms infecting *H. gammarus*.

Two species with compact genomes and a highly specific gene set in common

These two genomes are the first marine gregarine genomes to be sequenced and analyzed and the information generated considerably expands our knowledge of apicomplexan diversity. Both A and B genomes are very small compared to other apicomplexans, with a particularly high gene density. For example, for a similar genome size, *Cryptosporidium* spp. have only about 3900 protein-coding genes compared to the 5300 genes of *P. cf. gigantea*. This result could be partially explained by the absence of certain non-coding sequences in the assemblies, such as centromeres, telomeres and repeated sequences which are difficult to sequence and assemble, notably in *de novo* assembled genomes. However, the compaction is partially due to the comparatively short introns. Small introns with similar consensus sequences have been described in *Babesia microti*[31].

So far, we have not found any evidence of organellar genomes, whether from the mitochondrion or apicoplast. This needs to be investigated more definitively, especially the mitochondrial aspects. Indeed, the cystic stages from which DNA was collected are unlikely to have many mitochondrial genome copies. To address this issue, it would be more suitable to investigate trophozoite stages via single-cell genomics, for example. According to a recent study, mitochondrial genomes seem to have disappeared from eugregarines[32]. Instead of a distinct mitochondrial genome, the 129 mitochondrial proteins differentially conserved among the gregarine lineages are encoded in the nuclear genome. It would be interesting to identify how many of these nuclear-encoded proteins are conserved within the *P. cf. gigantea* genomes and to reconstruct their specific metabolism. Regarding the apicoplast genome, a recent study stated that it has probably been lost in all eugregarines, while archigregarines may have conserved a highly reduced plastid genome[10].

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 lower than was found for the *G. niphandrodes* genome (83%) and the 24 other representative species we evaluated (from 76.9% for *Cystoisospora suis* to 100% for *P. falciparum* (Figure S3)). This result also illustrates that the definition of "Apicomplexa core genome" is probably currently highly biased, notably towards *Plasmodium*. Gregarines should be taken into more consideration, as their divergence compared to other apicomplexan models was confirmed by the orthogroup analysis indicating a low percentage of genes conserved between A or B and other studied apicomplexans (<18%).

Even among gregarines the wide diversity is noted 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, and highlight the astonishing protein diversity brought about by 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 adaptations to parasitism in apicomplexans, which have been able to establish themselves in most Metazoan lineages, vertebrate or invertebrate, marine or terrestrial, in one or more hosts, intracellular or extracellular modes.

The gregarine glideosome(s)

An incomplete but operational machinery. Gliding motility is an essential feature of apicomplexans, and for some intracellular parasites among them, glideosome proteins have been shown to be crucial for host cell invasion and egress[22,23,26,63,74]. However, our sequence analysis of the glideosome components shows that the currently known mechanistic model based on *T. gondii* and *P. falciparum* does not fully account for gliding in all apicomplexans, as anticipated[26,63,67]. Moreover, the conservation of the proteins involved is very variable among the gregarines for which we have omic data. There is little evidence of key molecular components such as canonical adhesins or GAP45, implying that gregarines and *Cryptosporidium* species may have an at least partially alternative machinery dedicated to gliding (Figure 5.B), especially since *P. cf. gigantea* trophozoites are able to glide so rapidly.

The model 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[101]. 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* and in other apicomplexans, there are proteins with a structure similar to TRAPs (TRAP-like), that might replace the canonical TRAPs. Understanding the evolution of TRAP requires 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[76]. Similarly, the vWA domains, which are found in the canonical TRAPs, appear to be absent from the *Cryptosporidium* genomes. Since gliding is observed in *Cryptosporidium* 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 proteins 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*[75,80]. This is a plausible idea since ROM4, which cleaves adhesins from extracellular receptors of the host cell at the end of the gliding process, is extremely well conserved. GAP45 is thought to maintain the interaction between the IMC and the plasma membrane, and acts as an essential bridge between the two structures[102]. Deleting GAP45 has been proved to prevent glideosome assembly in *P. falciparum*[103]. Perhaps the absence of GAP45 in gregarines and *Cryptosporidium* could be compensated by other GAP-like proteins or it may not even be necessary. Indeed, a looser motor architecture has been proposed, in which actin-myosin motors push in a general backward direction, without necessarily being guided by GAP proteins[63]. 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[104], questioning again the real role of GAP45 and suggesting our comprehension of the intricacies of the glideosome is still incomplete.

Different structures for other forms of gregarine motility? Gregarines have other means of motility, presumably governed by other molecular mechanisms, and the relevance of the glideosome concept to gregarines has been questioned[27,105]. In particular, archigregarines use several modes of movement such as rolling and bending, but not gliding[6,19]. Coelomic and intestinal eugregarines, like crustacean gregarines, have longitudinal, drapery-like surface structures called epicytic folds, the most distinctive feature that differentiates eugregarine trophozoites and gamonts from other apicomplexans. These structures are considered to be involved in eugregarine 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 *Gregarina polymorpha*[106,107]*.* Epicytic folds and mucus, the substance often observed in the trace left by gliding eugregarines[6,27], are key components to integrate into an alternative model to the current glideosome more representative of eugregarine motility. A particularly interesting study of the crustacean gregarine *Cephaloidophora cf. communis* reported specific attachment apparatus structures[108]. 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]. This organism also has also a septum, a tubulin-rich filamentous structure that separates the epimerite from the protomerite at the cell apex. Together with microneme-like structures, these features suggest adhesion proteins are produced which could be threaded through the membrane by the numerous pores visible on the epimerite[108]. We were unable to identify alternative movements to gliding in *P. cf. gigantea* (such as peristaltic movement described in other coelomic eugregarines[6,109]). Additional observations are needed to fully document the range of potential motilities in this species, especially since the crustacean-infecting gregarine *C. cf. communis* is capable of jumping or jerking during discontinuous gliding[108]. The different structures invoked, or their absence must be evidenced; indeed, in eugregarines, subpellicular microtubules have never been observed, even though they are supposed to be involved in gliding motility in other apicomplexans[27,108].

Whatever the molecular mechanisms leading to gliding motility in *P. cf. gigantea*, there are likely to be unique molecular structures, which have evolved consecutive to the specific evolutionary path of gregarines, and which differ from what is currently documented in other apicomplexan lineages.

ACKNOWLEDGEMENTS

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

FUNDING

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), and 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 was funded in parts by the call EMBRC-France 2016 (Investments of the Future program, reference ANR-10-INSB-02, Agence Nationale de la Recherche).

AUTHORS' CONTRIBUTIONS

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 did the bioinformatics analyses. AL, IF and LD sequenced and assembled the complete ribosomal loci. JB, IF and EDB performed the glideosome expert annotation. JB performed the 18S phylogenetic analyses. JB, EDB and LP performed the phylogenomic analyses.

JB, EDB, LP and IF wrote the manuscript with contributions from all authors. All authors have read and approved the manuscript.

COMPETING 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 bar = 100µm). **B.** Zoom on A, showing trophozoite epimerite (scale bar = 10µm). **C**. Rectal ampulla showing cysts in folds (Lobster #4) (scale bar = 1 mm). **D.** Isolated cyst (Cyst #4, Lobster #12) (scale bar = 50µm). **E.** Broken cyst packed with gymnospores (Lobster #4) (scale = 10 µm). **F**. Section across a cyst illustrating radial arrangement of zoites in gymnospores (JS449 = Lobster #35) (scale bar = 2µm). **G., H.** Zoom on intact and broken gymnospores showing zoites (Lobster #4) (scale = 1µm). All images are scanning electronic micrographs except F which is a transmission electronic micrograph. See also Figure S1, Tables S2, S3, S4, S5 and S6.

Figure 2. Protocol for assembling 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 and S4.

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. Phylogeny of Apicomplexa. Maximum likelihood phylogeny of apicomplexans as retrieved from a 312 proteins dataset, merged from two previously published datasets[10,11,13]. Final concatenated alignment comprised 93,936 sites from 80 species. Bootstrap support values (n = 1000) followed by MrBayes posterior probabilities are shown on the branches. Black spots indicate 100/1 supports. *Porospora cf. gigantea* A and B sequenced in this study are bolded. See also Figures S6 and S7.

Figure 5. Comparative analysis of glideosome components. A. Table of presence/absence of genes encoding glideosome proteins, distributed into functional groups. Glideosome components have been described mainly in *T. gondii* and *P. falciparum*. Protein sequences were searched for in the genomes of both *Porospora* and a selection of representative species as well as in available gregarine transcriptomes. Green indicates the presence, while white indicates the absence of an orthologous protein-encoding sequence. Light red refers to cases where only partial sequences have been retrieved. Violet indicates the presence of at least one protein in multigenic family proteins. * refers to the GAP45 3' short

conserved domain found in some gregarines species. 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 with dotted lines.

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 S2.

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 and genome assemblies are available in the NCBI database (Bioproject PRJNA734792). Detailled protocols as well as complementary data (phylogenomics datasets, alignments, phylogenetic trees, blasts results and orthogroups) are available on Github at https://github.com/julieboisard/Marine_gregarines_genomes.git/.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Several specimens (n = 35) of the lobster species *Homarus gammarus* were collected in the English Channel at Roscoff (Brittany, France) between July 2015 and October 2017 (Table S2), either directly from the wild (Roscoff Bay) or from lobster tank facilities, in which crustaceans are maintained in captivity several weeks to months before their commercialization. According to UICN Red list, *Homarus gammarus* is not an endangered species[110]. The intestinal tract was carefully dissected from each freshly killed host specimen, and transferred to large Petri dishes filled with 0.22-µm filtered, 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). Trophozoites freely moving in the upper intestine lumen, and cysts loosely attached within the chitinous folds of the hosts' rectal ampullae (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 methods, trophozoites, cysts and host tissues were carefully washed several times in 0.22-µm filtered, 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 omics studies (i.e. DNA and RNA sequencing).

METHOD DETAILS

Electronic microscopy

For the scanning electron microscopy (SEM) studies, isolated trophozoites and cysts, or host intestines and rectal ampullas opened along their longitudinal axis, were washed as indicated above then fixed in 2.5% (v/v) glutaraldehyde in 0.1 M sodium cacodylate (pH 7.2) at 4°C for 6 to 12 hours. After two washing steps in 0.1 M sodium cacodylate (pH 7.2), biological specimens were transferred to microporous specimen capsules (30 µm porosity, 12 mm diameter, 11 mm high, ref #70187-20, Electron Microscopy Science) and dehydrated in a graded series of ethanol in double-distilled water (50, 70, 90, and 100%). Biological specimens in the capsules were critical point-dried in liquid CO_2 (Emitech K850, Quorum Technologies), then transferred to adhesive carbon-coated holders, and coated with 20 nm of gold (JEOL Fine Coater JFC-1200). Specimens were then examined with a Hitachi SU3500 Premium scanning electron microscope.
For the transmission electron microscopy (TEM) studies, samples were fixed for 2 h in 0.2 M sodium cacodylate buffer with 4% glutaraldehyde, 0.25 M 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 (0.25 M, 0.12 M, 0 M) for 15 min each time, followed by post-fixation for 1 h at 4 °C in 2% osmium tetroxide in 0.1 M sodium cacodylate buffer. After three rinses in 0.2 M sodium cacodylate buffer, samples were dehydrated by successive transfer through an increasing 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 3 min, were examined on grids with a Jeol 1400 transmission electron microscope (Jeol, Tokyo, Japan).

DNA/RNA isolations

Genomic DNA (gDNA) was isolated from 4 biological samples of pooled cysts taken from 3 specimens of the host *H. gammarus*: sample JS-470 from Lobster #7 (~70 cysts), sample JS-482 from Lobster #11 (~50 cysts), samples JS-488 and JS-489 from Lobster #12 (~100 cysts each). Lobster #7 was provided by the Roscoff lobster tank facility while Lobster #11 and Lobster #12 were caught from Roscoff bay. DNA was extracted from the pooled cysts using Macherey Nagel Tissue and Cells isolation kit (ref 740952.50) with yields of 4.1 µg for JS-470, 2 µg for JS-482), 4.5 µg for JS-488 and

6.7 µg (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 lysis step at 56°C was extended beyond the recommended to 1-3 hours with frequent microscopic (binocular) inspection to monitor cyst digestion until completion.

RNA was also isolated from 2 additional biological samples, both composed of pooled cysts taken from the rectal ampulla of their respective hosts: 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 ~155 ng of total RNA in 55 µl 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 ~50 ng of total RNA in 55 µl as assessed by Qbit quantification.

DNA/RNA sequencing and assembly

The gDNA extracted from the 4 biological samples (JS-470, JS-482, JS-488 and JS-489) was sequenced individually using Illumina NextSeq technology (2 × 151 bp; 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[111] (version 0.11.5). Reads were cleaned with Trim Galore[112] (version 0.4.4) which removed remnant Nextera adaptors, clipped 15 bp at 5'-ends and 1 bp at 3'-ends and trimmed low-quality ends (phred score < 30). The assembly was carried out using SPAdes[113] (version 3.9.1; options: careful mode, automatic k-mers) with the pooled libraries (Figure 2.A).

RNA was extracted from both samples (JS-555 and JS-575c) and treated with RNAse-free DNase. Libraries (Institut du Cerveau et de la Moelle, CHU Pitié Salpétrière, Paris) were prepared following the kit manufacturer's recommendations (SMART-Seq v4 Ultra Low Input RNA Kit from Takara). Samples were sequenced on a NextSeq 500 Illumina device with MidOutPut cartridge to generate a total of 2 × 87 M reads of 75 bp. The read quality was checked by using FastQC and cleaned by using Trim Galore to remove remnant Nextera adaptors, clipping 15 bp at 5'-ends and 1 bp at 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[114,115].

All genomic contigs longer than 1 kb were analyzed by principal component analysis (PCA) based on their 5-mer composition, which classified them into 6 groups 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 using Augustus[116] (version 3.3) and the Apicomplexa gene model for *T. gondii*. All the predicted proteins were thus compared with the NCBI non-redundant protein database using BLAST[117]. The analysis of the taxonomic groups corresponding to the best hits, enabled us to identify five clusters as putative bacterial contaminants whereas the sixth cluster which included 1745 contigs (18.0 Mb), was identified as organisms closely related to Apicomplexa, referred to as the “apicomplexa” cluster (Figure 2.B).

Identification of genomes A and B

Preliminary analysis of the “apicomplexa” cluster exhibit two sets of contigs with approximatively 10% of divergence and specific coverage values in the four libraries. The contigs of the “apicomplexa” cluster were split into genomes A and B by using the difference in coverage observed for the four gDNA libraries (Figure 2.C). Each gDNA library (JS-470, JS-482, JS-488 and JS-489) was individually mapped to the contigs using Bowtie2[118] and the median coverage was calculated for each contig and each library using Samtools[119] and Bedtools[120] suites. This coverage information was processed by PCA and a k-means algorithm which classified the contigs into 2 clusters. Then, a linear discriminant model was trained with the coverage information and the result of this first classification before applying it to all the contigs in order to improve the classification. The linear discriminant method (training and classification) was iterated 3 times until convergence. A similar analysis was carried out with 1-kb non-overlapping windows (instead of full-length contigs) to identify putative hybrid contigs. Contigs were thus classified to different genomes depending on the windows, then divided into sub-contigs which were re-assigned to their respective genomes. A detailed protocol with R scripts is available on github (see data and code availability).

The nucleotidic divergence between genome A and genome B was estimated from the alignment of contigs built with Mummer3.0[121]. All alignments of the syntenic regions were parsed to compute the divergence using a home-made script. Assembly metrics were assessed by using QUAST[122] (version 5.0).

Prediction and annotation

All *de novo* assembled transcripts were aligned against the “apicomplexa” cluster contigs with GMAP[123] within the PASA program[124]. Then, two *ab initio* gene prediction tools, SNAP[125] (version 2017-11-15) and Augustus were trained using a subset of the PASA transcriptome assemblies. A specific gene model was trained with Augustus, including meta-parameter optimization and prediction of introns (allowing small intron length >10bp) using our “apicomplexan” cluster repeat-masked genome assembly as reference (RepeatMasker[126], version 4.0.8). Gene predictions were then performed allowing for the prediction of alternative transcripts and noncanonical intron splice sites. An alternative model was also trained with SNAP (default protocol) and used for gene predictions. The Augustus and SNAP outputs showed that some gene predictions were slightly different, so the predictions were parsed with a home-made script to keep as many alternative genes and transcripts as possible for each prediction made. The completeness of the gene prediction was assessed using BUSCO (version 4.0.6).

The predicted proteins were 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[127] (version 2019-05-11) and iii) the signature domains obtained with Interproscan[128] (version 5.39-77.0).

The ortholog groups were identified with orthoMCL[129] (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. Then the coding sequences of the orthologous groups/quartets including a single gene each for genome A, genome B, *P. falciparum* and *P. reichenowi* were aligned using MacSE[130]. 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)[131] implemented in PAML4[132].

The Infernal software[133] (version 1.3.3) and the Rfam database[134] (version 14.2) were used together to search for transfer RNAs, spliceosomal RNAs and ribosomal RNAs. The snoReport software[135] (version 2) was used to search C/D and H/ACA small nucleolar RNAs.

Removal of contaminant sequences

Host contaminants. All “apicomplexa” cluster contigs were screened against the short reads available from the *Homarus americanus* (PRJNA486050) genome sequencing project, to identify closely-related host contaminants. This dataset was assumed to be free of sequences from apicomplexans, since it was obtained from DNA extracted from non-intestinal tissues (tail, leg or pleiopod appendices). Mapping was carried out with Bowtie2 and the coverages were calculated by using Samtools. The contigs thus identified that were covered over more than 60% of their length by *Homarus* short reads, were considered as host contaminants and were removed.

Prokaryotic and fungal contaminants. In parallel, predicted genes in the “apicomplexa” cluster contigs were deeply analyzed for the presence of bacterial and fungal sequences. For each scaffold containing at least one predicted protein, a BLASTP against the NCBI NR database was launched. If the resulting hit had an e-value lower than 1e-30 and more than 30% of the length of the contig was covered by prokaryote/fungi hits, an additional BLASTN against NCBI NR/NT was performed. For the remaining scaffolds without predicted proteins, a direct BLASTN vs NR/NT search was performed. At the end of this procedure, the contigs with prokaryotes/fungi hits covering more than 70% of the length were labeled as contaminants and were removed from the genome assembly.

Search for organellar genomes

Organellar genomes were searched using the mitochondrial or apicoplastic genomes available in VEupathDB (version 2019-20-01) as well as with the contigs described in Janouškovec et al. (2019)[10] as reference sequences. Firstly, a similarity search using a TBLASTX and these sequences as query was applied on all assembled contigs (identified as *P. gigantea* or not). All hits with a bit score above 100 were considered as organellar candidates and were extracted (with 100bp upstream and downstream). Secondly, these candidates were used in a reciprocal TBLASTX search against NCBI NR database to eliminate bacterial contamination. The regions exhibiting at least one hit against an eukaryotic sequence among the nine best hits were manually studied to check if the associated contigs could correspond to organellar genomes.

Experimental reconstruction of 18S/28S loci

First, a partial SSU rDNA locus was amplified by using JS-470 gDNA (i.e. genome A only) as template and WL1 and EukP3 primers (Table S7) in a conventional PCR reaction. The amplified bands were cloned and sequenced as previously described[40]. The resulting partial SSU rDNA sequence was further extended in the 3' direction still using JS-470 gDNA as template and novel primers designed or re-designed based on the molecular data published for *Cephaloidophora cf. communis* and *Heliospora cf. longissima*[39] (Figure S5A). 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. From this a complete ribosomal locus [18S – ITS1 – 5.8S – ITS2 –28S] of 5977 bp for genome A was reconstructed, which was validated by PCR amplification, cloning and sequencing (Figure S5B). In a similar clustering approach using all reads for JS-482, JS-488 and JS-489, the complete ribosomal locus for genome B was reconstructed *in silico*, which is the same length but has 30 polymorphisms compared to the genome A locus (Figure S5C). Next, 50% of the complete ribosomal locus for genome B was confirmed by PCR amplification, cloning and sequencing (positions 1187 to 4220, covering partial 18S-ITS1-5,8S-ITS2-partial 28S). This second round of clustering was also used to quantify the respective distributions of genomes A and B present in the latter three biological samples at the full ribosomal locus level. The validated sequence of 18S/28S was manually added to the genome assemblies of genomes A and B, respectively. Schematic representation of rRNA loci was done using BioRender (biorender.com).

Phylogeny

Phylogenomics of gregarines. The phylogenomic tree was built from a super matrix of 312 orthologues from two datasets published by Salomaki et al (2021)[13]. These two datasets are composed by 246 and 299 orthologues respectively. For all orthologues, corresponding genes have been searched in the proteomes of *P. cf. gigantea* A and B by using BLASTP and candidates were aligned with known orthologues using mafft[137]. Then, orthologous relationships were validated by visual inspection of all the single-protein phylogenetic tree using RaxML[138] with rapid bootstraps (-f a), -m PROTGAMMAAUTO. Orthologues for *P. cf. gigantea* A and/or B have been recovered for 201 and 256 orthologues in both initial datasets. Both datasets were grouped into a larger dataset composed by 312 non-redundant orthologues. All orthologues were I) filtered with Prequal[139] to remove non-homologous residues, ii) aligned with mafft, iii) filtered with divvier[140] to remove alignment errors, iv) trimed with trimAl[141] and v) merged into the super matrix by using the script *matrix_constructor.py* available with PhyloFisher[142]. The maximum likelihood tree was built with IQ-Tree2 under LG+C60+G+F[143]. The reliability of the phylogenetic tree was tested by the SH-aLRT and ultrafast bootstrap methods (repeated 1 000 times). Bayesian phylogenetic tree was constructed with MrBayes[144] (version 3.2.3) using a LG+G+F model on a partitioned alignment: prset applyto=(all) aamodelpr=fixed(lg); prset applyto=(all) statefreqpr=fixed(empirical); lset applyto=(all) rates=gamma; unlink shape=(all) pinvar=(all) statefreq=(all); mcmc ngen=500000 samplefreq=1000 printfreq=10000 nchains=4 nruns=2 savebrlens=yes; sump burnin=25000; sumt burnin=25000 contype=allcompat. All trees were visualized and edited using FigTree[145] (version 1.4.4) and Inkscape (www.inkscape.org).

18S phylogeny of gregarines. The 100-sequence phylogeny was built from the 18S SSU rDNA sequences of the two genotypes of *P. cf. gigantea*, which were aligned with 84 sequences from a diversity of gregarines, either marine or terrestrial, as well as 12 other apicomplexan sequences. Two chromerid sequences were used as the outgroup[146] but several trees including more than 20 sequences selected from a large diversity of outgroups (from Cryptosporidians, Coccidians, Hematozoans, Colpodellids, Chromerids, Perkinsids, Dinoflagellates, Ciliates, Colponemids, Heterokonts and/or Rhizaria) were built based on Schrével et al (2016)[40] and conducted to the same conclusions. A total of 1614 sites were found to be conserved after selecting conserved blocks as defined by Gblocks[147] (version 0.91b) with the following parameters: minimum number of sequences for a conserved position, 51; minimum number of sequences for a flanking position, 51; maximum number of contiguous non-conserved 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[148]. Maximum likelihood analyses were performed using RAxML (version 8.2.12) with bootstraps estimated from 1,000 replicates. A Bayesian phylogenetic tree was constructed with MrBayes (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.

Environmental 18S phylogeny focused on crustacean gregarines. The 189-sequence phylogeny was built from the 18S SSU rDNA sequences from genomes A and B aligned with 14 from crustacean gregarines, and 154 environmental sequences from several projects described in Rueckert et al. (2011)[41] or gathered from NCBI Genbank. The sequences from the Gregarinoidae clade (n = 19) were used as the outgroup, as this clade has been placed as a sister group to the crustacean gregarine clade in recent literature[10–12]. A total of 1135 sites were found to be conserved after selecting conserved blocks as defined by Gblocks with the following parameters: 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 18S phylogeny.

Expert annotation for glideosome proteins

A reference apicomplexan glideosome protein dataset was written based on glideosome protein repertoires described in the literature mainly for *T. gondii* and *P. falciparum*[26,63,67]. This reference dataset was used as a seed for parsing the orthogroups established for 25 reference proteomes (Table S1) and the predicted proteomes of the two *P. cf. gigantea* genomes. These reference proteomes were selected by considering the most recent data and associated publications to have the most complete panorama of apicomplexan proteins and key functions/structures documented to date. We also searched for potential orthologues within all recently published proteomes of gregarines using BLASTP (seed: reference proteins in *T. gondii* and *P. falciparum*).

For each orthogroup containing at least one of the reference proteins, the list of proteins was extracted, and the protein sequences were recovered with their respective coding sequences for both *P. cf. gigantea* genomes. BLASTP was performed for extracted proteins against the proteomes of *P. cf. gigantea*, as well as for the candidate proteins from each *P. cf. gigantea* genome against the 25 species reference proteomes. 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. Maximum likelihood molecular phylogeny was deduced from each alignment using RAxML. 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[149]. For each protein, the results of all the analyses were examined to validate the candidate proteins within the proteomes of the two *P. cf. gigantea* genomes. A table summarizing the presence or absence of glideosome proteins was visualized using R using the tidyverse package[150]. Putative TRAP-like proteins were identified by searching for sequences encoding the TSP1 molecular domain (IPR000884) within the two *P. cf. gigantea* genomes. The predicted structure of each candidate protein was studied, and if necessary partially predicted proteins were re-edited with Genewise[151]. Schematic representation of TRAP-like proteins was done using BioRender (biorender.com).

REFERENCES

1. 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.

2. 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.

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. Boisard, J., and Florent, I. (2020). Why the –omic future of Apicomplexa should include gregarines. Biol. Cell *112*, 173–185. 10.1111/boc.202000006.

5. 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.

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. Russell, D.G. (1983). Host cell invasion by Apicomplexa: an expression of the parasite's contractile system? Parasitology *87*, 199–209. 10.1017/S0031182000052562.

21. King, C.A. (1988). Cell motility of sporozoan protozoa. Parasitology Today *4*, 315–319. 10.1016/0169-4758(88)90113-5.

22. 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.

23. 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.

24. 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.

25. 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.

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. Mathur, V., Wakeman, K.C., and Keeling, P.J. (2021). Parallel functional reduction in the mitochondria of apicomplexan parasites. Current Biology, S0960982221005418. 10.1016/j.cub.2021.04.028.

33. 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.

34. 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.

35. Ricklefs, R.E., and Outlaw, D.C. (2010). A Molecular Clock for Malaria Parasites. Science *329*, 226–229. 10.1126/science.1188954.

36. 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.

37. 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.

38. 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.

39. 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.

40. 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.

41. 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.

42. 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.

43. 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.

44. 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.

45. 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.

46. 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.

47. 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.

48. 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 *Toxoplasma gondii*. Elife, 12;8:e42669. doi: 10.7554/eLife.42669.

49. 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.

50. 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.

51. Hunt, A., Russell, M.R.G., Wagener, J., Kent, R., Carmeille, R., Peddie, C.J., Collinson, L., Heaslip, A., Ward, G.E., and Treeck, M. (2019). Differential requirements for cyclase-associated protein (CAP) in actin-dependent processes of Toxoplasma gondii. eLife *8*, e50598. 10.7554/eLife.50598.

52. Ganter, M., Schüler, H., and Matuschewski, K. (2009). Vital role for the *Plasmodium* actin capping protein (CP) beta-subunit in motility of malaria sporozoites. Molecular Microbiology *74*, 1356–1367. 10.1111/j.1365-2958.2009.06828.x.

53. 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.

54. 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.

55. 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.

56. 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.

57. 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.

58. 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.

59. 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.

60. 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.

61. 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.

62. 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.

63. 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.

64. 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.

65. 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.

66. 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.

67. 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.

68. 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.

69. 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.

70. 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.

71. 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.

72. 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.

73. 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.

74. 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.

75. 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.

76. 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.

77. 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.

78. 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.

79. 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.

80. 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.

81. 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.

82. 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.

83. 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.

84. 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.

85. 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.

86. 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.

87. 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.

88. 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.

89. 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.

90. 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.

91. 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.

92. 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.

93. 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.

94. 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.

95. 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.

96. 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.

97. 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.

98. 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.

99. 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.

100. 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.

101. 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.

102. 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.

103. Schlott, A.C., Knuepfer, E., Green, J.L., Hobson, P., Borg, A.J., Morales-Sanfrutos, J., Perrin, A.J., Maclachlan, C., Collinson, L.M., Snijders, A.P., et al. (2021). Inhibition of protein N-myristoylation blocks Plasmodium falciparum intraerythrocytic development, egress and invasion. PLoS Biol *19*, e3001408. 10.1371/journal.pbio.3001408.

104. Rompikuntal, P.K., Kent, R.S., Foe, I.T., Deng, B., Bogyo, M., and Ward, G.E. (2021). Blocking palmitoylation of *Toxoplasma gondii* myosin light chain 1 disrupts glideosome composition but has little impact on parasite motility. mSphere *6*. doi/10.1128/mSphere.00823-20.

105. 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.

106. Heintzelman, M.B. (2004). Actin and myosin in *Gregarina polymorpha*. Cell Motility and the Cytoskeleton *58*, 83–95. 10.1002/cm.10178.

107. 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.

108. 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.

109. 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.

110. 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.

111. Andrews, Simon (2010). FastQC: a quality control tool for high throughput sequence data.

112. Krueger, F. (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. https://github.com/FelixKrueger/TrimGalore.

113. 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.

114. 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.

115. 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.

116. 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.

117. 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.

118. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357–359. 10.1038/nmeth.1923.

119. 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.

120. 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.

121. 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.

122. 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.

123. 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.

124. Haas, B.J. (2003). Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. Nucleic Acids Research *31*, 5654–5666. 10.1093/nar/gkg770.

125. Korf, I. (2004). Gene finding in novel genomes. BMC Bioinformatics *5*, 59. 10.1186/1471-2105-5-59.

126. Smit, Hubley, and Green (2015). RepeatMasker Open-4.0. http://www.repeatmasker.org. http://www.repeatmasker.org.

127. 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.

128. 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.

129. 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.

130. 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.

131. 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.

132. Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. Molecular Biology and Evolution *24*, 1586–1591. 10.1093/molbev/msm088.

133. 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.

134. Burge, S.W., Daub, J., Eberhardt, R., Tate, J., Barquist, L., Nawrocki, E.P., Eddy, S.R., Gardner, P.P., and Bateman, A. (2013). Rfam 11.0: 10 years of RNA families. Nucleic Acids Research *41*, D226–D232. 10.1093/nar/gks1005.

135. 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.

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. 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.

138. 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.

139. Whelan, S., Irisarri, I., and Burki, F. (2018). PREQUAL: detecting non-homologous characters in sets of unaligned homologous sequences. Bioinformatics *34*, 3929–3930. 10.1093/bioinformatics/bty448.

140. Ali, R.H., Bogusz, M., and Whelan, S. (2019). Identifying Clusters of High Confidence Homologies in Multiple Sequence Alignments. Mol Biol Evol *36*, 2340–2351. 10.1093/molbev/msz142.

141. Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*, 1972–1973. 10.1093/bioinformatics/btp348.

142. Tice, A.K., Žihala, D., Pánek, T., Jones, R.E., Salomaki, E.D., Nenarokov, S., Burki, F., Eliáš, M., Eme, L., Roger, A.J., et al. (2021). PhyloFisher: A phylogenomic package for resolving eukaryotic relationships. PLoS Biol *19*, e3001365. 10.1371/journal.pbio.3001365.

143. Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods

for Phylogenetic Inference in the Genomic Era. Molecular Biology and Evolution *37*, 1530–1534. 10.1093/molbev/msaa015.

144. Ronquist, F., and Huelsenbeck, J.P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics *19*, 1572–1574. 10.1093/bioinformatics/btg180.

145. Rambaut (2018). FigTree. tree.bio.ed.ac.uk/software/figtree/. tree.bio.ed.ac.uk/software/figtree/.

146. 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.

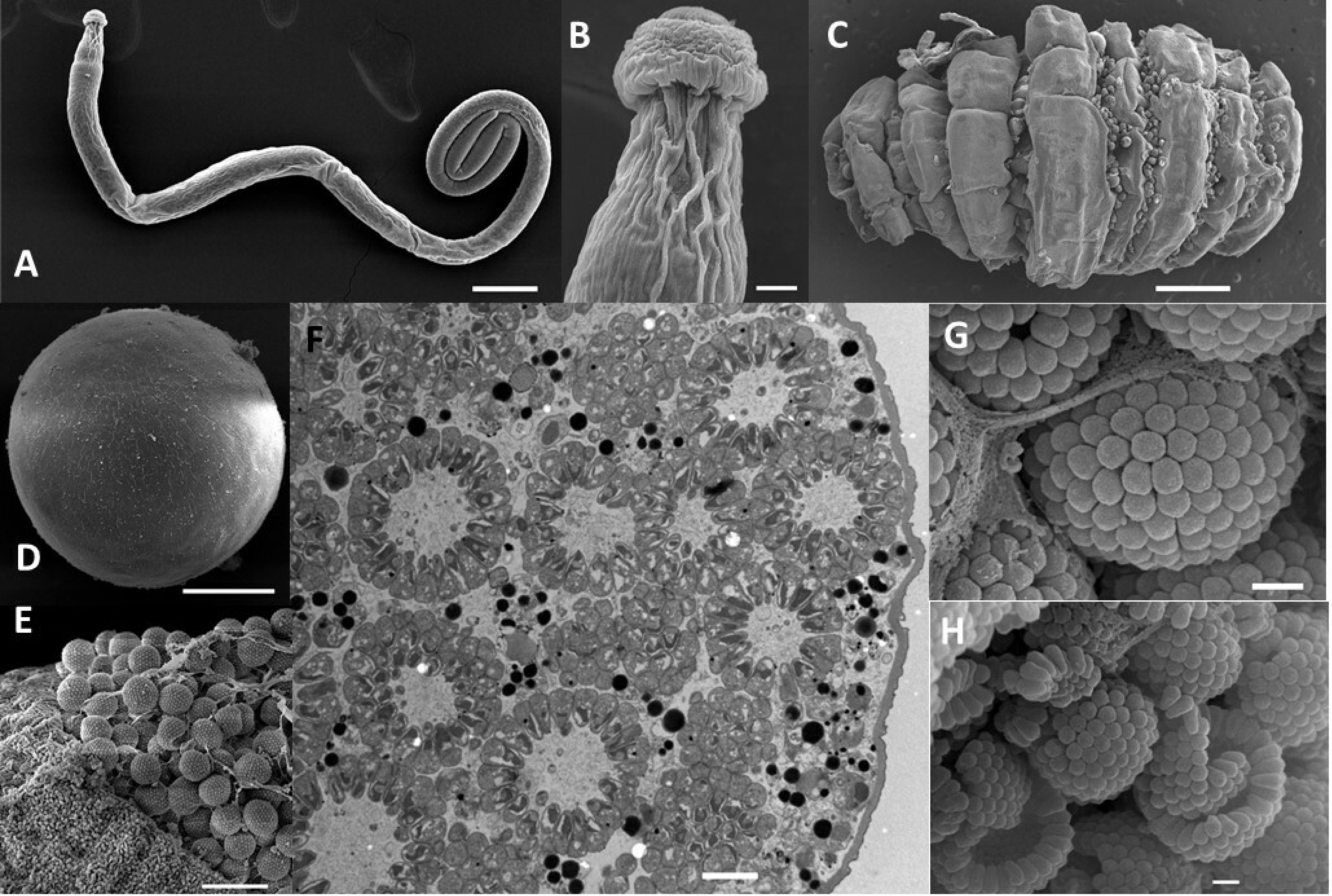
147. Castresana, J. (2000). Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. Molecular Biology and Evolution *17*, 540–552. 10.1093/oxfordjournals.molbev.a026334.

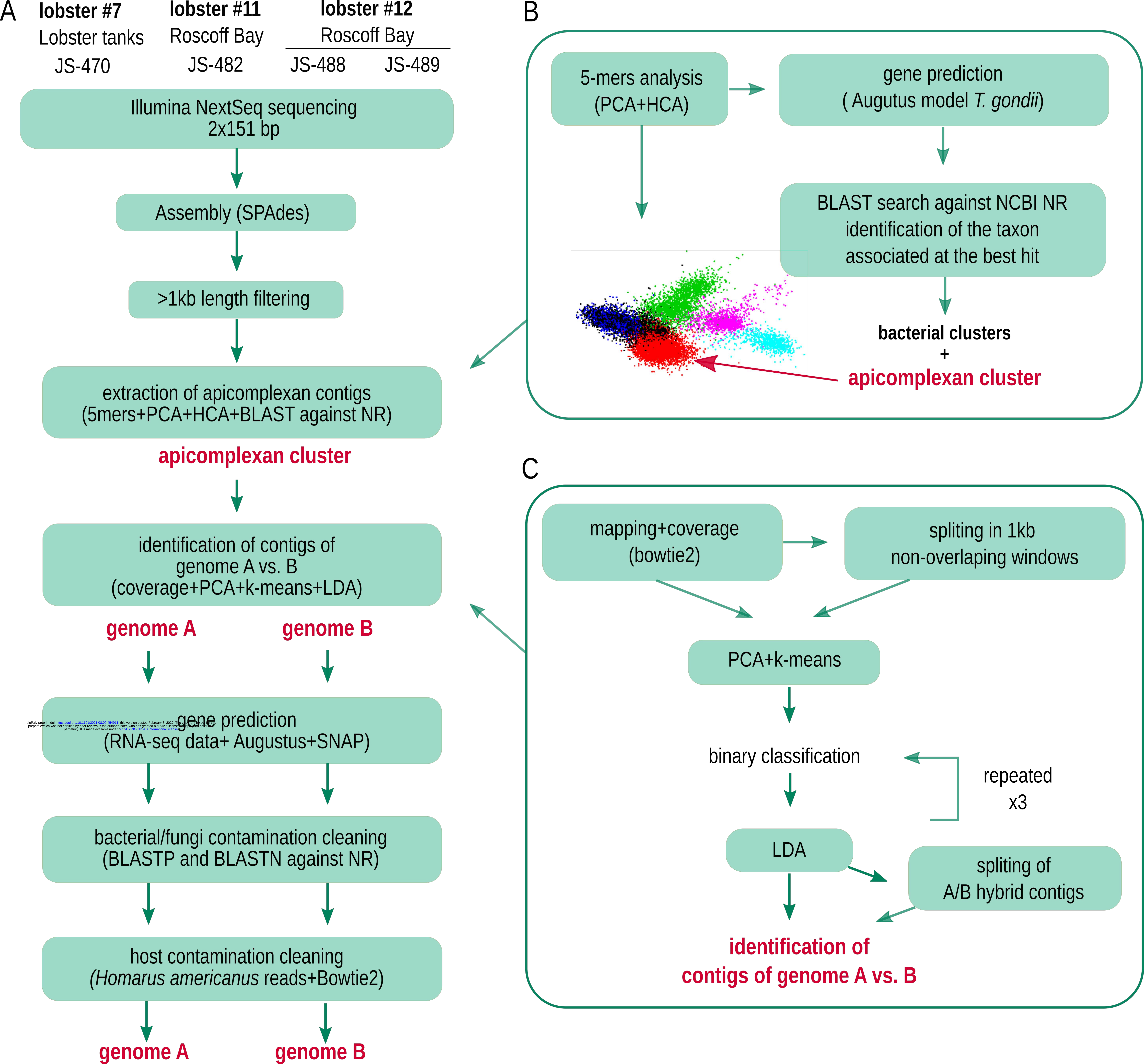
148. Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Molecular Biology and Evolution *35*, 1547–1549. 10.1093/molbev/msy096.

149. 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.

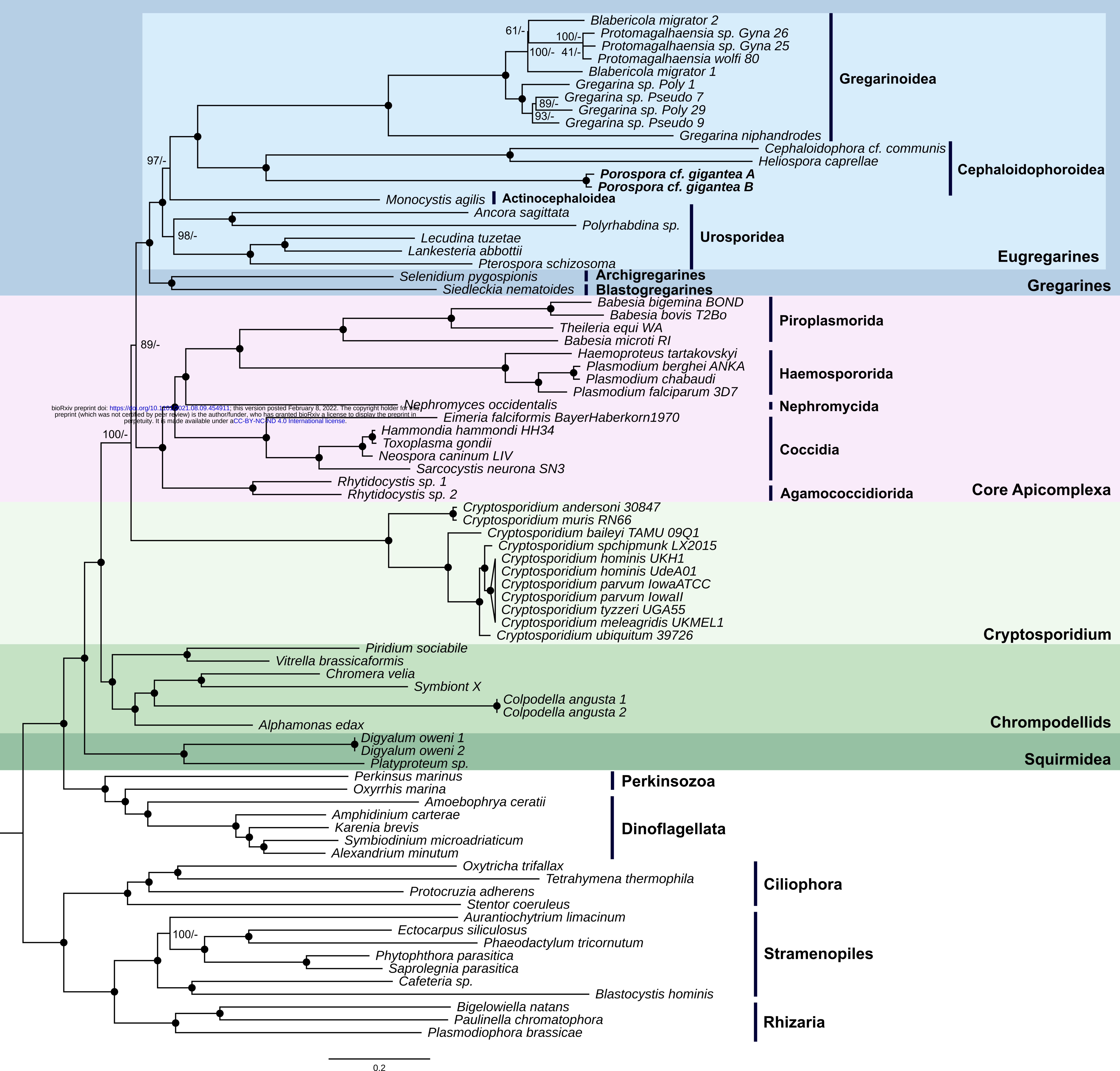
150. Wickham, H. (2009). Ggplot2: elegant graphics for data analysis (Springer).

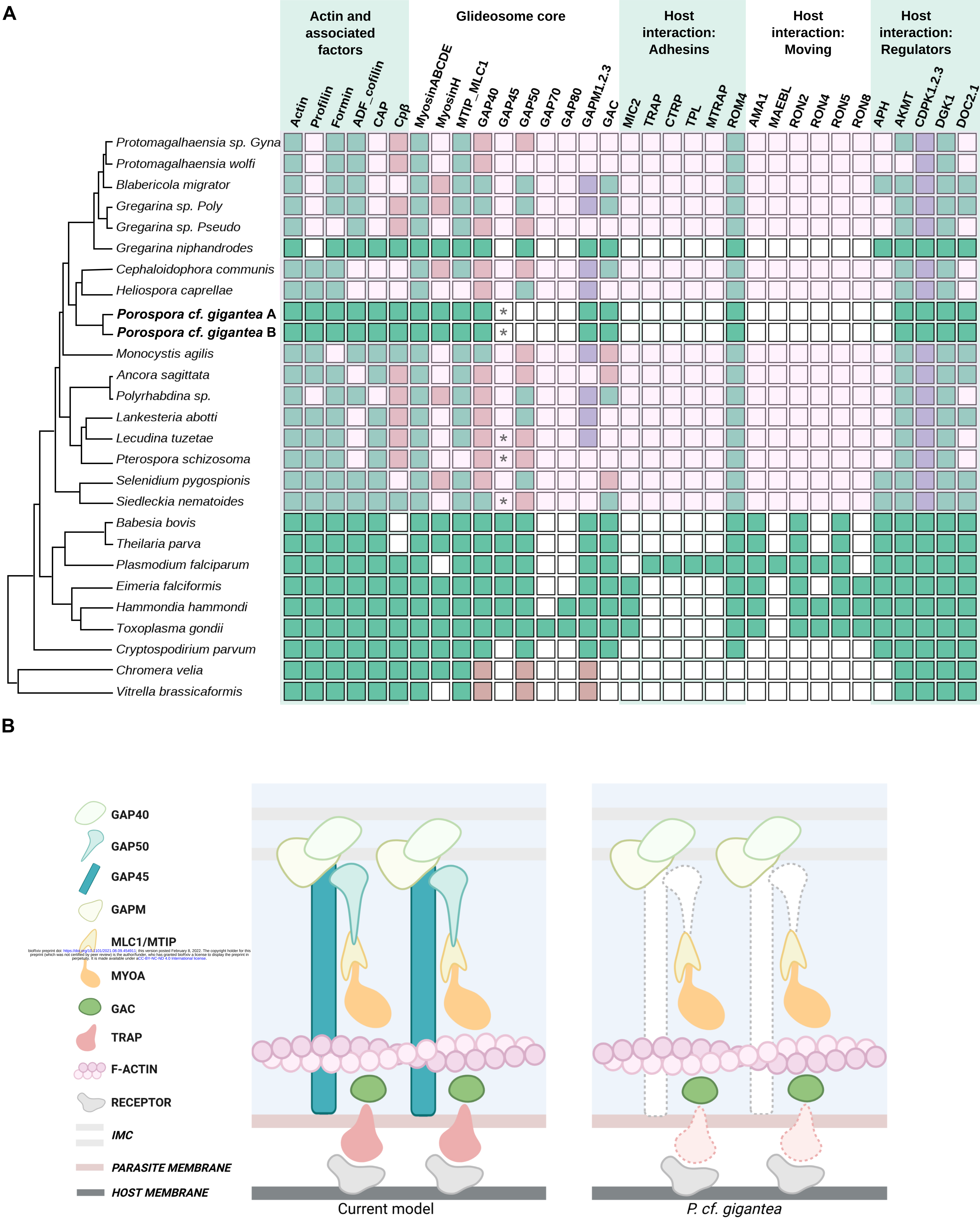
151. Birney, E. (2004). GeneWise and Genomewise. Genome Research *14*, 988–995. 10.1101/gr.1865504.

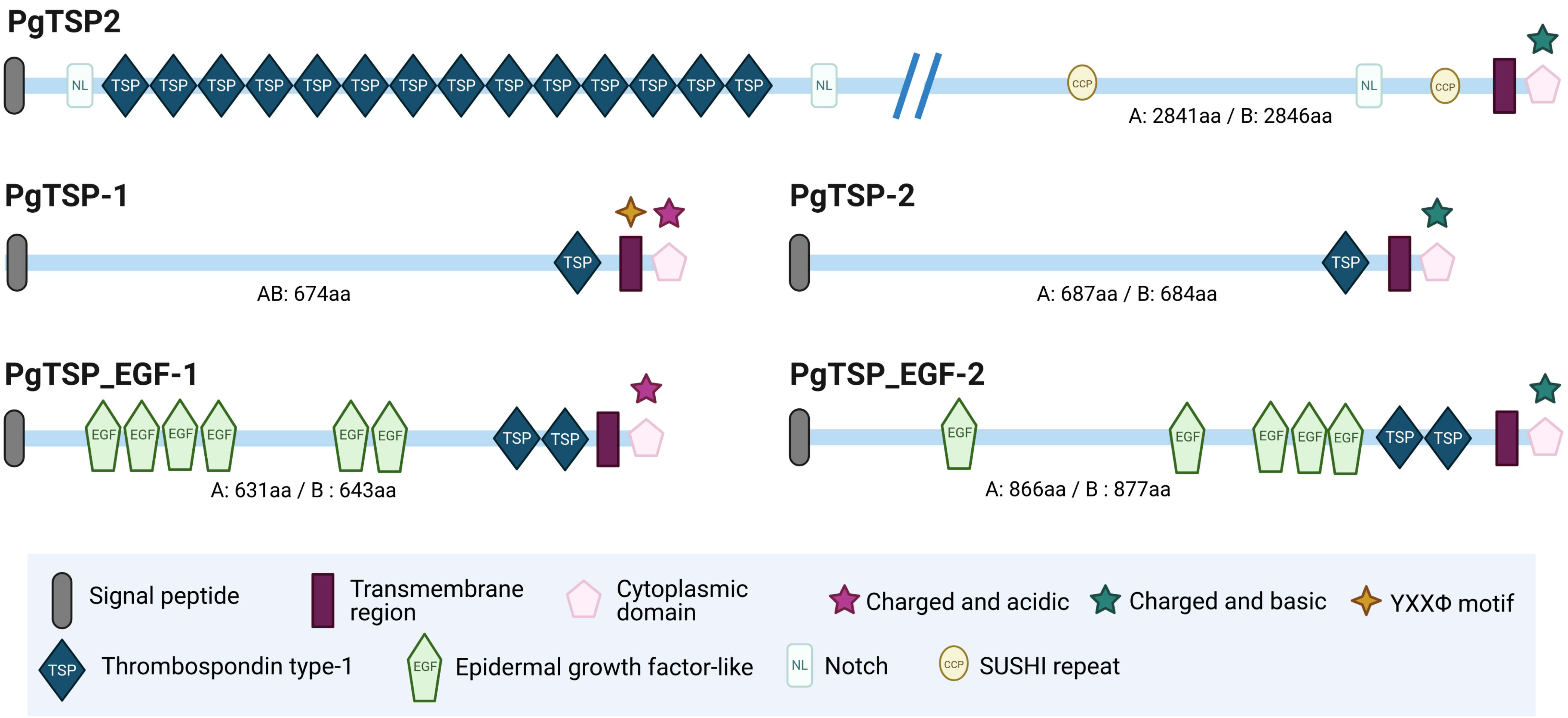

A
B
C
D
E
F
G
H


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

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
Pfalciparum
Tgondii
Cparvum
Gniphandrodes
PcfgiganteaB
PcfgiganteaA
4821
7664
3816
4860
5071
5031
7500
5000
2500
0
OrthoGp Per Species


61/-
100/-
100/-
41/-
89/-
93/-
97/-
98/-
89/-
100/-
100/-
Blabericola migrator 2
Protomagalhaensia sp. Gyna 26
Protomagalhaensia sp. Gyna 25
Protomagalhaensia wolfi 80
Blabericola migrator 1
Gregarina sp. Poly 1
Gregarina sp. Pseudo 7
Gregarina sp. Poly 29
Gregarina sp. Pseudo 9
Gregarina niphandrodes
Gregarinoidea
Cephaloidophora cf. communis
Heliospora caprellae
Porospora cf. gigantea A
Porospora cf. gigantea B
Cephaloidophoroidea
Monocystis agilis
Actinocephaloidea
Ancora sagittata
Polyrhabdina sp.
Lecudina tuzetae
Lankesteria abbottii
Pterospora schizosoma
Urosporidea
Eugregarines
Selenidium pygospionis
Archigregarines
Siedleckia nematoides
Blastogregarines
Gregarines
Babesia bigemina BOND
Babesia bovis T2Bo
Theileria equi WA
Babesia microti RI
Piroplasmorida
Haemoproteus tartakovskyi
Plasmodium berghei ANKA
Plasmodium chabaudi
Plasmodium falciparum 3D7
Haemospororida
Nephromyces occidentalis
Nephromycida
Eimeria falciformis BayerHaberkorn1970
Hammondia hammondi HH34
Toxoplasma gondii
Neospora caninum LIV
Sarcocystis neurona SN3
Coccidia
Rhytidocystis sp. 1
Rhytidocystis sp. 2
Agamococcidiorida
Core Apicomplexa
Cryptosporidium andersoni 30847
Cryptosporidium muris RN66
Cryptosporidium baileyi TAMU 09Q1
Cryptosporidium spchipmunk LX2015
Cryptosporidium hominis UKH1
Cryptosporidium hominis UdeA01
Cryptosporidium parvum IowaATCC
Cryptosporidium parvum IowaII
Cryptosporidium tyzzeri UGA55
Cryptosporidium meleagridis UKMEL1
Cryptosporidium ubiquitum 39726
Cryptosporidium
Piridium sociabile
Vitrella brassicaformis
Chromera velia
Symbiont X
Colpodella angusta 1
Colpodella angusta 2
Alphamonas edax
Chrompodellids
Digyalum oweni 1
Digyalum oweni 2
Platyproteum sp.
Squirmidea
Perkinsus marinus
Perkinsozoa
Oxyrrhis marina
Amoebophrya ceratii
Amphidinium carterae
Karenia brevis
Symbiodinium microadriaticum
Alexandrium minutum
Dinoflagellata
Oxytricha trifallax
Tetrahymena thermophila
Protocruzia adherens
Stentor coeruleus
Ciliophora
Aurantiochytrium limacinum
Ectocarpus siliculosus
Phaeodactylum tricornutum
Phytophthora parasitica
Saprolegnia parasitica
Cafeteria sp.
Blastocystis hominis
Stramenopiles
Bigelowiella natans
Paulinella chromatophora
Plasmodiophora brassicae
Rhizaria
0.2


A
Actin and associated factors
Glideosome core
Host interaction: Adhesins
Host interaction: Moving
Host interaction: Regulators
Actin
Profilin
Formin
ADF_cofilin
CAP
Cpβ
MyosinABCDE
MyosinH
MTIP_MLC1
GAP40
GAP45
GAP50
GAP70
GAP80
GAPM1.2.3
GAC
MIC2
TRAP
CTRP
TPL
MTRAP
ROM4
AMA1
MAEBL
RON2
RON4
RON5
RON8
APH
AKMT
CDPK1.2.3
DGK1
DOC2.1
Protomagalhaensia sp. Gyna
Protomagalhaensia wolfi
Blabericola migrator
Gregarina sp. Poly
Gregarina sp. Pseudo
Gregarina niphandrodes
Cephaloidophora communis
Heliospora caprellae
Porospora cf. gigantea A
Porospora cf. gigantea B
Monocystis agilis
Ancora sagittata
Polyrhabdina sp.
Lankesteria abotti
Lecudina tuzetae
Pterospora schizosoma
Selenidium pygospionis
Siedleckia nematoides
Babesia bovis
Theilaria parva
Plasmodium falciparum
Eimeria falciformis
Hammondia hammondi
Toxoplasma gondii
Cryptospodirium parvum
Chromera velia
Vitrella brassicaformis
*
B
GAP40
GAP50
GAP45
GAPM
MLC1/MTIP
MYOA
GAC
TRAP
F-ACTIN
RECEPTOR
IMC
PARASITE MEMBRANE
HOST MEMBRANE
Current model
P. cf. gigantea


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***	**Iowall**	**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)