Characterization of gregarine genomes and their deduced proteomes to understand the diversification of apicomplexans and their adaptation to parasitic lifestyle

Julie Boisard

MUSÉUM NATIONAL D'HISTOIRE NATURELLE

École Doctorale 227
Sciences de la Nature et de l'Homme : évolution et écologie

Année 2021

N°attribué par la bibliothèque
|_|_|_|_|_|_|_|_|_|_|_|_|_|

THÈSE

pour obtenir le grade de

DOCTEUR DU MUSÉUM NATIONAL D'HISTOIRE NATURELLE

Spécialité : Biologie des Organismes, Parasitologie et Génomique

présentée et soutenue publiquement par

Julie BOISARD

le 4 Octobre 2021

Characterization of gregarine genomes and their deduced proteomes to understand the diversification of apicomplexans and their adaptation to parasitic lifestyle

---

sous la direction de: **Pr Isabelle FLORENT, PR MNHN, Directrice de thèse**
Dr Loïc PONGER, MCF MNHN, Co-encadrant de thèse

devant le jury: **Dr Franck PANABIERES, DR INRAE, rapporteur**
Pr Philippe SILAR, PR Université Paris, rapporteur
Dr Laura EME, CR CNRS, examinatrice
Dr Gwenaël PIGANEAU, DR CNRS, examinatrice
Dr Isabelle TARDIEUX, DR CNRS, examinatrice

Contents

List of Figures

List of Tables

Chapter 1

Introduction : On the importance of gregarine genomics

1.1 Gregarines: well described yet forgotten at molecular level

Apicomplexa are unicellular eukaryotes (protists) collectively corresponding to ∼350 genera and ∼6000 named species, the wide majority of which have adopted a strict parasitic lifestyle in a very wide diversity of metazoan hosts (Portman and Šlapeta, 2014; Adl et al., 2019). Apicomplexa, together with the two sister groups Dinoflagellata and Ciliata, form Alveolata; which itself forms with Rhizaria, Stramenopiles and Telonemia the recent TSAR supergroup (Adl et al., 2019; Burki et al., 2019) (Figure 1.1).

Apicomplexa are mostly known for comprising infamous intracellular parasites of vertebrates responsible for important human diseases such as malaria due to *Plasmodium* spp., cryptosporidiosis due to *Cryptosporidium* spp. and toxoplasmosis due to *Toxoplasma gondii*. Apicomplexa also comprise diverse other intracellular parasites of vertebrates, with economical or veterinary importance such as *Eimeria* spp., *Babesia* spp. and *Theileria* spp. These apicomplexan parasites have simple to very complex life cycles. Some are restricted to single hosts (monoxenous parasites, e.g. *Cryptosporidium*, *Eimeria*). Other alternate between two successive hosts (dixenous parasites such as *Plasmodium*, *Babesia*

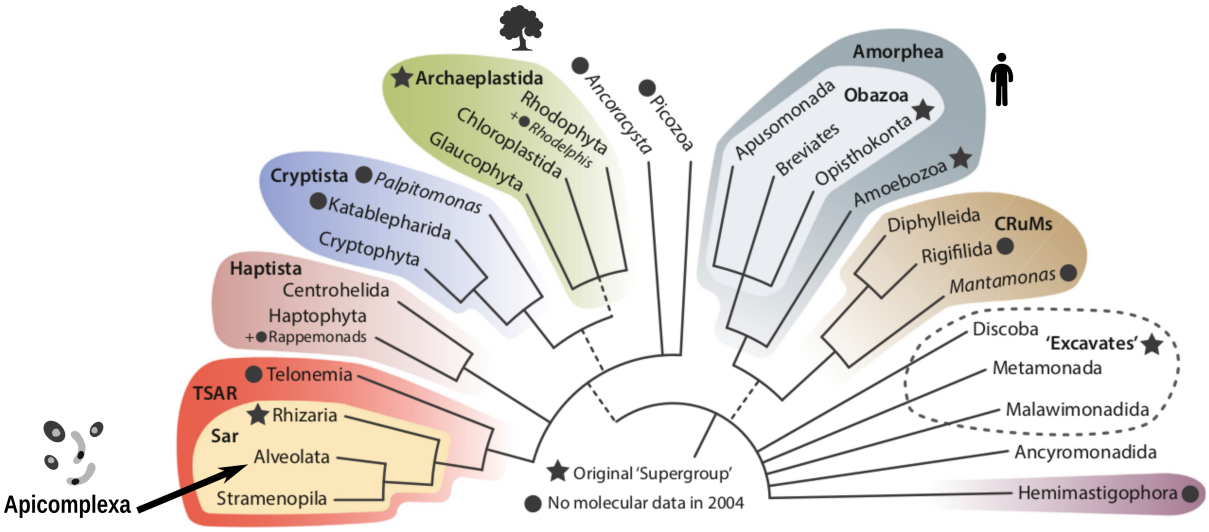


Figure 1.1: **The New Tree of Eukaryotes.** Modified from Burki et al. (2019).

and *Theileria*, completing sexual reproduction in various insects or arthropods and asexual phases in various tissues of vertebrates). Few have the capacity to infect multiple hosts (polyxenous parasites such as *Toxoplasma*, completing its sexual reproduction in cats and several asexual phases in various tissues of diverse warm blooded vertebrates).

Due to their medical, veterinary or economical importance, and because it has been possible to cultivate most of them in laboratory conditions, their genomes have been deciphered and ∼125 of them are deposited into the VEupathDB database (Aurrecoechea et al., 2017) (Figure 1.2, page 14). These genomes constitute major references for medical investigations, comparative genomics studies and exploration of evolutionary history of apicomplexan parasites (Janouškovec et al., 2015; Woo et al., 2015; Janouškovec et al., 2019; Kwong et al., 2019; Mathur et al., 2019, 2021a; Salomaki et al., 2021).

But Apicomplexa also comprise another group of organisms collectively known as "gregarines", that are principally monoxenous parasites of a wide diversity of non-vertebrate metazoan hosts, ranging from Polychaeta annelids to tunicates, arthropods and mollusks. They develop mostly extracellularly in the intestinal and coelomic cavities of their host - although some gregarines, the neogregarines, can have an intracellular development phase (Desportes and Schrével, 2013). These endoparasites are mostly considered as being non-pathogenic, with a few reported cases of recognised pathogenicity. However, experimental studies that can assess the actual pathogenicity of gregarines are still lacking (Rueckert

et al., 2019a).

There are several consequences to the mostly extracellular lifestyle displayed by gregarines, a feature that indeed distinguishes them from their intracellular parasites relatives (Desportes and Schrével, 2013; Schrével and Desportes, 2015). First, gregarines can reach very large sizes even for unicellular eukaryotes, from less than a micrometer to more than a millimeter for respectively the zoite and trophozoite forms of the marine eugregarine *Porospora gigantea*, intestinal parasite of the lobster *Homarus gammarus* (Desportes and Schrével, 2013; Schrével and Desportes, 2015). Insect eugregarines, such as *Gregarina garnhami*, intestinal parasite of the desert locust *Schistocerca gregaria*, display trophozoite forms reaching a dozen to several hundred micrometers (Desportes and Schrével, 2013; Schrével and Desportes, 2015). Archigregarine trophozoites, such as *Selenidium pendula*, intestinal parasite of the Polychaeta *Scolelepis (Nerine) squamata* also reach several dozen micrometers (Schrével et al., 2016). In comparison, *Toxoplasma*'s tachyzoite (developmental phase equivalent to the trophozoite) mean length is about 7 micrometers (Weiss and Kim, 2014). The large sizes of these developmental stages and their very common occurrence in a large diversity of non-vertebrate metazoan hosts have facilitated the discovery and biological studies on gregarines, resulting in an abundant literature on their morphologies, ultrastructures and life cycles, that have been examined by photonic and electron microscopy (SEM Scanning Electron Microscopy, TEM Transmission Electron Microscopy) imaging or by using cryoelectron microscopy or immunofluorescence (see Desportes and Schrével (2013) for exhaustive bibliography). Dynamic recordings are also available notably concerning their movements or behaviours (Desportes and Schrével, 2013). This rich literature offers a wide panorama of the adaptive capacities of these organisms to their hosts and environments, awaiting now exploration by -omic approaches to decipher the molecular bases of their functioning and variations, as it has been developed so far for their intracellular apicomplexan cousins (Rueckert et al., 2019b).

Indeed, the genomic knowledge on apicomplexan is currently highly biased in favour of intracellular parasites of vertebrate hosts, belonging to Haematozoa, Coccidia and to a lesser extent, the genus *Cryptosporidium* (Figure 1.2). Gregarines have been so far mostly

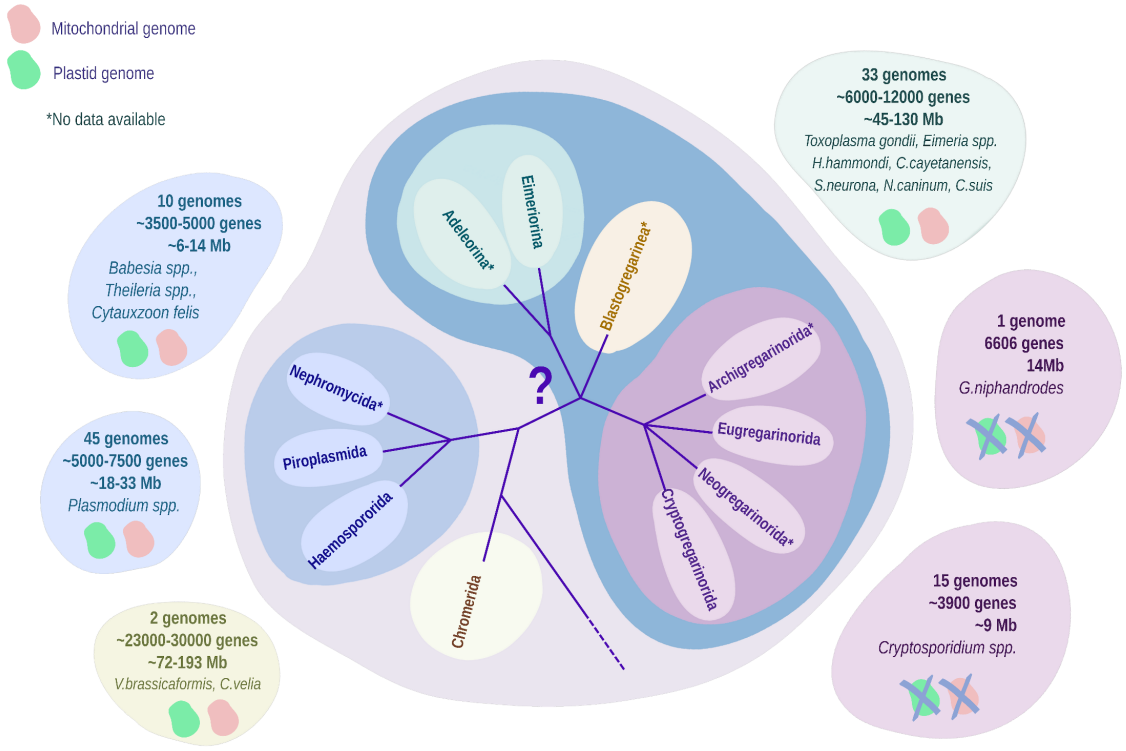


Figure 1.2: **The genomic panorama of Apicomplexa.** On this schematic representation inspired by Portman and Šlapeta (2014) and using the then most recent taxonomy by Adl et al. (2019), we have compiled the genomic information currently available on Apicomplexa and proto-Apicomplexa, mostly available from VEupathDB (Aurrecoechea et al., 2017). We have indicated for each group of data: the number of available genomes, specifying the concerned species, the number of protein-coding genes, the nuclear genome size and the presence (or absence) of mitochondrial or plastid genomes. The question mark symbolises the currently unresolved branching order of the various apicomplexan groups. In Boisard and Florent (2020).

excluded from this -omic exploration, to the exception of (unpublished) genome of the terrestrial insect eugregarine *Gregarina niphandrodes*, intestinal parasite of the mealworm beetle *Tenebrio molitor*. Indeed, this genome has no associated scientific publication describing it to this day, while being accessible through the database VEupathDB, section CryptoDB (Aurrecoechea et al., 2017). There is also very partial genomic data on insect eugregarine *Ascogregarina taiwanensis* (Templeton et al., 2010) intestinal parasite of the tiger mosquito *Aedes albopictus*, as well as partial and recently emerging transcriptomic data for a selection of terrestrial and marine gregarine species (Omoto et al., 2004; Janouškovec et al., 2019; Mathur et al., 2019; Salomaki et al., 2021; Mathur et al., 2021b) (see Figure 1.2 for illustration on Apicomplexa genomic data knowledge).

There are several reasons why the acquisition of -omic knowledge on gregarines is lagging behind that of their intracellular vertebrate parasite cousins:

1) as there are infecting "only" non-vertebrates hosts and are mostly considered non-pathogenic (Rueckert et al., 2019a), they have been neglected;

2) the current lack of *in vitro* culture methods for these parasites complicates the isolation of biological material in adequate amounts and quality for accurate usage in molecular investigations.

Indeed, gregarine biological studies must rely on field collections, mostly from infected hosts (alternatively their feces), which expose the collected material to contaminations by host cells and environmental microorganisms. The ability to maintain the hosts of a diversity of gregarines in laboratory conditions, offers a good compromise as it allows regular access to different developmental stages amenable to a variety of cellular (microscopy, test of inhibitors) and molecular (-omics) studies (see Desportes and Schrével (2013) for exhaustive descriptions). For example, the ragworm *Hediste (Nereis) divesicolor*, host of the marine eugregarine *Lecudina tuzetae*, can be maintained for several months in natural or artificial sea water in the laboratory (Kuriyama et al., 2005; Desportes and Schrével, 2013). In the same line of thought, several insects raising facilities may be used to get access to their infecting gregarines (Clopton, 2009; Desportes and Schrével, 2013; Florent et al., 2021). Finally, it can be expected that the concomitant progression of -omics and microscopic methodologies, allowing using increasingly reduced amounts of biological material, will also facilitate in a near future the bridging of this molecular knowledge gap, between the currently very poorly documented non cultivable gregarines and the increasingly well documented cultivable intracellular parasites of vertebrates (Gawad et al., 2016).

1.2 Why should we study gregarines?

Studying gregarines at the molecular level would yield novel knowledge about apicomplexan parasites. Gregarines display unique characteristics, notably their essentially ex-

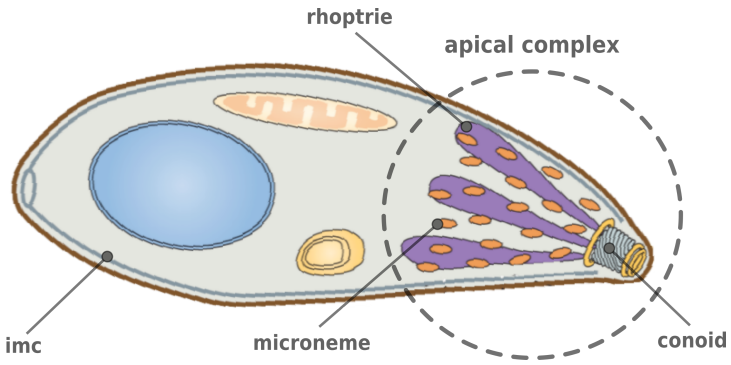


Figure 1.3: **Schematic representation of the zoite, the infective stage common to all apicomplexans.** The apical complex is circled. "imc" stands for inner membrane complex, a three lamellar membrane made from flattened membrane sacs termed alveoli. Adapted from Frénal et al. (2017).

tracellular life style and its biological consequences (Desportes and Schrével, 2013; Schrével and Desportes, 2015; Adl et al., 2019; Rueckert et al., 2019a), which we propose to expose in detail in this section (see detailed biological cycles of three different gregarines in Figure 1.4, page 18).

Like all Apicomplexa, gregarines present at least once during their life cycle a developmental form called zoite (Figure 1.3), a polarised cell comprising the so-called "apical complex" composed of scaffolding cytoskeletal elements enclosing specialized apical organelles (rhoptries, micronemes and dense granules) that gave the name to this phylum replacing the former Sporozoa (Morrissette and Sibley, 2002; Tardieux and Baum, 2016; Adl et al., 2019). As displayed in Figure 1.2, Gregarines do have a conoid, composed of spirally arranged array of microtubules as found in Coccidia and Cryptospodia (forming Conoidasida), but secondarily lost in Haemosporidia (Portman and Šlapeta, 2014; Adl et al., 2019) - forming Aconoidasida, although recent studies have shown that a divergent and greatly reduced form of the conoid is in fact conserved in *Plasmodium* species (Kořený et al., 2021; Bertiaux et al., 2021).

In Apicomplexa having intracellular developmental phase(s), this apical complex has been clearly involved in the recognition and invasion of host cells, allowing parasites establishment and development in this novel ecological niche (Tardieux and Baum, 2016; Hakimi et al., 2017). In Apicomplexa displaying extracellular lifestyle as most intestinal gregarines do, the apical complex appears to have a different role. Yet involved in par-

asite attachment to host cells at sporozoite stage, it subsequently appears mostly used for parasite feeding, sustaining spectacular growth phases, rather than to achieve tissue penetration or parasite internalisation within host cells (Valigurová and Koudela, 2008; Simdyanov and Kuvardina, 2007; Schrével et al., 2016; Valigurová and Florent, 2021). This is with the notable exception of coelomic (eu)gregarines, which are capable of intestinal barrier crossing as well as neogregarines that can reach intracellular niches in some of their hosts tissues (Desportes and Schrével, 2013; Schrével and Desportes, 2015).

There are several consequences to this extracellular growth of these trophozoite forms in gregarines:

1) an extremely wide diversity of shapes and sizes, as mentioned above and largely used for taxonomical purposes (Desportes and Schrével, 2013; Valigurová and Florent, 2021);

2) a sexual phase (gamogony followed by sporogony) that also occurs extracellularly, producing developmental forms that are particular to gregarines, starting with the syzygy (Desportes and Schrével, 2013; Schrével and Desportes, 2015).

Syzygy is the named given to the developmental stage that precludes the gregarine sexual reproduction. It corresponds to the bi-association of two trophozoites after they have detached from host cell and which are future gamonts (Figure 1.4, stages (a) and (b/b')). Although morphologically similar both in size and shape, the two partners of the syzygy are committed to evolve into respectively male and female gamonts (Figure 1.4, stages (c)).

Depending on the species, the bi-association may be caudo-caudal (Figure 1.4.A, stage (c)), lateral (Figure 1.4.B, stage (c)), fronto-frontal (not shown, see Desportes and Schrével (2013)) or caudo-frontal (Figure 1.4.C, stage (c)). In this latter case, found in *Gregarina garnhami*, the anterior partner of the syzygy is called primite, the other being known as the satellite (Figure 1.4.C, stage (c)). Occasionally, syzygy associations may involve more than two partners but the evolution of such *ménage à trois* has not been yet examined at molecular level (Desportes and Schrével, 2013).

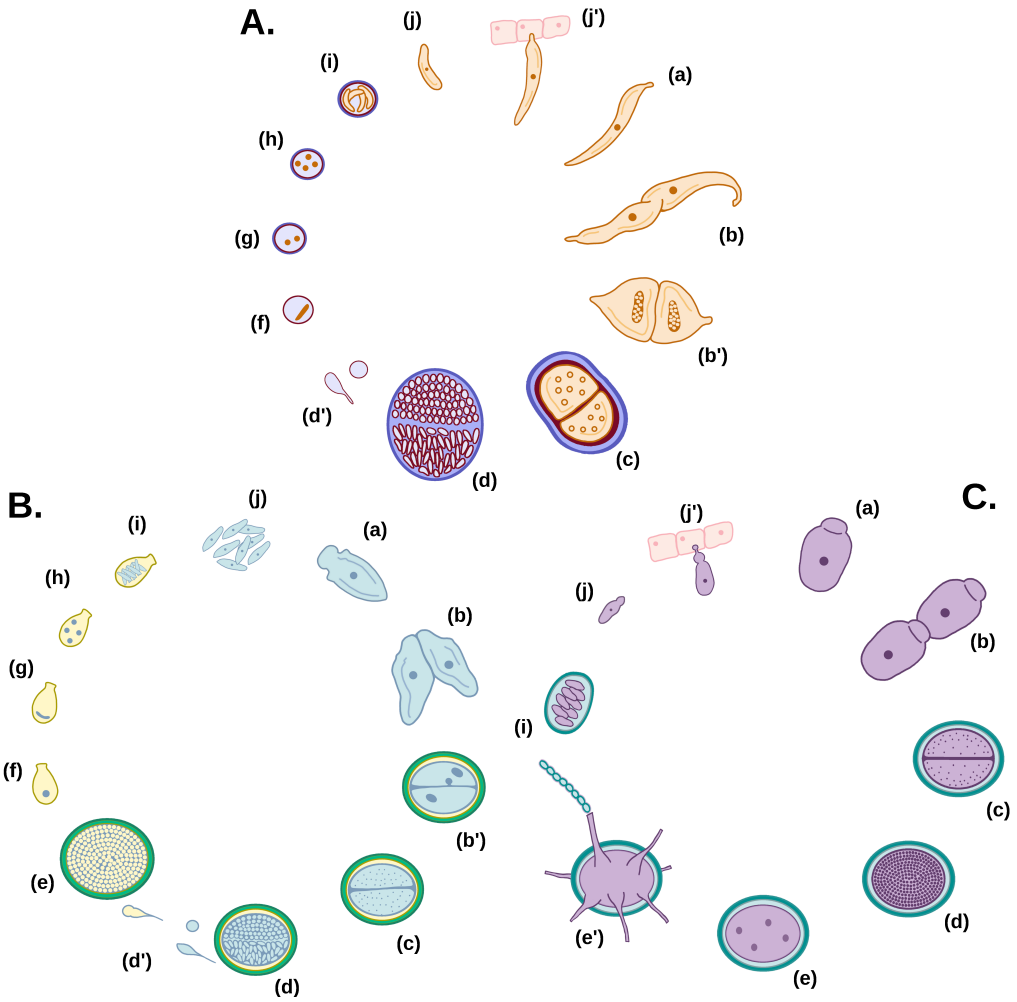


Figure 1.4: **Representative development cycles for three gregarines**. The developmental cycles of: (A) the marine archigregarine *Selenidium pendula*, intestinal parasite of the Polychaeta worm *Scolelepis (Nerine) squamata*, adapted from Schrével and Desportes (2015); (B) the marine eugregarine *Lecudina tuzetae*, intestinal parasite of the Polychaeta worm *Hediste (Nereis) diversicolor*, adapted from Schrével and Desportes (2015); (C) the terrestrial eugregarine *Gregarina garnhami*, intestinal parasite of the desert locust *Schistocerca gregaria*, based on Canning (1956) and personal observations. The drawings use identical legend letters to designate similar developmental stages across the three cycles. (A) *Selenidium pendula.* (a) detached trophozoite; (b) caudal syzygy; (b') particularity in syzygy for this species (nuclear modifications before encystment); (c) gametocyst undergoing gamogony; (d) gametocyst with fully differentiated gametes; (d') details of male (flagellated) and female (ovoid) gamete; (f) zygote ready to undergo sporogony yielding stages with two nuclei (g), then four nuclei (h); (i) spore containing four sporozoites; (j) released sporozoite (in host) starting vegetative phase. (B) *Lecudina tuzetae.* (a) detached trophozoite; (b) lateral syzygy; (c) gametocyst undergoing gamogony; (d) gametocyst with fully differentiated gametes; (d') details of male (flagellated) and female (ovoid) gamete; (e) sporokyst enclosing $\sim$ 5000 zygotes ready to evolve into spores (f) eventually undergoing sporogony yielding stages with two nuclei (g), four nuclei (h); (i) spore containing eight sporozoites; (j) released sporozoite (in host) starting vegetative phase including attachment to host epithelial cell (not shown). (C) *Gregarina garnhami.* (a) detached trophozoite; (b) caudo-frontal syzygy (primite ahead, satellite following); (c) gametocyst undergoing gamogony; (d) gametocyst with fully differentiated gametes; (e) sporokyst enclosing zygotes ready to evolve into spores undergoing sporogony (details not shown); (e') these spores are released in the environment as spore chains through sporoducts emerging from the sporokyst; (i) spore containing eight sporozoites; (j) released sporozoite (in host) starting vegetative phase including attachment to host epithelial cell (j'). Cyst or spore walls surround developmental stages from (c) to (i). In Boisard and Florent (2020).

The evolution of the syzygy bi-association is a globular structure called gametocyst, initially composed of two hemispheres of similar shape and volume around which a thick wall is elaborated (Figure 1.4, stages (c)). Series of nuclear divisions with final cytokinesis (gamogony) then occur within each hemisphere producing male gametes within one hemisphere and female gametes within the other one (Figure 1.4, stages (d)). A clear anisogamy is commonly observed between male gametes - more pyriform and usually flagellated - and female gametes - more globular and non-flagellated (Figure 1.4.A and .B, stages (d')). It is therefore only after complete gametes production that the "sex" of gamonts may be deduced. Numerous imaging recordings have been performed to study the cellular events occurring during this first phase of the gregarine sexual reproduction called gamogony, with a remarkable confocal imaging analysis performed in the case of the marine eugregarine *Lecudina tuzetae*, intestinal parasite of the Polychaeta *Hediste (Nereis) diversicolor*, in which ∼5000 male and as many female gametes are produced per gametocyst (Kuriyama et al., 2005).

It is interesting to notice that, depending on the species, the length of the syzygy phase might be particularly long; in the atypical case of *Diplauxis hatti*, coelomic parasite of the Polychaeta *Perinereis cultrifera*, this bi-association remains stable for more than two years, awaiting host's sexual maturation to engage into gamogony (Prensier et al., 2008).

Once the male and female gametes have been produced within their respective compartments, their mixing occurs within the gametocyst. Thousands of fertilisations then take place simultaneously during a process called "gamete dance" in *L. tuzetae*, which lasts ∼4h and produces ∼5000 zygotes per gametocyst (Kuriyama et al., 2005) (Figure 1.4.B, stage (e)). Sporogony then begins and the gametocyst takes the name of sporokyst (Figures 1.4B. and .C, stages (e/e')).

Each zygote secretes a cyst wall (stage called immature oocyst) and undergoes meiosis and additional mitosis (in eugregarines and neogregarines) leading to sporozoites formation (Figure 1.4, stages (f) to (i)). Each spore, also called oocyst, therefore possesses a thick wall and is the form of dissemination of the gregarine in the environments (De-

sportes and Schrével, 2013; Schrével and Desportes, 2015). It contains four sporozoites in the case of archigregarines and eight sporozoites in eugregarines and neogregarines (Figure 1.4, stages (i)) (Desportes and Schrével, 2013; Schrével and Desportes, 2015).

In marine eugregarines, spores are eventually released in the environment with the breaking down of the sporokyst (Figure 1.4.B, stage (e) in the case of *L. tuzetae*). In terrestrial gregarines, spores are released in the environments via sporoducts that are formed at the surface of the sporokyst (see for example *Gregarina garnhami*, Figure 1.4.B, stage (e')). The progeny in terms of number of spores, resulting from the evolution of a single syzygy, is considerable and can reach few thousands to several millions depending on the gregarine species (Desportes and Schrével, 2013; Schrével and Desportes, 2015). Sporokyst are therefore obviously a material of choice to isolate genomic DNA. Oocysts on the other hand, are the developmental forms that are likely collected within soils and sediments from environments.

Once ingested by hosts, the oocysts undergo dehiscence after passage through the host's digestive system and sporozoites are released (four in the case of archigregarines, eight in the case of eugregarines and neogregarines, Figure 1.4, stages (j)). These will be able to attach to their host's intestinal cells (Figure 1.4.C, stage (j')), using their apical end displaying typical apical complex features, and these attached sporozoites will start to grow dramatically, evolving into trophozoites. In the case of *G. garnhami* for example, trophozoites grow from less than 10 µm to over 400 µm in length within a single host. The size range of trophozoites may therefore cover two to three orders of magnitude depending on the species (Desportes and Schrével, 2013; Schrével and Desportes, 2015). This gigantism, achieved for a large number of gregarine species, offers a remarkable material for cell biology explorations and immunofluorescence imaging for example (Kuriyama et al., 2005; Valigurová et al., 2013; Valigurová and Florent, 2021).

This radical difference in lifestyle including gamogonic and sporogonic phases contrasts from those observed in Hematozoa and Coccidia and results also in very different interactions between gregarines and their hosts. These particularities raise many questions about the adaptive pathways gregarines have developed to survive over the course

of evolution. Which molecular solutions have they then developed to survive within the host's intestinal tracts or other cavities, maintain survival, acquire nutriments, complete (a)sexual reproduction ? Furthermore, it must be noticed that they have done so with a remarkable success if one considers their wide occurrence in such a high diversity of endoparasitic non-vertebrate hosts contexts (Desportes and Schrével, 2013; Schrével and Desportes, 2015; Valigurová and Florent, 2021).

The following sections introduce some possible topics of exploration, focused on gregarines, to broaden our view of host adaptation patterns in Apicomplexa. They are obviously not exhaustive; however, they gather the questions that we are particularly curious about, and that partly underlie the aims of this PhD, which will be outlined thereafter.

1.2.1 The apical complex

Gregarines, as apicomplexan parasites, do possess a fully developed apical complex, at least in sporozoites and trophozoites developmental stages (Desportes and Schrével, 2013). Biological and morphological studies have established that in gregarines, the apical complex is used for host cell attachment to allow the parasite to feed from its host cell by a process known as myzocytosis (Schrével et al., 2016; Simdyanov and Kuvardina, 2007; Valigurová and Florent, 2021). The host cell penetration by the parasite is not complete as the gregarine remains extracellular with only its apical end intimately engaged in a host–parasite interplay that has been studied at microscopic level but whose molecular actors are poorly defined (Valigurová et al., 2007; Schrével et al., 2016; Simdyanov et al., 2017; Valigurová and Florent, 2021); see also Desportes and Schrével (2013) for exhaustive descriptions of several additional examples. Therefore, the biological function of "gregarine apical complex" only partly overlaps the biological function currently attributed to "apicomplexan parasite apical complex", that is recognition and invasion of their host cell (Tardieux and Baum, 2016).

Several questions therefore emerge: to which extent does the molecular architecture of "gregarine apical complexes" compares to that of their best known cousins, that are

Toxoplasma and *Plasmodium* (Boucher and Bosch, 2015; Tardieux and Baum, 2016)? Are the scaffolding of the apical complex, and its recognition, invasive and nutrition functions fulfilled by the same molecular components or by other ones? What are the composition and functional roles of micronemes, rhoptries and dense granules in gregarines?

In apicomplexan parasites of vertebrates with intracellular developmental phases (*Plasmodium* and *Toxoplasma*), these secretory organelles are documented to intervene sequentially in an orchestrated manner, with first micronemes secreting parasite proteins involved in host cell recognition as well as AMA-1, which, when combined to a defined cortege of RON proteins (secreted by the neck of rhoptries) will assemble into the so-called mobile junction essential to the invasion process of host cells by *Plasmodium* merozoites or *Toxoplasma* tachyzoites (reviewed in Tardieux and Baum (2016)). Subsequently, ROP proteins (secreted by the bulb of rhoptries) and GRA proteins (secreted by the dense granules) will allow establishing the intracellular niche for these parasites, either at the parasitophorous vacuole level (facilitating metabolite exchanges) or beyond this border to manipulate the host cell program to the benefit of the parasite (Hakimi et al., 2017). The apical complex is also involved in the motility of apicomplexan parasites, through adhesion proteins secreted by micronemes as part of an apicomplexan-specific form of displacement, so-called gliding, to which we shall return later (Boucher and Bosch, 2015; Tardieux and Baum, 2016).

It may be expected that gregarine micronemes, rhoptries and dense granules will have their own protein repertoires; indeed, there are currently very limited overlaps of these repertoires between currently described apicomplexan genera (Counihan et al., 2013; Boucher and Bosch, 2015; Hakimi et al., 2017). Thus it will be interesting to decipher their specific roles and how they contribute (or not) to the establishing of the host-parasite interface (Valigurová and Florent, 2021). It is important here to indicate that there are alternative invasive modes in apicomplexan parasites such as *Theileria* and *Cryptosporidium* that differ from the better described *Toxoplasma/Plasmodium* mode (Gubbels and Duraisingh, 2012). For instance, *Theileria*'s non-motile invasive stages don't possess secretory organelles such as micronemes. Therefore invasion doesn't rely on

the parasite's cytosqueleton and apical complex but its contact with a host cell rather happens by chance (Shaw, 2003; Woods et al., 2021; Jalovecka et al., 2018). For their part, *Cryptosporidium* spp. are characterized by an epicellular localization (Barta and Thompson, 2006; Bartošová-Sojková et al., 2015). What are the molecular similarities between the *T. gondii* and *P. falciparum* parasitophorous vacuole make up and the food vacuole of gregarines, that forms at the gregarine-host cell interface (Valigurová et al., 2007; Schrével et al., 2016; Simdyanov et al., 2017; Valigurová and Florent, 2021)? Or is the similarity stronger to the feeder organelle of epicellular *Cryptosporidium* (Barta and Thompson, 2006; Bartošová-Sojková et al., 2015; Valigurová and Florent, 2021)?

1.2.2 Their extracellular development

A first morphological and biological consequence of this particular behaviour selected over evolution is the fact mentioned above that gregarine trophozoites can reach very large sizes (up to several millimeters) contrary to intracellular parasites of vertebrates (1-10 micrometers at most).

A second is that their sexual phase is also extracellular, starting with the syzygy that evolves into gametocysts producing oocysts (Figure 1.4, page 18). These developmental forms are strikingly distinct from the developmental forms encountered in *Toxoplasma*, *Plasmodium* and even *Cryptosporidium* (Aly et al., 2009; Robert-Gangneux and Dardé, 2012; Bouzid et al., 2013). Interestingly, the oocysts forms of *Toxoplasma* and *Cryptosporidium*, which are also extracellular and disseminated with their hosts' feces as resistant forms in the environments, are however elaborated intracellularly, within their hosts' intestinal cells (Robert-Gangneux and Dardé, 2012; Bouzid et al., 2013). Indeed, in *Toxoplasma* and *Cryptosporidium* the gamogony remains intracellular, whereas it is extracellular in gregarines.

An important consequence is that gregarines thus display totally different types of host-parasite interactions, having other constraints to face such as surviving in host-gut environment. Several studies have explored the permeability of the trophozoite membrane in link with the question of their nutrition mode (Desportes and Schrével, 2013; Schrével

and Desportes, 2015; Valigurová and Florent, 2021). This questions the molecular nature and the biological role of their inner membrane complex (see imc in Figure 1.3), which, interestingly, may be acting as a continuous "shield" all around the trophozoite as it is only interrupted at the conoidal opening through which nutrition occurs; see Schrével et al. (2016) for the case of *S. pendula* or Kuriyama et al. (2005) for the case of *L. tuzetae*; for a review on nutrition in apicomplexan, see Valigurová and Florent (2021).

Interestingly, as the syzygy evolves towards the gametocyst form, this imc appears to be disassembled concomitantly with the gametocyst secretion of the protective cyst wall (see Kuriyama et al. (2005) for details on *L. tuzetae*). This suggests that in gregarines, one form of shielding (imc) in trophozoites and syzygies gives place to another form of shielding (cyst wall) during gamogony then sporogony, both of which being intended to isolate the gregarine from its hostile (gut) environment.

Obviously, the molecular exploration of such a parasite–environment (host) interplay will reveal novel adaptive features developed by Apicomplexa over evolution. To which extent the host-gut environment is less hostile in an invertebrate host rather than in a vertebrate one, notably regarding immune system response and microbiota regulation, should certainly be taken into account. First, invertebrate hosts rely mainly on innate immunity to fight intruders while vertebrates also have adaptive defence mechanisms (Buchmann, 2014). In addition, the co-existing microbiome is notoriously less complex and diverse in invertebrates than vertebrates (Bahrndorff et al., 2016).

These differences could explain the capacity of gregarines to self-maintain in such host-gut environment for extended life cycle times while parasites of vertebrates have been constrained to invade host cells to achieve their maintenance in hosts. Further studying not only gregarines but also their host's immune and microbiota responses will certainly clarify the contribution of these host-specific features to the diversity of gregarines behaviours and life traits over evolution.

1.2.3 A greater diversity of motility modes

Gregarines have developed a wider diversity of motility modes than what is mostly described (and deeply studied at molecular level) for intracellular parasites of vertebrates: the gliding motility (Frénal et al., 2017).

This movement, governed by an acto-myosin motor, involves at least ∼35 proteins that appear (so far) well conserved between apicomplexan species (Boucher and Bosch, 2015; Frénal et al., 2017; Mueller et al., 2017). Whether the gliding components are conserved also in gregarines that move by gliding (i.e., most eugregarines) remains to be established.

However, gregarines display other modes of motility such as rolling or bending, the molecular bases of which are currently totally unknown (Desportes and Schrével, 2013). Do these alternative modes involve molecular components shared with those of the glideosome? Do they involve other components, inherited from the putative ancestor and possibly lost secondarily in intracellular parasites of vertebrates (Janouškovec et al., 2015; Woo et al., 2015; Füssy and Oborník, 2017)? Or do they involve novel components, refunctionalized from the ancestor heritage or acquired by horizontal gene transfer? All this remains to be established for the diversity of known and to-be-discovered gregarines.

Answers to these questions will be precious to understand how such a diversity of motility modes may have emerged for apicomplexan from a common ancestral genetic heritage, as apicomplexan are thought to derive from ancestral biflagellated organisms with repositioning of some of the former flagellar components into the apical complex structure and functioning (Janouškovec et al., 2015; Woo et al., 2015; Füssy and Oborník, 2017). This knowledge will also be precious to widen our current knowledge of the adaptive capacities to hosts developed by these remarkable apicomplexan parasites.

Indeed, evolutionary molecular studies on this point have established that gliding components partially existed in the common ancestor of Apicomplexa (Janouškovec et al., 2015; Woo et al., 2015; Füssy and Oborník, 2017) but have been repositioned to be functionally operational in intracellular parasites of vertebrates. What paths of specialisation did they follow to generate such a diversity of movements in gregarines? And, corollary to these observations, we can formulate the hypothesis that the lack of host-cell fully

invasive capacities of gregarines may be due either to:

- absence of gliding capacities despite a developed/expected to be functional apical complex (in the case of archigregarines)
- or, conversely, to an under-developed apical complex despite operational gliding capacities (in the case of eugregarines)

It is therefore time to focus on gregarines to explore these adaptive traits at the molecular level. However, which ones to select and from which extent of diversity? Indeed, a recent convergence of novel data clearly indicates that the current inventory of the true gregarine diversity is dramatically underestimated - and therefore, corollary, all the relevant biological models may have not yet been discovered.

1.3 The true gregarine biodiversity

Our current understanding of gregarine biodiversity comes uniquely from three sources of information that only partly overlap:

1) the number of formally inventoried species (Portman and Šlapeta, 2014);
2) the number of species theoretically computable based on the inventory and diversity of their hosts (Desportes and Schrével, 2013);
3) environmental or host-associated metagenomic or metabarcoding approaches, that have revealed novel molecular signatures, sufficiently related to gregarines to allow taxonomic affiliation to this group but sufficiently divergent to strongly suggest novel taxonomic species (de Vargas et al., 2015; Mahé et al., 2017; del Campo et al., 2019).

The regular cross-referencing of these three sources of data leads to a permanent readjustment of both the taxonomy and phylogeny of these species, so that it is safe to say that the current biodiversity of gregarines is a field of investigation whose physiognomy is likely to evolve considerably in the coming years.

1.3.1 Formally described species

Regarding the first point, there are currently ∼1770 formally described gregarine species, unequally distributed between archigregarines (∼20), eugregarines (∼1700) and neogregarines (∼50) (Portman and Šlapeta, 2014). In parallel, taxonomic and phylogenetic revisions concerning gregarines are a currently very active field with a diversity of successive proposals regarding their phylogenetic inter-relations (Cavalier-Smith, 2014; Schrével et al., 2016; Rueckert and Horák, 2017; Simdyanov et al., 2017, 2018) as well as with other apicomplexan parasites (Janouškovec et al., 2019; Mathur et al., 2019, 2021a).

Molecular phylogenies are nowadays mostly based on usage of SSU rDNA marker, more rarely complete ribosomal loci (Diakin et al., 2016, 2017; Simdyanov et al., 2018). Studies based on the SSU rDNA marker alone are fairly effective in defining monophyletic groups at the genus or family levels, but fail to robustly resolve the respective branches' relationships at higher taxonomic level. Attempts to improve phylogenies using the full ribosomal marker (18S SSU + 28S LSU rDNA) have provided some progress, but have the important disadvantage to be currently available for only very few gregarine species (∼20) (Diakin et al., 2016, 2017; Simdyanov et al., 2018).

Phylogenies relying on multiple genes (or more accurately proteins) sequences derived from available apicomplexan genomes as well as recent gregarine transcriptome investigations are now emerging but remain restrained to a few gregarine species, namely *Selenidium pygospionis*, *Pterospora schizosoma*, *Lankesteria abbottii*, *Lecudina tuzetae*, *Polyrhabdina* sp. WS-2016, *Ancora sagittata*, *Monocystis agilis*, *Cephaloidophora cf. communis*, *Heliospora caprellae*, *Blabericola migrator*, *Protomagalhaensia* sp. Gyna, *Protomagalhaensia wolfi*, *Gregarina* sp. Poly, *Gregarina* sp. Pseudo (Janouškovec et al., 2019; Mathur et al., 2019, 2021b; Salomaki et al., 2021). Although they are indeed expected to be more resolving, the number of concerned species is even smaller and ambiguities remain in the interrelationships between groups, since the position of the genus *Cryptosporidium* is for example unstable between Janouškovec et al. (2019) and Mathur et al. (2019). Hovewer, a recent study used advanced alignment comparison methods to confront the different topologies currently proposed for apicomplexan phylogeny (Salomaki et al., 2021). The

team seems to achieve a congruent phylogeny of the two phylogenomic datasets; nevertheless, as the authors specify, many lineages exhibit long branches that the lack of taxa fails to resolve. We plainly concur with the authors, who conclude that a more accurate and stable picture of the evolutionary history of gregarines requires a drastic increase in sampling effort.

These genome/transcriptome studies have however also shown another interest: some species historically described as gregarines (*Platyproteum* spp., *Filipodium phascolosomae*, *Piridium sociabile* and *Digyalum oweni*) do not actually appear to be part of Apicomplexa anymore (Janouškovec et al., 2019; Mathur et al., 2019, 2021a). This suggests many taxonomic revisions to come concerning the multitude of species described as gregarines on the basis of similar morphological characters (extracellularity, gigantism), but which might in fact encompass polyphyletic lineages. Indeed, these studies have already established the multiple and independent origins of apicomplexan-like parasites (Janouškovec et al., 2019; Mathur et al., 2019).

1.3.2 Extrapolation based on host diversity and host-parasite behaviours

Concerning the second point, it is clearly documented that gregarines parasite virtually all non-vertebrate metazoan groups, from Polychaeta annelids to tunicates, arthropods and mollusks (Desportes and Schrével, 2013; Schrével and Desportes, 2015). For a long time, experts in the gregarine field have argued that, given the currently documented diversity of gregarines, their lifestyle principally monoxenous and their apparently narrow host-range specificity, the real biodiversity of gregarines is probably several orders of magnitude underestimated in particular for those infecting insects, that represent ∼half of metazoan diversity according to Mayhew (2007) (Desportes and Schrével, 2013; Schrével and Desportes, 2015).

In addition, most known hosts are infected by several gregarine species, as for example the mealworm *Tenebrio molitor* that is infected by at least three gregarine species (*Gregarina cuneata*, *Gregarina polymorpha* and *Gregarina steini*) (Clopton et al., 1992) in addi-

tion to *G. niphandrodes*. In consequences, some experts have even proposed that the real gregarine diversity should exceed that of their hosts, making it one of the most widespread groups of organisms in the environment (Desportes and Schrével, 2013; Schrével and Desportes, 2015).

1.3.3 Environmental metadata uncovering a wide and mostly undescribed diversity

Thirdly, several environmental and host-associated metagenomics/metabarcoding surveys have recently started to document this tremendous diversity at molecular level (de Vargas et al., 2015; Mahé et al., 2017; del Campo et al., 2019). A very wide diversity of apicomplexan-related sequences is present in terrestrial soils and marine sediments, and ∼80% of which appears more closely related to gregarines (Clopton et al., 1992; Mahé et al., 2017; del Campo et al., 2019).

This diversity may be paralleled to the biological cycles of gregarines for which cyst forms, enclosing each thousand to millions of oocyst progenies, are frequently released with the feces of their infected hosts, easily contaminating therefore such soils. In addition, the high resistance of these oocyst forms to environmental conditions certainly explains the high abundance and maintenance in environment of their enclosed genetic material (Desportes and Schrével, 2013; Schrével and Desportes, 2015). However, and even more remarkably, the diversity of apicomplexan-related sequences with gregarine-like affinities is also high in marine environments (both pelagic and benthic). This suggests the presence of either developing or resistant (oocyst) forms of these gregarines in association with planktonic elements - biological or even mineral - that remain to be identified or "freely" floating in these marine environments (de Vargas et al., 2015).

However, the flip side of these discoveries is that there is a remarkably high number of gregarine-like molecular data that cannot be related to formally described species (del Campo et al., 2019). This is in part due to the very low molecular knowledge we currently have of taxonomically and morphologically described species. Only talking about the most commonly used molecular marker, SSU rDNA, it is available in databases for only

about one hundred of the ∼1770 formally described gregarine species, which corresponds to just ∼5% of them, not taking into account the many gregarine species to be still discovered. We need to generate a much higher number of molecular markers for these known species. Furthermore, we must deploy appropriate strategies to morphologically and biologically describe the increasing number of "molecular-species" emerging from metagenomics studies, pointing to entirely novel phylogenetic groups within gregarines (de Vargas et al., 2015; Mahé et al., 2017; del Campo et al., 2019).

1.3.4 Missing data blur apicomplexan phylogeny

The missing information regarding gregarines is not restricted to this lack of connection between taxonomic data derived from morphological studies and molecular data emerging from metagenomic approaches. The confrontation of these data indicates that our understanding of the apicomplexan biodiversity remains very limited. Indeed, it remains biased because we have too long neglected the large spectrum of their pathogenicity focusing preferentially on the deadliest ones and ignoring the vast majority of poorly pathogenic ones (Rueckert et al., 2019a). This biased position has probably blurred not only our comprehension of the extent of their extraordinary host-adaptive capacities, but altogether a full section of their evolutionary history. So that we still do not know where the emergence of these species is taking root; is it within a group of non-pathogenic symbionts that has become even more diversified than we imagine, advocating for multiple emergence within a radiation that is still incompletely understood (Janouškovec et al., 2019; Kwong et al., 2019; Mathur et al., 2019; Rueckert et al., 2019b)?

A close examination of the recent taxonomy of Apicomplexa clearly mentions that most groups are currently polyphyletic or paraphyletic, notably: Aconoidasida, Coccidia, Gregarinasina, Archigregarinorida and Eugregarinorida (Adl et al., 2019). This is mostly due to lack of sufficiently informative and resolving data concerning these organisms, still mostly based on SSU rRNA phylogenies (Cavalier-Smith, 2014; Simdyanov et al., 2017, 2018) and recent emerging phylogenomic analyses (Janouškovec et al., 2019; Kwong et al., 2019; Mathur et al., 2019). Certainly, as stated above, insufficient sampling also prevents

a solid and integrated vision of phylogenetic inter-relationships for all of these species (del Campo et al., 2019; Janouškovec et al., 2019; Mathur et al., 2019; Salomaki et al., 2021).

1.4 The genomic panorama of Apicomplexa : a state of the art

Before going further, it is worth providing an precise overview of the currently documented apicomplexan genomes; although their sampling is biased in favor of intracellular and pathogenic parasites, these genomes are highly valuable since they are the only available references on which we can rely.

Currently, the available apicomplexan genomes are registered in the VEupathDB database which is dedicated to the inventory of eukaryotic pathogen genomic resources (Aurrecoechea et al., 2017).

At the beginning of this project in October 2018, 87 apicomplexan and related genomes were deposited in 4 subdatabases within VEupathDB (release 41):

- 16 on CryptoDB: 13 *Cryptosporidium* spp., 1 *Gregarina niphandrodes*, 1 *Vitrella brassicaformis*, 1 *Chromera velia*
- 32 on ToxoDB: 18 *Toxoplasma gondii* strains, 8 *Eimeria* spp., 2 *Sarcocystis neurona*, 1 *Neospora caninum*, 1 *Hammondia hammondi*, 1 *Cyclospora cayetanensis*, 1 *Cystoisospora suis*
- 9 on PiroplasmaDB: 4 *Babesia* spp., 4 *Theileria* spp., 1 *Cytauxzoon felis*
- 30 on PlasmoDB: 30 *Plasmodium* spp.

It should be noted that there is a significant redundancy in the sampling of these genomes, regarding the many strains of parasites of the species of the genus *Plasmodium* as well as *Toxoplasma gondii*, which are the most pathogenic for humans and therefore the most studied. These are also the species that can be cultured in the laboratory, facilitating

Species & Strain	**Genome release**	**Publication**	**Nb of proteins**	**Nb of contigs**	**Total length (Mb)**	**%GC**
Cryptosporidium hominis 30976	2017-11-16	Guo et al. (2015)	3949	53	9	30.13
Cryptosporidium meleagridis UKMEL1	2018-02-02	Ifeonu et al. (2016)	3753	57	8.9	30.97
Cryptosporidium muris RN66	11-03-2015	-	3938	75	9.2	28.47
Cryptosporidium parvum Iowa II	2007-02-26	Abrahamsen et al. (2004)	3941	8	9.1	30.22
Chromera velia CCMP2878	2014-05-15	Woo et al. (2015)	30604	5953	193.8	49.11
Vitrella brassicaformis CCMP3155	2014-05-15	Woo et al. (2015)	23412	1064	72.7	58.09
*Gregarina niphandrodes*Unknown	2014-04-15	-	6375	468	14	53.78
*Cyclospora cayetanensis*CHN_N01	2016-09-22	Liu et al. (2016)	7455	2297	44	51.84
Cystoisospora suis Wien I	2017-10-20	Palmieri et al. (2017)	11543	7880	81	49.32
Eimeria falciformis Bayer Haberkorn 1970	2014-02-10	Heitlinger et al. (2014)	5876	753	43.6	49.86
Eimeria tenella Houghton	2013-11-05	Reid et al. (2014)	8597	4664	51.8	51.33
Hammondia hammondi HH34	2014-06-30	-	8004	3676	64	52.83
Neospora caninum LIV	2015-02-27	Reid et al. (2012)	7122	66	59	54.82
Sarcocystis neurona SN3	2015-04-13	-	6965	873	124	51.41
Toxoplasma gondii ME49	2015-08-11	-	8322	2075	65.5	52.30
Babesia bovis T2Bo	2010-03-10	Brayton et al. (2007)	3706	14	8	41.59
Babesia microti RI	2017-07-12	Cornillot et al. (2012)	3601	6	6.4	36.17
Babesia ovata Miyake	2017-12-16	Yamagishi et al. (2017)	5031	91	14.4	49.27
Theileria equi WA	2014-08-21	Kappmeyer et al. (2012)	5332	12	11.6	39.48
Theileria orientalis Shintoku	2012-09-06	Hayashida et al. (2012)	4002	6	9	41.55
Theileria parva Muguga	2010-03-10	Gardner (2005)	4082	10	8.3	34.04
Cytauxzoon felis Winnie	2015-01-20	Tarigo et al. (2013)	4323	357	9.1	31.81
Plasmodium berghei ANKA	2017-01-09	Otto et al. (2014)	5067	21	18.7	22.04
*Plasmodium falciparum*3D7	2015-06-18	Gardner et al. (2002)	5460	16	23.3	19.34
Plasmodium vivax P01	2018-02-28	Auburn et al. (2016)	6677	242	29	39.78

Table 1.1: **Metrics and publications of the selected reference genomes**. Metrics were calculated with QUAST (Gurevich et al., 2013).

both the sequencing of their genomes and functional genetic studies (Limenitakis and Soldati-Favre, 2011).

In order to build a set of genomes that can serve as references for the assembly of gregarine genomes, but also in a comparative genomics perspective, for data-mining of their deduced proteomes, 25 genomes representative of the currently known diversity of apicomplexan were selected from the available genomes. The selection was made taking into account the most recent data and techniques used as well as the presence of an associated publication, in order to have the most complete panorama of apicomplexan proteins and key functions/structures currently documented and suitable to be used as primers to search for their homologs within the gregarine genomes. The metrics of all available genomes were calculated with QUAST (Gurevich et al., 2013) from contigs deposited on VEupathDB (release 41). The genomes as well as their metrics and information are gathered in Table 1.1, page 32 and illustrated in Figure 1.2, page 14; the selection is as follows:

- 4 genomes for the Cryptosporidia (*Cryptosporidium hominis* 30976, *C. meleagridis* UKMEL1, *C. muris* RN66 and *C. parvum* Iowa II)
- 1 genome for *Gregarina niphandrodes* Unknown
- 8 genomes for Coccidia (*Toxoplasma gondii* ME49, *Cystoisospora suis* Wien I, *Cyclospora cayetanensis* CHN_N01, *Sarcocystis neurona* SN3, *Hammondia hammondi* HH34, *Neospora caninum* LIV, *Eimeria falciformis* Bayer Haberkorn 1970 and *E. tenella* Houghton)
- 3 genomes for Hemosporidia (*Plasmodium falciparum* 3D7, *P. berghei* ANKA and *P. vivax* P01)
- 7 genomes for Piroplasma (*Theileria equi* WA, *T. orientalis* Shintoku, *T. parva* Mugaga, *Babesia bovis* T2Bo, *B. microti* RI, *B. ovata* Miyake and *Cytauxzoon felis* Winnie)
- 2 genomes for Chromerida (*Chromera velia* CCMP2878 and *Vitrella brassicaformis*

CCMP3155)

At present, the only available genomic data of gregarines concern terrestrial eugregarines:

- *Gregarina niphandrodes* - unpublished genome, deposited on VEupathDB and GenBank (accession number: AFNH00000000.2)
- *Ascogregarina taiwanensis* (Templeton et al., 2010), very partial genome, not annotated and deposited on GenBank (accession number: PRJNA27765)

The *Ascogregarina taiwenensis genome* is unfortunately very partial and from a very low coverage assembly (<0.2X, source: GenBank Bioproject PRJNA27765). For these reasons, we did not retain it in our analyses.

The main problem with the *Gregarina niphandrodes* genome is that there is no publication associated with it; on the other hand, it is the only other existing reference for a full-scale genome of a gregarine. Therefore, from a scientific perspective, we could not reasonably disregard these data, but we remained cautious in their use, as it is currently impossible to evaluate their completeness and quality.

In parallel, the inventory of currently known organelle genomes was also performed, for mitochondrial and apicoplastic genomes, and their metrics were again calculated with QUAST (Table 1.2, page 35 and Table 1.3, page 36).

Species/Strain	**Nb contigs**	**Length (Mb)**	**%GC**	**N50**	**N's per 100 kbp**
Babesia bovis	1	6005	29.51	6005	0.00
Babesia microti	1	10547	35.18	10547	0.00
Theileria equii	1	9001	29.17	9001	0.00
Theileria orientalis	1	2595	31.25	2595	0.00
Theileria parva	1	5895	29.99	5895	0.00
Plasmodium berghei	1	5957	30.94	5957	0.00
Plasmodium cynomolgi	1	6017	30.28	6017	0.00
Plasmodium falciparum 3D7	1	5967	31.59	5967	0.00
Plasmodium falciparum IT	1	6616	31.65	6616	755.74
Plasmodium gallinaceum	1	6747	32.58	6747	0.00
Plasmodium knowlesi MalayanPk1A	1	3833	32.19	3833	0.00
Plasmodium malariae UG01	1	5969	29.87	5969	0.00
Plasmodium reichenowi CDC	1	5966	31.63	5966	0.00
Plasmodium relictum SGS1-like	1	6092	31.68	6092	0.00
Plasmodium vivax P01	1	5989	30.52	5989	0.00
Plasmodium vivax Sal1	1	5990	30.50	5990	0.00
Plasmodium yoelii yoelii 17X	1	6083	31.27	6083	0.00
Eimeria falciformis BayerHaberkorn1970	1	6280	34.49	6280	0.00

Table 1.2: **Metrics of available mitochondrial genomes**. Metrics were calculated with QUAST (Gurevich et al., 2013).

Species/Strain	**Nb contigs**	**Length (Mb)**	**%GC**	**N50**	**N's per 100 kbp**
Babesia bovis	1	35107	22.02	35107	0.00
Babesia microti	1	28657	14.06	28657	0.00
Theileria equi	1	47880	29.02	47880	0.00
Theileria orientalis	1	24173	19.48	24173	14892.65
Theileria parva	1	39579	19.48	39579	5.05
Plasmodium berghei ANKA	1	34403	15.11	34403	0.00
Plasmodium cynomolgi	1	34521	14.24	34521	0.00
Plasmodium falciparum 3D7	1	34250	14.22	34250	0.00
Plasmodium falciparum IT	1	29686	13.32	29686	168.43
Plasmodium fragile Nilgiri	1	35616	13.09	35616	18688.23
Plasmodium gaboni SY75	1	29387	13.03	29387	340.29
Plasmodium gallinaceum 8A	1	29456	12.90	29456	0.00
Plasmodium knowlesi H	1	30638	14.03	30638	0.00
Plasmodium knowlesi MalayanPk1A	1	32097	14.08	32097	0.00
Plasmodium malariae UG01	1	34324	13.70	34324	0.00
Plasmodium ovale curtisi GH01	1	28304	12.46	28304	215.52
Plasmodium reichenowi CDC	1	29226	13.04	29226	171.08
Plasmodium relictum SGS1-like	1	29365	13.06	29365	0.00
Plasmodium vivax P01	1	29582	13.30	29582	338.04
Plasmodium yoelii yoelii 17X	1	34324	15.02	34324	0.00
Eimeria falciformis BayerHaberkorn1970	1	33174	22.97	33174	6963.28
Hammondia hammondi HH34	1	29684	17.15	29684	25006.74
Sarcocystis neurona SN3	1	35004	21.97	35004	0.00
Toxoplasma gondii ARI	11	20862	19.36	1975	0.00
Toxoplasma gondii FOU	2	29760	19.28	24410	0.00
Toxoplasma gondii GAB2-2007-GAL-DOM2	13	10789	15.31	758	0.00
Toxoplasma gondii GT1	15	21338	21.47	1408	0.00
Toxoplasma gondii MAS	3	29272	19.06	24390	0.00
Toxoplasma gondii ME49	2	36351	20.82	35372	6533.52
Toxoplasma gondii p89	3	29490	19.20	24386	0.00
Toxoplasma gondii pRUB	5	29279	19.11	24358	0.00
Toxoplasma gondii TgCATBr9	2	29725	19.27	24364	0.00
Toxoplasma gondii TgCatPRC2	4	30392	19.29	18165	0.00
Toxoplasma gondii VAND	4	29731	19.28	24393	0.00
Toxoplasma gondii VEG	19	45270	19.78	2688	0.00

Table 1.3: **Metrics of available apicoplast genomes**. Metrics were calculated with QUAST (Gurevich et al., 2013).

1.5 Aims and Objectives

With just above a hundred genomes deciphered for ∼350 genera and ∼6000 described species, Apicomplexa is a group for which there is still a lot to discover. It is certainly not the poorest documented branch of the tree of life (Sibbald and Archibald, 2017), but it is far from having revealed all the secrets of the diversity of its molecular innovations, developed during its evolutionary history as it had to adapt to such a wide diversity of hosts and environments (Desportes and Schrével, 2013; Schrével and Desportes, 2015).

As to date, molecular exploration of apicomplexan parasites have mainly concerned a very small number of species that have in common:

1) to infect humans causing threatening and global diseases such as malaria or less threatening but worldwide spread diseases such as toxoplasmosis and cryptosporidiosis;

2) to be cultivable in the laboratory, at least for some developmental stages;

3) to have been the subject of extremely sophisticated methodological developments such as genetic manipulation, -omics in all their variations and static and dynamic microscopy (Limenitakis and Soldati-Favre, 2011).

In this panorama, gregarines, full members of the Apicomplexa phylum, have been so far left on the side of the road for exactly corollary reasons:

1) they do not infect humans;

2) they are not easy to cultivate;

3) while there is a very abundant literature of their life cycles, morphologies and ultrastructure, they are almost unknown at the genomic/transcriptomic levels and have been the subject of very few biochemical studies.

But their future is now open for exploration, as the stage is set for the emergence of genomic data. A first move should be in favour of known species, selected either for

their biological characteristics (intestinal, coelomic, motile, non-motile) or their particular phylogenetic position (archigregarines vs eugregarines as a broad distinction). As previously stated, the main obstacle for the production of gregarine genomes is the difficulty to collect enough biological material for genomic sequencing. Indeed, no species of gregarine is currently cultivable in the laboratory. There are two possibilities to overcome these problems: the first is to maintain hosts infected with gregarines under laboratory conditions. While this solution does not solve the problem of contamination by the host and surrounding microorganisms during the collection, it does guarantee regular access to the targeted developmental forms and thus enables the acquisition of an adequate amount of biological material. The second solution involves identifying biological models that are capable, through their inherent characteristics, of providing sufficient biological material. The next chapter of this thesis will demonstrate how we were able to exploit both of these solutions that results in the selection of three biological models: 2 marine gregarines, *Porospora gigantea*, parasite of the European lobster *Homarus gammarus* and *Diplauxis hatti*, parasite of the Polychaeta marine worm *Perinereis cultrifera*; and 1 terrestrial gregarine, *Gregarina acridiorum*, parasite of the locust *Locusta migratoria*.

The collection of sufficient biological material is only the first of the challenges raised by the genomics of non-model, non-culturable organisms. Indeed, in the absence of data close enough to serve as reference, the removal of contaminants as well as the prediction of genes from genomic assemblies are other obstacles. I have addressed the issues raised by each of the gregarines for which we were able to produce genomic data differently, by tailoring to their specific situation. Again, these protocols and their results will be outlined and discussed in the following chapter.

Another issue of concern is that the currently documented diversity of gregarines is lagging, by probably several orders of magnitude, far beyond the true diversity of these organisms in open and host-associated environments (Desportes and Schrével, 2013; Schrével and Desportes, 2015; Mahé et al., 2017; del Campo et al., 2019; Mathur et al., 2019). As mentioned earlier, there are several examples of taxonomic revisions following the introduction of molecular methods to complement traditional morphological approaches

(Cavalier-Smith, 2014; Schrével et al., 2016; Rueckert and Horák, 2017; Simdyanov et al., 2017, 2018). In this context, it is particularly difficult to estimate the extent of sampling needed to properly document the full diversity of gregarines. We were indeed confronted with two examples illustrating the magnitude of the upcoming taxonomic revisions concerning gregarines, and the need to turn to molecular markers, and likely on a genomic scale, to properly assess the diversity of these organisms.

In a first case, we studied the gregarines infecting two species of locusts, *Schistocerta gregaria* and *Locusta migrotaria*. The associated morphological descriptions and the existing bibliography, both of which were extensive and conflicting, did not permit to decide on the identity of the infecting species. Through a combined analysis of morphology and molecular phylogeny, we were able to accurately determine the presence of two distinct species each infecting one or the other locust - *Gregarina garnhami* infecting *Schistocerca gregaria* while *Gregarina acridiorum* infects *Locusta migratoria* (Florent et al., 2021).

The second example derives from the genomic data generated for *Porospora gigantea*, a marine gregarine infecting the European lobster *Homarus gammarus*. Indeed, as detailed in the second chapter, we have highlighted the presence of two genomes in samples taken from some of the lobsters. Based on this discovery, we have carried out an in-depth study of these genomes in order to determine the taxonomical status of these gregarines. Here again, molecular data proved to be essential, as morphological data did not allow us to distinguish the two taxa. After a careful review of the current standards and a comparative analysis of all available data, we were able to demonstrate the co-infection of European lobsters by two species of gregarines of the genus *Porospora*, named *P. cf. gigantea* A and *P. cf. gigantea B*. The third chapter of this thesis is dedicated to those two complementary examples, in which I demonstrate the contribution of molecular data in the species delimitation and biodiversity assessment of gregarines, as well as the perspectives on the future of gregarine taxonomy and the available means to update it.

Gregarines' diversity is also likely to reveal even more original and unknown adaptive mechanisms for this fascinating group, the Apicomplexa. Gregarine genomes have the potential to teach us novel adaptive aspects of the group (relative to intracellular parasites

of vertebrates), particularly in relation to their extracellular mode of living and the associated specific constraints to survival capacities. In this perspective, the fourth and last chapter of the thesis is devoted to the first comparative genomic analyses of the two annotated genomes generated during this PhD, i.e. the genomes of *P. cf. gigantea* A and *P. cf. gigantea* B. There we present the main structural characteristics of these genomes, and compare their deduced proteome with a selection of apicomplexan proteomes, thus highlighting the unsuspected diversity of the gregarine gene pool. Indeed, the identification of multiple proteins specific to gregarines that are totally absent from other apicomplexan lineages supports and describes the extent of the still unknown molecular diversity of this group.

The study of gregarines can also provide a better understanding of the adaptive mechanisms set up by apicomplexan parasites. The study of certain key functions or structures, as presented above, will undoubtedly allow a better understanding of the range of possible adaptations and of the genetic inheritance of the common ancestor of all apicomplexans. I chose to focus on the glideosome, a complex molecular structure responsible for gliding, a specific and emblematic form of motility in the Apicomplexa, which is essential to the expression of their pathogenicity. Expert annotation of the glideosome proteins revealed their differential conservation at the apicomplexan scale (i.e. some functional groups fully conserved and others partially retained), suggesting a diversity of adaptations to the challenges of motility and host cell invasion during the evolutionary history of the group.

Finally, I'd like to state that this thesis manuscript is essentially derived from the three articles published or submitted during my PhD: Boisard and Florent (2020); Florent et al. (2021); Boisard et al. (submitted), and therefore must by considered as a thesis by publication. Nevertheless, in order to highlight my own contributions to each of them, and for the sake of fluidity and ease of reading, they have been reformatted, completed and arranged in a structure more suitable for a thesis manuscript. The three papers are reproduced in their entirety at the end of the volume.

Therefore, this introductive first chapter was partly based on the review I co-authored with I. Florent (mainly parts 1.1, 1.2 and 1.3) (Boisard and Florent, 2020), and ex-

tended by an inventory of available apicomplexan genomes and the selection of 25 references among them, which I used to conduct the comparative analyses presented in this manuscript. Finally, this last part exposed, on the basis of the established observations, the objectives and the means implemented during this thesis to address the outlined problematics.

Chapter 2, on the other hand, consists mainly of unpublished data concerning the genomes of the coelomic marine eugregarine *Diplauxis hatti*, and of the intestinal terrestrial eugregarine *Gregarina acridiorum*. The parts from Chapter 2 devoted to *Porospora gigantea* genomes are partially derived from Boisard et al. (submitted); however, I made a point of detailing the protocols implemented in order to provide a critical look at the challenges imposed by the genomics of non-model and non-cultivable organisms.

Chapter 3 is based on the phylogenetic analyses published in Florent et al. (2021) and submitted in Boisard et al. (submitted), to illustrate the contribution of molecular analyses to species delimitation, and their absolute requirement to document the full extent of gregarine biodiversity.

Finally, Chapter 4 is for the most part grounded in the submitted paper Boisard et al. Here, I was willing to highlight and elaborate on my own work concerning comparative genomics analyses. I have chosen to develop further these analyses, both in their methodology and in the resulting discussions. Indeed, in order to satisfy the current editorial constraints of the journal to which this work was submitted, some topics have not been developed in the article. It seemed important to me to take advantage of this thesis manuscript to present them more thoroughly.

This thesis is part of a larger research project conducted by Isabelle Florent at the Muséum national d'Histoire naturelle, aimed at studying gregarines to further explore apicomplexan evolutionary history. It involves numerous collaborators, some of whom are co-authors of the articles reproduced on pages 165, 178 and 193. I have chosen, in agreement with my supervisors, to explicitly mention the work done by the collaborators while refering the readers to the original articles. Whenever it was necessary to reproduce some results relevant to the understanding of my thesis work, I explicitly mentioned the

collaborators' contribution at the beginning of the concerned sections.

Chapter 2

The challenge of deciphering genomes for non-model and non-cultivable species

2.1 Selecting suitable biological models

The MCAM laboratory undertook the sequencing of the genomic DNA of gregarines in 2017, relying on a unique expertise at the Muséum national d'Histoire naturelle through the presence of two worldwide experts on gregarines, Pr J. Schrével and Dr. I. Desportes.

The challenge was both methodological and conceptual: to isolate biological material in sufficient quantity and quality and to develop specific bioinformatics methods to reconstruct these genomes entirely *de novo*, in the absence of data on species close enough to serve as references. As explained in the previous chapter, there are two possibilities to overcome the un-cultivability of gregarines: the first is to maintain hosts infected with gregarines under laboratory conditions, while the second solution is to identify biological models with developmental forms that can provide enough biological material to allow adequate genomic DNA isolation.

In 2016, during a campaign in Roscoff (Brittany, France), a marine gregarine caught the attention of I. Florent and J. Schrével. The species in question is *Porospora gigantea*

Van Beneden, 1869, a parasite of the European lobster *Homarus gammarus*. According to the literature, this gregarine has particular cystic forms located in the rectal ampulla of these crustaceans and specific to the Porosporidae family (Hatt, 1931; De Bauchamp, 1910). These were then observed and sampled: they contained thousands of gymnospores, consisting of monolayers of "naked" zoites (without envelope) (Figure 3.1, page 83 and Figure 3.2, page 84). These cystic forms appeared an exceptional source of biological material for genomic DNA isolation to be sequenced, providing millions of copies of the genome for each cyst and a low risk of contamination by the host or surrounding microorganisms.

Access since 2012 to the MNHN vivarium breeding of a species of locust, *Locusta migratoria*, infected with a terrestrial gregarine, *Gregarina acridiorum*, allowed isolation of sufficient biological material for genomic sequencing in 2017.

Finally, genomic data for a second marine gregarine, *Diplauxis hatti*, a coelimic gregarine of the polychaete annelid *Perinereis cultrifera*, were acquired by the lab in late 2018. *Diplauxis hatti* has a unique life cycle adaptation to its host. Indeed, observations on natural populations in the English Channel have shown that the release of parasite spores is concomitant with polychaete spawning (Prensier et al., 2008). Thus, by collecting hosts during their breeding season, late March or early April, J. Schrével, G. Prensier and L. Guillou were able to collect *Diplauxis hatti* cysts. Detailled informations on the life cycle of *D. hatti* is provided in Prensier et al. (2008).

It is thus the original genomic data for 3 gregarines that have been studied during this PhD: 2 marine gregarines, *Porospora gigantea*, parasite of the European lobster *Homarus gammarus* and *Diplauxis hatti*, parasite of the marine worm *Perinereis cultrifera*; and 1 terrestrial gregarine, *Gregarina acridiorum*, parasite of the locust *Locusta migratoria*.

2.2 Methods

2.2.1 Biological sampling

Collection of hosts, isolation of parasites and morphological studies on *P. cf. gigantea* was carried out by I. Florent, J. Schrével and L. Guillou. See detailled methods in Boisard et al. (submitted). *Homarus gammarus* lobster were collected in the British Channel at Roscoff (Britany, France) between July 2015 and October 2017, either directly from the field (Roscoff bay) or through lobster tanks facilities (lobster provenance: South of England), in which crustaceans are maintained in captivity several weeks to months before their commercialization.

Concerning *Porospora gigantea*, 4 isolates were made and DNA was extracted from cysts collected in the rectal ampulla of the host *Homarus gammarus*. Detailled methods are available in Boisard et al. (submitted).

- JS-470, from Lobster #7 in Roscoff Lobster tank facility, 70 cysts, 50µL (41.4 ng/µL)
- JS-482, from Lobster #11 caught in Roscoff Bay, 50 cysts, 50µL (19.8 ng/µL)
- JS-488, from Lobster #12 caught in Roscoff Bay, 100 cysts, 50µL (44.8 ng/µL)
- JS-489, from Lobster #12 caught in Roscoff Bay, 100 cysts, 50µL (66.6 ng/µL)

RNA was also isolated from 2 additional biological samples, both composed of cysts' pools taken from the rectal ampulla of their respective host specimens:

- JS-555, from Lobster #26 caught in Roscoff bay, 35 cysts, 55µL (2.81ng/µL)
- JS-575c, from Lobster #34 in Roscoff Lobster tank facility, 40 cysts, 55µL (0.92ng/µL)

Concerning *G. acridiorum*, 1 isolate was used on 27/02/2015 by A. Labat. DNA was extracted from the cyst pools collected in the intestines and feces of the host *Locusta migratoria*; using MasterPureTM Complete DNA and RNAPurification kit (Epicentre, Illumina Inc. USA) with a yield of:

- JS-313, 32 cysts, 50µL (20ng/µL)

Concerning *D. hatti*, 5 isolates were made on 30/04/2018 by J. Schrével, G. Prensier et L. Guillou, each containing 100 cysts. DNA was extracted from two cyst pools collected in the hosts' coelome; using Macherey Nagel Tissue and Cells isolation kit (ref 740952.50) with a yield of respectively:

- JS 626a, 100 cysts, 50µL(63.8ng/µL)
- JS 627a, 100 cysts, 50µL (35.6ng/µL)

2.2.2 DNA/RNA sequencing and assembly

All DNA isolates were sequenced individually by using Illumina NextSeq technology (2*151bp; NextSeq 500 Mid Output Kit v2; Institut du Cerveau et de la Moelle Epinière - CHU Pitié-Salpêtrière - Paris).

Genomic assemblies were performed by E. Duvernois-Berthet (*P. gigantea*, *G. acridiorum*) and myself (*D. hatti*).

Reads were cleaned by using Trim Galore (version 0.4.4) (Krueger, 2015) removing remnant Nextera adaptors, clipping 15 bp in 5'-end and 1 bp in 3'-end and trimming low-quality ends (phred score < 30). For each species' isolate, reads were assembled with SPAdes (version 3.9.1; options: careful mode, automatic k-mers), first independently for each library, and then with pooled libraries if applicable. (Bankevich et al., 2012). See later the detailed workflow for *P. gigantea*: Figure 2.1, page 53; for *G. acridiorum*: Figure 2.5, page 60 and for *D. hatti*: Figure 2.6, page 61.

RNA isolates were sequenced individually by using NextSeq technology (librairy preparation: SMART-Seq v4 Ultra Low Input RNA Kit from Takara; 2*75bp; NextSeq 500 Mid Output Kit v2; Institut du Cerveau et de la Moelle – CHU Pitié Salpétrière - Paris). Reads were cleaned by using Trim Galore to remove remnant Nextera adaptors, clipping 15 bp in 5'-end and 1 bp in 3'-end and trimming low-quality ends (phred score < 30). The sequence reads of both samples were merged into one library which was assembled using Trinity (Haas et al., 2013).

2.2.3 Extraction of putative apicomplexan sequences

I performed the extraction of putative apicomplexan contigs using a method developed by L. Ponger. The same method was applied for the 3 assemblies, i.e. *P. gigantea*, *G. acridiorum* and *D. hatti*. All genomic contigs longer than 1kb were analyzed by using a principal component analysis (PCA) based on their 5-mer composition, which allowed classifying them into n groups by using a hierarchical clustering method (HCA) based on the Ward criterion (see for *P. gigantea*: Figure 2.1, page 53; for *G. acridiorum*: Figure 2.5, page 60 and for *D. hatti*: Figure 2.6, page 61).

Concerning *P. gigantea*: for all contigs, the putative protein coding genes were then predicted by using Augustus (version 3.3) (Stanke et al., 2006) and the only gene model natively implemented for an Apicomplexa: *T. gondii*.

Concerning *G. acridiorum*: for all contigs, the putative protein coding genes were then predicted by using Augustus, first using *T. gondii* model and then the gene model of *G. niphandrodes* that we built from the genome annotation available on VEupathDB (see below section 2.2.5, page 49).

Concerning *D. hatti*: for all contigs, the putative protein coding genes were then predicted by using Augustus, first using *T. gondii* model and then the gene model of *Cryptosporidium hominis* that we built from the genome annotation available on VEupathDB (see below section 2.2.5, page 49).

In all three cases, the predicted proteins were then compared with the NCBI non-redondant protein database (NR) by using BLASTP (Altschul et al., 1990). The analysis of the taxonomic groups associated to the corresponding best hits, enabled us to identify putative bacterial/fungi contaminants clusters as well as the clusters gathering sequences from closely related apicomplexan organisms (see for *P. gigantea*: Figure 2.1, page 53; for *G. acridiorum*: Figure 2.5, page 60 and for *D. hatti*: Figure 2.6, page 61).

2.2.4 *P. gigantea*: Identification of genomes A and B

The identification of the A and B genomes was achieved through a multiple step process, together by E. Duvernois-Berthet, L. Ponger and myself.

Metrics were calculated for the assemblies of the four independent libraries with QUAST (Gurevich et al., 2013) and revealed an apparent redundancy of the data. The first gene predictions were carried out on these 4 independent assemblies using *T. gondii* gene model natively implemented in Augustus; they also showed a duplication of the predicted proteins in the 3 assemblies from lobsters caught in Roscoff Bay, suggesting the presence of one genome in the "Roscoff tank facility" librairy (JS-470) while two similar genomes were present in the "Roscoff Bay" librairies (JS-482, JS-488 and JS-489).

A difference of coverage was indeed observed for each of the 4 gDNA libraries (Figure 2.2, page 54). Each gDNA library (JS-470, JS-482, JS-488 and JS-489) was individually mapped on the contigs by using Bowtie2 and the coverages' medians were calculated for each contig and each library by using the Samtools (Li et al., 2009) and the Bedtools (Quinlan and Hall, 2010) libraries. Finally, the contigs of the "apicomplexan" cluster from the final pooled assembly were splitted into genomes A and B using this difference of coverage as a discriminant (analyses done by E. Duvernois-Berthet).

This coverage information was processed with a principal component analysis and a k-means algorithm which allowed classifying the contigs into 2 clusters. Then, a linear discriminant model was trained with the coverage information and the result of this first classification; it was then be applied to all the contigs in order to improve the classification. The linear discriminant method (training and classification) was iteratively repeated 3 times until convergence. A similar analysis was carried out with 1kb non-overlapping windows (instead of full length contigs) to identify some putative hybrid contigs. Then, contigs classified to different genomes (depending on the windows) were divided into sub-contigs which were re-assigned to their respective genomes (analyses done by L. Ponger, see Figure 2.3.C, page 56).

The nucleic divergence between genome A and genome B was calculated by L. Ponger following methods described in Boisard et al. (submitted).

2.2.5 Prediction of coding genes

RNAseq model for *P. gigantea*

All *de novo* assembled transcripts were aligned against the "apicomplexan" cluster contigs with GMAP (Wu and Watanabe, 2005) within the PASA program (Haas, 2003).

Then, two *ab initio* gene prediction tools, SNAP version 2017-11-15 (Korf, 2004)) 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 construction of hints from introns (allowing small introns length >10bp) using our "apicomplexan" cluster repeat-masked genome assembly as reference using RepeatMasker 4.0.8 (Smit et al., 2015)). Gene predictions were then performed allowing the prediction of alternative transcripts and non-canonical intron bounds. An alternative model was also trained with SNAP (default protocol) and used for gene predictions.

The Augustus and SNAP outputs having sometimes predicted genes slightly differently, the predictions were then parsed by a home-made script (by L. Ponger) in order to keep, for each prediction made, as many alternative genes and transcripts as possible.

Exonerate gene model for *G. acridiorum* and *D. hatti*

The Exonerate (Slater and Birney, 2005) software was used to produce two gene models from the genomic contigs of *Gregarina acridiorum* and *Diplauxis hatti*, for which we currently have no RNA data.

Gene models from public available apicomplexan genomes

Using a selection of reference apicomplexan genomes available in VEupathDB (release 41) (Aurrecoechea et al., 2017), Augustus was trained from each genome with a subset of its own genes. Thus, several gene models were constructed and implemented in Augustus so that comparative genetic predictions can be conducted.

Comparative gene models analyses

Comparative analyses of gene prediction metrics from the different available or re-created gene models were performed with QUAST (Gurevich et al., 2013) or using home-made scripts. The completeness of all gene predictions performed were assessed by using BUSCO version 4.0.6 (Seppey et al., 2019).

2.2.6 Removal of contaminants in both *P. gigantea* genomes

Host contamination

All the "apicomplexan" cluster contigs were screened by using the short reads available for the genome sequencing project of the closely related *Homarus americanus* species (PRJNA486050) in order to identify host contaminants. This Lobster dataset was supposed to be free of sequences from apicomplexan species, since it has been obtained from DNA extracted from the non-intestinal tissues (the tail, the leg or the pleiopod appendices). The mapping was carried out with Bowtie2 (Langmead and Salzberg, 2012) and the coverages were calculated by using the Samtools library (Li et al., 2009). Contigs thus identified that were covered for more than 60% of their length by the Homarus' short reads were considered host contaminants and were removed (analyse done by L. Ponger; see Figure 2.3.A, page 2.3).

Prokaryotes contaminants

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

Isolate Identifier	Nb of contigs >1 Kb	Length (Mb)	%GC	N50 (Kb)	Host Origin
JS-470 *Lobster #7*	938	9.7	51.6	38	Roscoff Tank facility
JS-482 *Lobster #11*	8900	31	48	2.3	caught in Roscoff Bay
JS-488 *Lobster #12*	5001	27	47.5	12	caught in Roscoff Bay
JS-489 *Lobster #12*	4025	25	48.5	18.5	caught in Roscoff Bay

Table 2.1: **Characteristics of the 4 assemblies from the 4 isolates of *Porospora gigantea* cysts.** The N50 value indicates the scaffold size for which 50% of the scaffolds are smaller or larger.

were removed from the genome assembly. See Figure 2.3.A, page 2.3.

2.3 *P. cf. gigantea*

2.3.1 Identification and characterization of 2 genomes

As soon as the first analyses of the genomic assemblies from the 4 sequenced biological isolates were performed, it became evident that the processing of the genomic data of *Porospora gigantea* would be much more challenging than expected. Indeed, the first analyses revealed a redundancy of the majority of the data in 3 assemblages, whose origin had to be understood (Table 2.1, page 51).

The metrics we calculated (notably the genome size and the number of predicted proteins) revealed a difference between the datasets (Table 2.1). Indeed, the JS-482, JS-488 and JS-489 assemblies displayed a two- to three-times greater total length of assembled nucleotides and a two- to four-times greater amount of predicted proteins, compared to the assembly obtained with JS-470.

In order to extract the apicomplexan sequences from the environmental contaminants, we carried out a demultiplexing based on the k-mers composition of the pooled assembly (Figure 2.1, page 53). The frequency of k-mers was used here as a signature of the subsequences occurring in the genome assembly. The comparison of these frequencies allowed to discriminate patterns within the contigs of each assembly, whose distribution

were then visualized using descriptive statistics. Principal component analysis (PCA) is a multivariate statistical method that allows to transform correlated variables into new variables that are decorrelated from each other (i.e. the principal components). PCA helps to summarize the data by pointing out the variables that best explain the variance in the data. The hierarchical cluster analysis (HCA) enables the regroupment of individuals from a given set into different classes. The results of the PCA were here used as a measure of dissimilarity to determine the different classes that were then vizualised on the PCA in different colors (Figure 2.1, page 53).

We thus performed a PCA and then a HCA which highlighted 6 clusters on the scatter plot, corresponding to 6 groups of contigs within our assembly. We predicted genes on each of these clusters using the Augustus natively implemented gene model of *T.gondii*, and carried out BLASTP searches to identify the taxa associated with the best hits. The analysis of the taxonomic groups associated to the corresponding best hits, enabled us to identify five clusters as putative bacterial contaminants whereas the sixth cluster which included 1745 contigs (18.0Mb), was identified as organisms closely related to Apicomplexa, referred as "apicomplexan" cluster later (see Figure 2.1, page 53 as well as Figure 2.3, page 56 to get an overview of the whole protocol). Furthermore, the "apicomplexan" cluster, when submitted to a similarity search against itself, revealed the presence of many contigs in two similar copies, within the assemblies JS-482, JS-488 and JS-489; on the other hand, the JS-470 assembly did not displayed this redundancy pattern.

We calculated the median coverage of the cluster 6 contigs for each of the libraries (Figure 2.2, page 54). It is expected that the contigs of a same species are covered homogenously, with a modal distribution. However, the analysis of contigs coverage by each individual library revealed a bimodal distribution in three librairies, suggesting a mixture of genomes with a proportion depending on the biological sample (Figure 2.2). More precisely, while only one set of contigs displayed a significant coverage for the lobster tank parasite sample (JS-470, peak around 250X), the three other parasite samples, from freshly captured hosts (JS-482, JS-488, JS-489) showed two distinct sets of contigs with

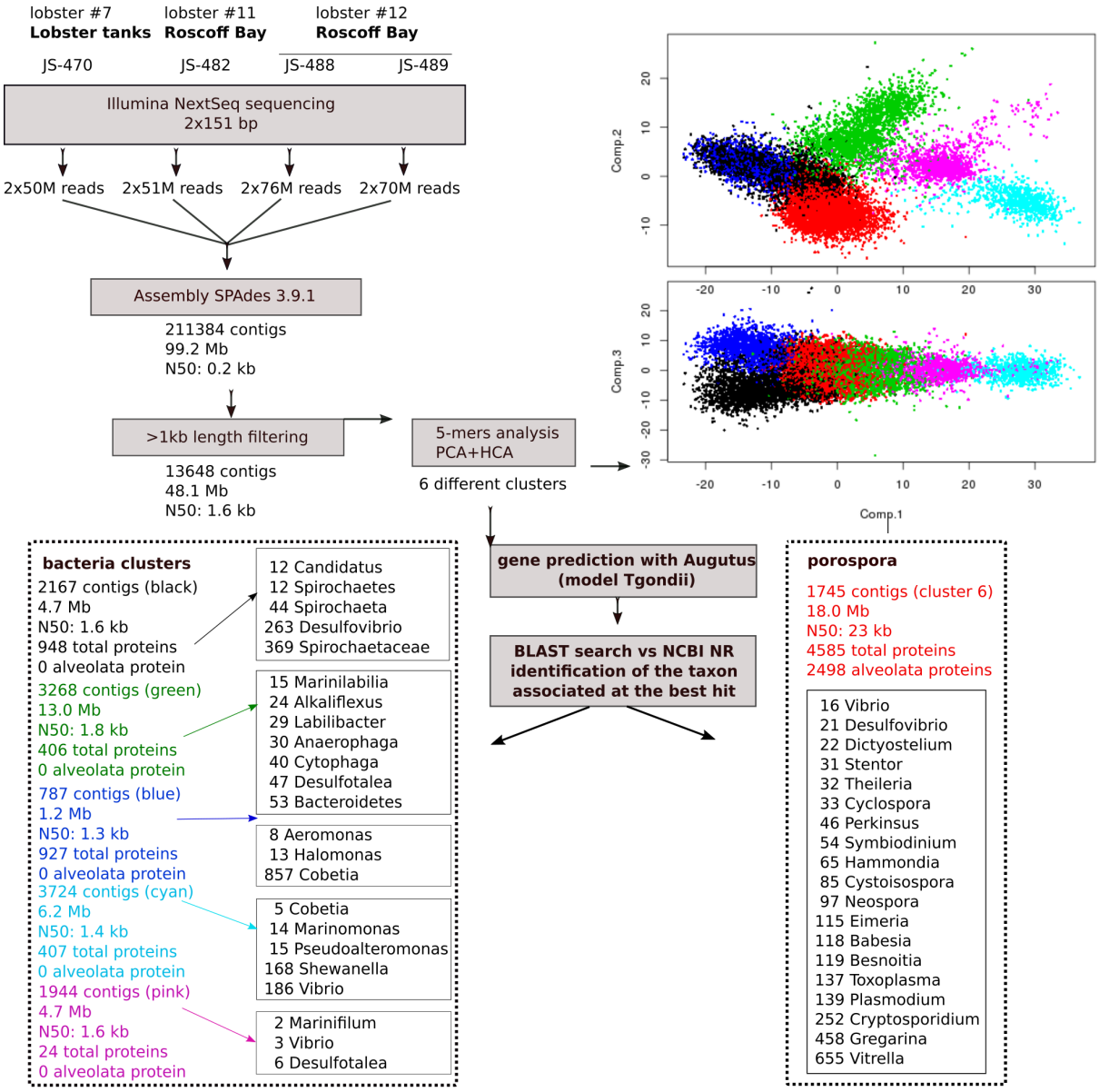


Figure 2.1: **Assembly protocol of *P. gigantea* genomic data** Identification of apicomplexan vs. contaminant contigs based on k-mer composition.

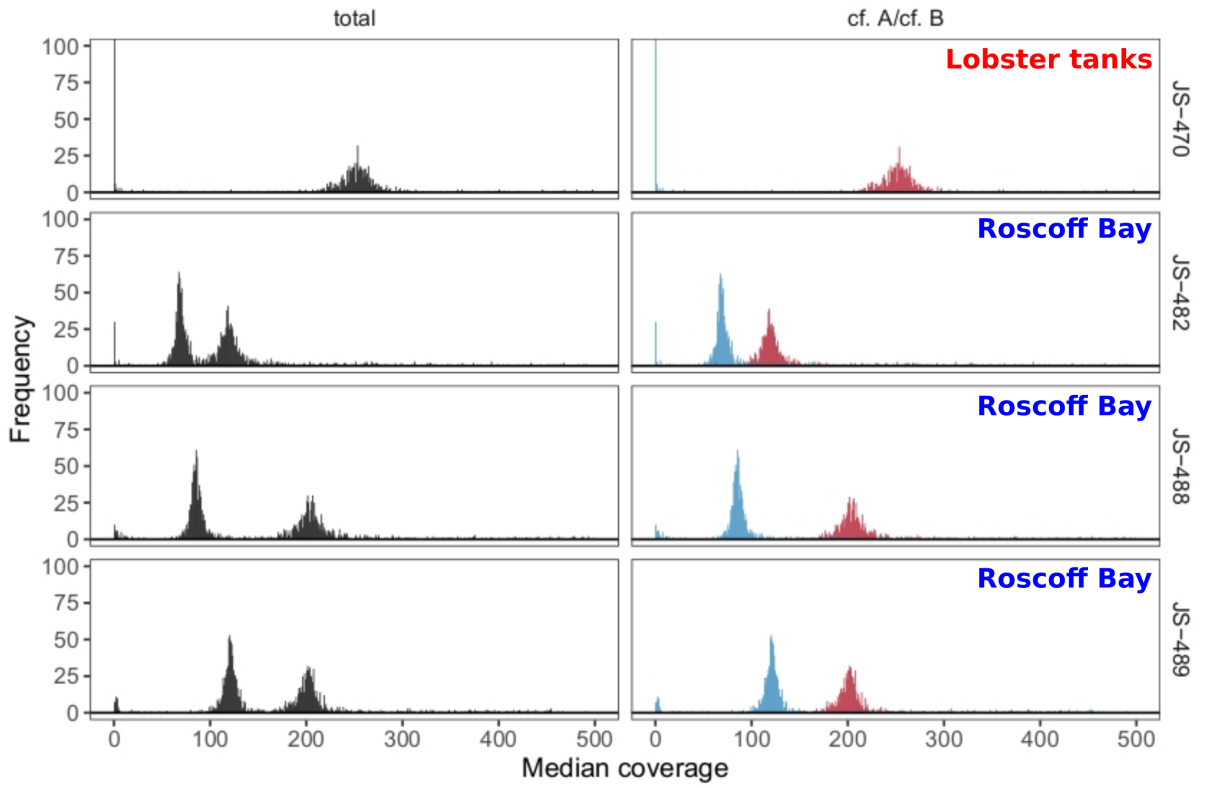


Figure 2.2: **Distribution of the median coverage in each individual library calculated for each contig from the raw assembly**. Total coverage are presented in black (left side). After genomic attribution of each contig, plot is presented again in red for *Porospora cf. gigantea* A and in blue for *Porospora cf. gigantea B* (right side). Analyse done by E. Duvernois-Berthet, in Boisard et al. (submitted).

different coverage values. The in-silico analysis of these two sets revealed an equivalent genome size of ∼9Mb.

This coverage difference was used to split the contigs into two sets that were given the names A (for the set present in all four samples) and B (for the set present only in three lobsters freshly captured in the wild) (Figure 2.3.C, page 56).

Three hypotheses were formulated following the discovery of redundant data in 3 of the 4 assemblies of the *Porospora gigantea* genome (JS-482, JS-488, JS-489), while such redundancy was not found in the JS-470 assembly.

- **Assembly error.** The first, which assumes an assembly error, was ruled out by the experimental validation by PCR amplification and RFLP discrimination of the presence of two copies of the aminopeptidase gene within the genomic DNAs of JS-482, JS-488, and JS-489, with only one copy found in the genomic DNA of JS-

470 (analyse done by A. Labat, I. Florent and myself), and later the experimental validation of two different ribosomal loci (see Chapter 3, section 3.2.2, page 75).

- **Genomic duplication or polyploidy.** The second hypothesis was that the three isolates JS-482, JS-488 and JS-489 corresponded to a species (B) with a duplicated or diploid genome, while isolate JS-470 corresponded to another species (A) without duplication. Under this assumption, the blue and red contig clusters potentially corresponding to the two copies of the B genome should had approximately the same coverage within each isolate JS-482, JS-488 and JS-489 (Figure 2.2, page 54), and a constant ratio was expected between the coverage rates of these two groups of contigs. However, the proportion of genomes A and B in each biological DNA sample has been estimated as 100%-0% for JS-470, 63.2%-36.8% for JS-482, 70.5%-29.5% for JS-488 and 62.4%-37.6% for JS-489, based on medium coverage levels (Figure 2.2, page 54).

- **Two different taxa** The third hypothesis suggested the presence of two different taxa in isolates JS-482, JS-488 and JS-489 and only one taxon in isolate JS-470. This hypothesis also assumed that one of the two taxa present in JS-482, JS-488 and JS-489 was the same as the taxon found in isolate JS-470. This hypothesis was in accordance with the observed median coverages: the bimodal coverage could thus be explained by the relative (and variable) proportions of the two taxa within the different isolates, which are always different. Sample JS-470 would therefore contain only one of the two taxa present in the isolates JS-482, JS-488 and JS-489, named taxon A; a second taxon, named B would be present only in isolates JS-482, JS-488 and JS-489.

We were never able to refute this last hypothesis; furthermore, it is also supported by the different ecological origin of the hosts from which these isolates were collected. Indeed, isolates JS-482, JS-488 and JS-489 were taken from lobsters caught in the wild in Roscoff Bay and quickly dissected, while isolate JS-470 was collected from a lobster from a lobster tank maintained in artificial living conditions for several weeks after its capture

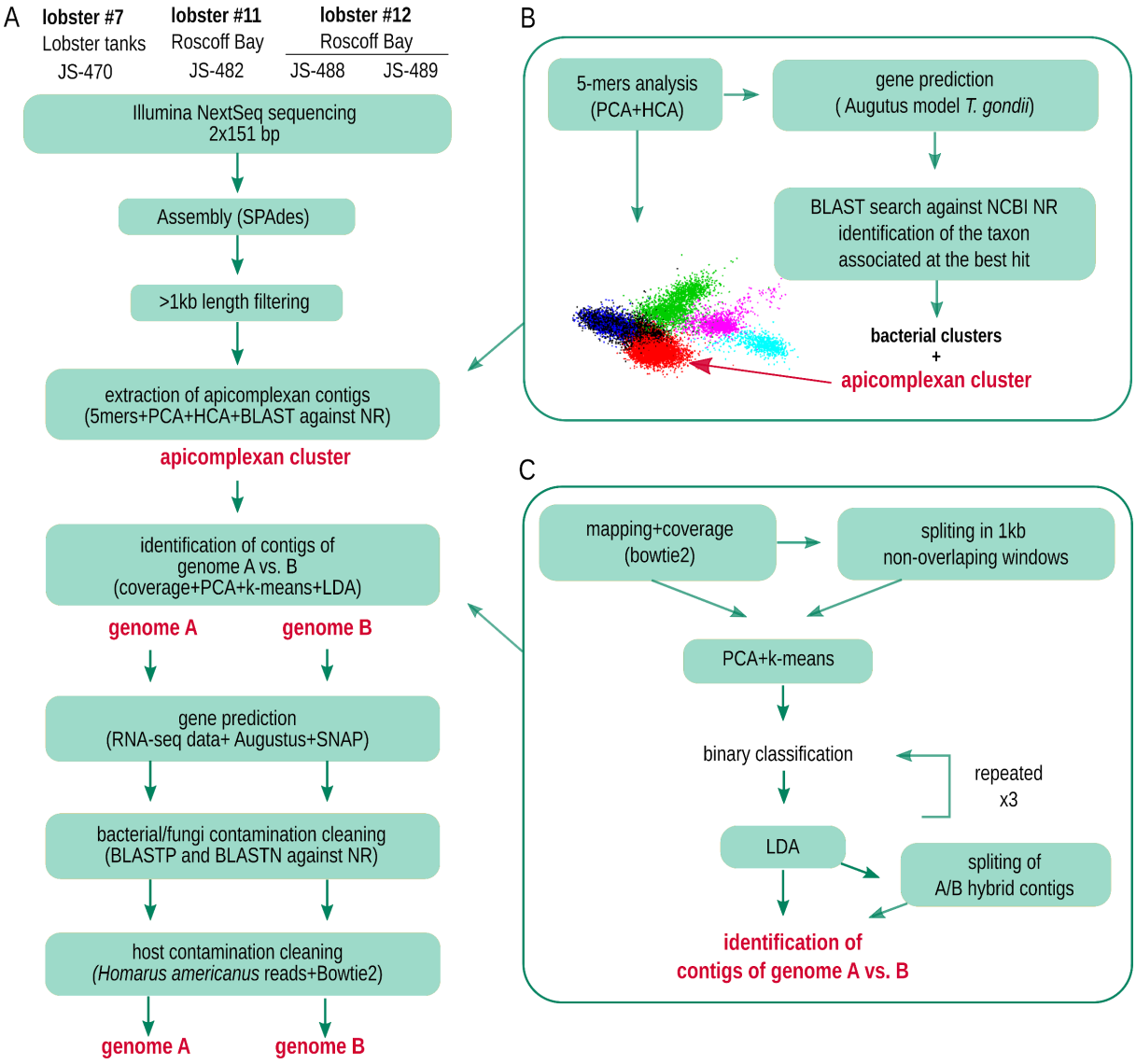


Figure 2.3: **Assembly protocol of the two genomes of *P. gigantea*.** A. Overview of the full protocol. B. Identification of apicomplexan vs. contaminant contigs based on k-mer composition. C. Identification of contigs from genomes A and B based on coverage data for each individual library. In Boisard et al. (submitted).

in the wild.

2.3.2 Creating a RNAseq based gene model

During the extraction of apicomplexan contigs phase of the protocol, we used the Augustus natively implemented gene model of *T. gondii*. We used these predicted proteins to perform a BLASTP search to identify the taxon associated with the best result and thus detected contamination contigs, as explained in the previous section and in Figure 2.1, page 53. But in a taxonomic group with this level of divergence, one simply cannot use

a gene model as distant as *T. gondii* can be from the gregarines (see schematic overview of apicomplexan phylogeny in Figure 1.2, page 14). We quickly realized how few genes were predicted for A and B with the *T. gondii* gene model: just about 2500 genes for each genome, while the existing data for *G. niphandrodes* suggested at least the double (6375 predicted genes for a genome of about 14Mb, see Chapter 1, Table 1.1, page 32). In addition, many predicted genes appeared to be fusions of at least two unrelated genes.

In the absence of genomes close enough to serve as references, we were left in a situation in which gene predictions had to be achieved *ab initio*. *Ab initio* predictions, also called *de novo* or intrinsic, are methods that operate only from the nucleotide sequence itself. However, *ab initio* gene finding in eukaryotes is complex, especially in less well studied species as apicomplexan (Scalzitti et al., 2020). Indeed, promoters and other regulatory signals can vary and thus are difficult to recognize automatically. Splicing mechanisms also rely on the specific characteristics of introns, which can differ from one lineage to another. The codon usage is also highly variable, due to GC% variations. Therefore, the use of transcriptomic data in parallel is of great value: mRNAs can provide considerable information on the characteristics of transcripts which in turn can be used to build a gene prediction model.

Thankfully we were able to generate RNAseq data for *P. gigantea*. Reads from the two sequenced RNA libraries were pooled and assembled with Trinity (Haas et al., 2013); the assembled transcripts were then aligned to the "apicomplexan cluster" contigs from *P. gigantea* pooled asembly using GMAP (Wu and Watanabe, 2005) implemented in the PASA pipeline (Haas, 2003). We first trained Augustus on a subset of transcripts, creating a first gene model which was efficient in predicting multi-exonic genes, but which was prone to create protein fusions in our proteomes. We therefore decided to use in parallel another gene prediction software called SNAP, and merged the two predictions; the whole gene prediction worflow is presented in Figure 2.4, page 58.

Following this protocol, we greatly improved gene predictions, with about 5300 predicted proteins for A and B. In order to refine these predictions, we analyzed them in order to more precisely detect any contamination that might have escaped us during the

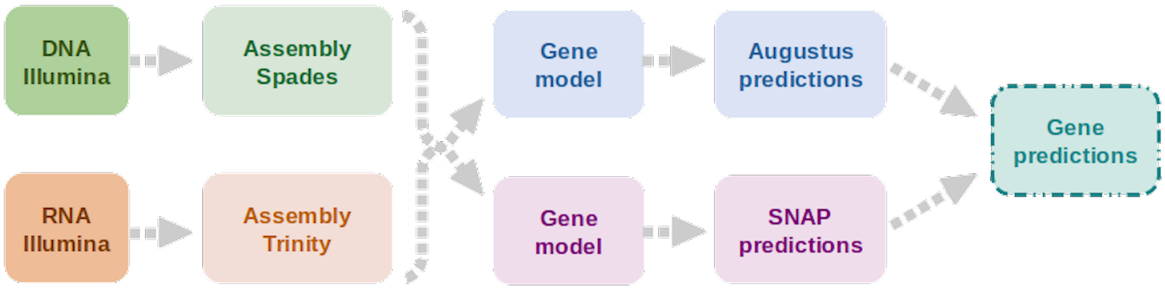


Figure 2.4: **Gene prediction workflow for *P. cf. gigantea* genomes.**

demultiplexing process. For microorganism-related contaminants, we queried NR Genbank database to identify both bacteria and fungi remaining sequences. We also mapped reads from the American Lobster Sequencing Project to our contigs to further identify host-related contamination.

Finally, at the end of all the analyses carried out and summarized in Figure 2.3, the genome A is composed of 786 scaffolds for a total of 8.8Mb whereas the genome B is composed of 933 contigs for a total of 9.0Mb (Table 2.2). The contigs from both genomes can be aligned over 7.7Mb, with a percentage of divergence around 10.8% at nucleotide level. These highly related genomes A and B are associated with the species name *P. cf. gigantea*; the detailed process of species delineation, using integrative taxonomy, is presented further in Chapter 3, part 3.3, while detailed genomes characteristics are discussed in Chapter 4, part 4.3.1.

Genomes	**Nb of Contigs >1kb**	**Length (Mb)**	**GC%**	**Gene count**
P. cf. gigantea A	787	8.8	54.3	5270
P. cf. gigantea B	933	9	54.3	5361

Table 2.2: **Characteristics of *Porospora cf. gigantea* A and B assemblies**.

2.4 *D. hatti* and *G. acridiorum*

2.4.1 Assembly and apicomplexan contigs extraction

The extraction of putative apicomplexan contigs from *G. acridiorum* and *D. hatti* genomic assemblies were performed following the same protocol as established for *P. gigantea*.

G. acridiorum PCA and HCA analyses revealed 3 clusters of contigs (Figure 2.5, page 60). BLASTP vs NR analysis of gene predictions on each cluster evidenced one cluster gathering mainly apicomplexan hits while the two others were composed of a majority of proteins with bacteria/fungi hits. These analyses indicated that cluster 2 contained the apicomplexan contigs among the *G. acridiorum* assembly. Therefore, we retained all cluster 2 contigs for further analysis.

The scatterplot from the PCA derived from *D. hatti* pooled assembly did not clearly show any clusters of contigs; though we decided to perform a HCA in order to distinguish 3 possible classes (Figure 2.6, page 61). However, BLASTP vs. NR analysis of gene predictions did not reveal any differential profile since the majority of best hits were associated with apicomplexan species for each cluster. Since these analyses were similar with the independent libraries (JS-626a and JS-627a), we continued with the pooled library assembly. We therefore retained the entire set of contigs assembled from the pooled *D. hatti* librairies, as no clear pattern of contamination emerged from our analysis.

Metrics for both *G. acridiorum* and *D. hatti* assemblies are displayed in Table 2.3.

Isolate Identifier	**Nb of Contigs >1kb**	**Length (Mb)**	**GC%**	**N50 (Kb)**
G. cridiorum JS-313	4741	25	43.11	57
D. hatti JS-626a & JS-627a	3888	13	27.14	4.9

Table 2.3: **Characteristics of the assemblies from the isolates of *G. acridiorum* and *D. hatti* cysts after the extraction of apicomplexan contigs.** The N50 value indicates the scaffold size for which 50% of the scaffolds are smaller or larger.

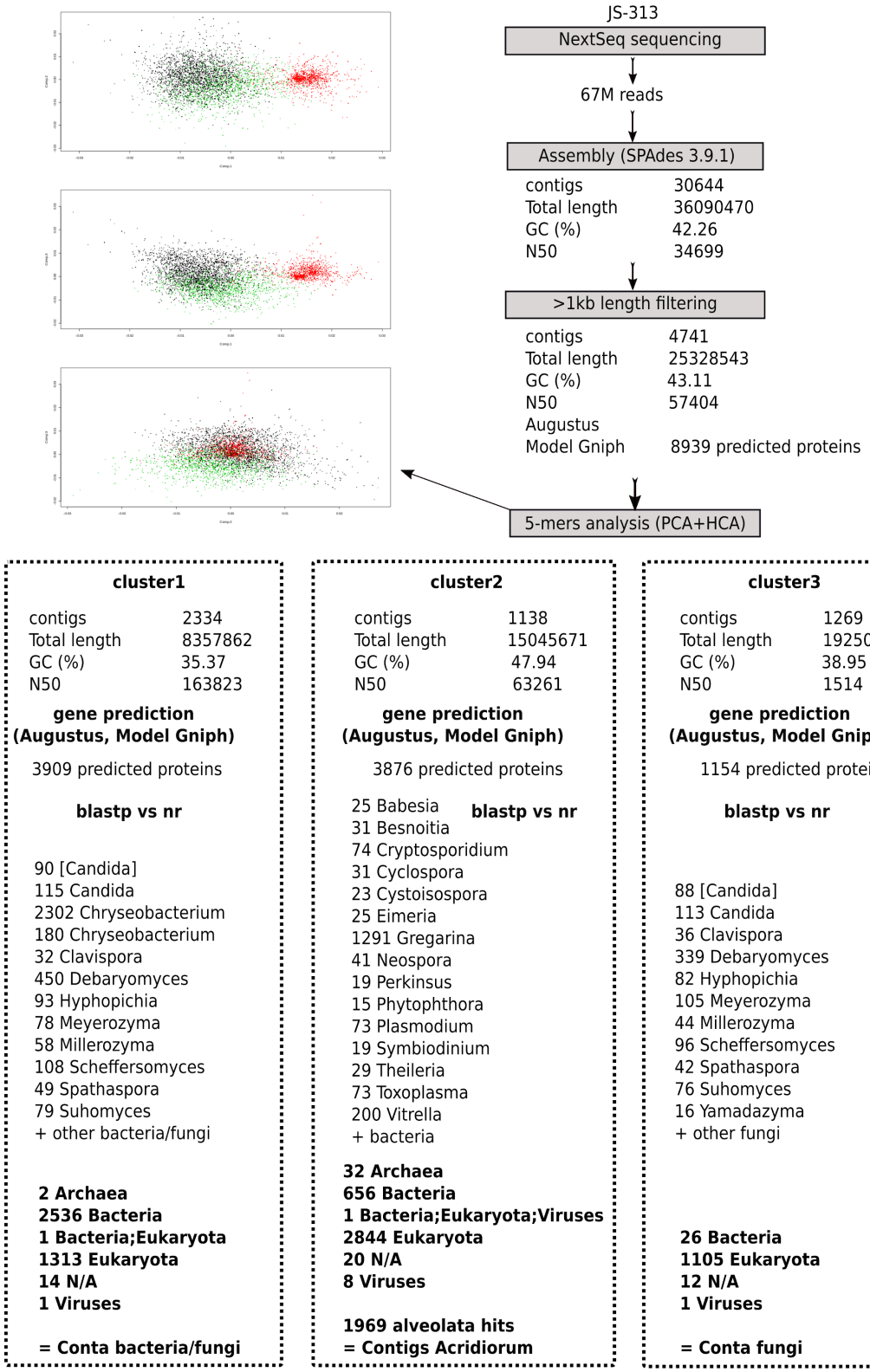
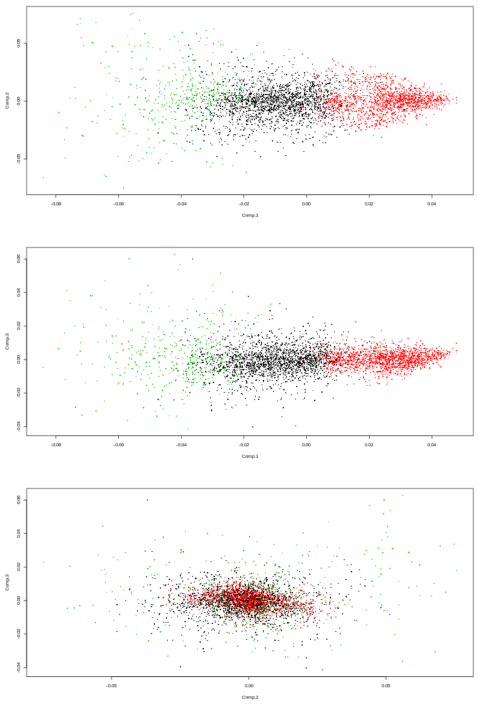


Figure 2.5: **Assembly protocol of *G. acridorum* genomic data** Identification of apicomplexan vs. contaminant contigs based on k-mer composition. Cluster 1 is colored in black on the scatter plot; cluster 2 in red and cluster 3 in green.



JS-626a
NextSeq sequencing
67M reads

JS-627a
NextSeq sequencing
70M reads

Assembly (SPAdes 3.9.1)

contigs	93972
Largest contig	194429
Total length	34018625
GC (%)	28.02
N50	3076

>1kb length filtering

contigs	3888
Total length	13381681
GC (%)	27.14
N50	4948
Augustus on Ch	6260 predicted proteins

5-mers analysis (PCA+HCA)

cluster1

contigs	1910
Total length	6275128
GC (%)	23.33
N50	4311

gene prediction (augustus, ModelCh)
2549 predicted proteins

blastp vs nr

18 Ascogregarina
25 Babesia
23 Besnoitia
85 Cryptosporidium
29 Eimeria
50 Gregarina
32 Neospora
8 Perkinsus
80 Plasmodium
17 Theileria
44 Toxoplasma
126 Vitrella
+ other bacteria

30 Archaea
672 Bacteria
1270 Eukaryota
17 N/A
13 Viruses

about 550 apicomplexan hits = Contigs Diplauxis

cluster2

contigs	1522
Total length	6121377
GC (%)	32.66
N50	10740

gene prediction (augustus, ModelCh)
3402 predicted proteins

blastp vs nr

77 Ascogregarina
47 Babesia
56 Besnoitia
162 Cryptosporidium
38 Cyclospora
39 Cystoisospora
46 Eimeria
153 Gregarina
50 Neospora
20 Perkinsus
120 Plasmodium
14 Symbiodinium
59 Theileria
98 Toxoplasma
293 Vitrella
+other bacteria/host

23 Archaea
663 Bacteria
2289 Eukaryota
16 N/A
20 Viruses

about 1200 apicomplexan hits = Contigs Diplauxis

cluster3

contigs	456
Total length	985176
GC (%)	17.14
N50	2394

gene prediction (augustus, ModelCh)
309 predicted proteins

blastp vs nr

3 Babesia
3 Besnoitia
7 Cryptosporidium
7 Cyclospora
3 Cystoisospora
2 Eimeria
4 Gregarina
3 Neospora
13 Plasmodium
4 Toxoplasma
12 Vitrella
+ other bacteria

3 Archaea
3 Bacteria
139 Eukaryota
3 N/A

about 70 apicomplexan hits = Contigs Diplauxis

Figure 2.6: **Assembly protocol of *D. hatti* genomic data** Identification of apicomplexan vs. contaminant contigs based on k-mer composition. Cluster 1 is colored in black on the scatter plot; cluster 2 in red and cluster 3 in green.

2.4.2 Dealing with the absence of RNAseq data

As discussed in the previous section, predicting the genes of species whose genomes have been *de novo* assembled and for which we do not have reference genomes is a challenge. While we were able to generate RNAseq data for *P. gigantea*, this was unfortunately not possible for the other two gregarine species *G. acridiorum* and *D. hatti*. My goal here has been to characterize these first gregarine genomes as best as possible while waiting to be able to generate RNA data for these species. Thus, I aimed to generate the best possible predictions from these genomic data. This section describes a method that is not intended to produce publishable genomes, but rather outlines the different options to provide significance from currently available data, while not relying on yet to be produced data.

Gene models

The first step was to inventory the gene models for apicomplexan species available and natively implemented in the Augustus software. These are those of *Toxoplasma gondii* and *Plasmodium falciparum*. For convenience, all models have been renamed with a short identifier:

TgAug model *Toxoplasma gondii* model implemented in Augustus

PfAug model *Plasmodium falciparum* model implemented in Augustus

Two new gene models were also created from the VEupathDB deposited annotations for these two species in order to confront them with the models natively implemented in Augustus.

Tg model retraining (new model) from *T. gondii ME49*

Pf model retraining (new model) from *P. falciparum 3D7*

In addition, a gene model for *Gregarina niphandrodes* and *Cryptosporidium hominis* was also created from the VEupathDB deposited annotations for these two genomes:

Genomes	Contigs	Length (Mb)	GC (%)
T. gondii ME49	2075	65	52.30
P. falciparum 3D7	16	23	19.34
C. hominis 30976	53	9	30.13
G. niphandrodes Unknown	468	14	53.78
P. cf. gigantea A	737	8.8	54.3
P. cf. gigantea B	933	9	54.3
D. hatti	3888	13	28.02
G. acridiorum	4741	25	43.11

Table 2.4: **Metrics of genomes used for gene prediction comparisons**

Gniph model training (new model) from *G. niphandrodes Unknown*

Ch model training (new model) from *C. hominis 30976*

We also included the RNAseq based gene model created for *P. gigantea* as part of our comparisons.

Pg model training (new model) from *Porospora gigantea* RNAseq data

Finally, two gene models for *Diplauxis hatti* and *Gregarina acridiorum* were created with Exonerate (Slater and Birney, 2005). This software allows to search for proteins conserved in the 25 references species within genomic contigs using similarity and to build a gene model from them, in order to overcome the lack of RNA data.

Dhatti model training (new model) with Exonerate

Gacri model training (new model) with Exonerate

The metrics of all genomic assemblies used for gene prediction comparisons are summarized in Table 2.4.

Gene predictions metrics

A total of 36 gene predictions were performed for the 4 available gregarine genomes (contigs >1kb) according to the 9 available gene models. 27 control predictions were also performed on the genomes of *T. gondii*, *P. falciparum* and *C. hominis*. A total of 63 gene predictions were made from 9 gene models and 7 genomes.

In order to evaluate the quality of these numerous predictions and to be able to compare them, we considered the following metrics:

1. **Number of predicted genes** This metric should be analyzed in relation to the number of genes theoretically expected for a given taxon. However, since references are not available for closely related species, one must be cautious in interpreting the number of predicted genes. Few genes is problematic, but conversely a very high number of predicted genes is probably a sign of many fragmented genes predicted by an inadequate model.

2. **Average length of predicted genes (CDS coding sequence)** The average CDS length can be compared with the average gene lengths predicted for the reference genomes. For CDS (i.e. excluding introns which can be subject to a lot of variability), a similar length is expected for a given protein across a same taxonomical group.

3. **Search for fused predicted genes.** The aim is to search for the number of predicted proteins that appears to have merge several adjacent proteins from the reference proteomes. A BLASTP of the proteins of each gene prediction against each of the 25 reference proteomes was performed. The best hits were selected each time, and merged for the same prediction-proteome couple, allowing to identify multiple adjacent hits. Thus, a protein found to align with to 2 or more adjacent proteins in more than 20 of the 25 reference proteomes is considered to be the result of a gene fusion occurring during prediction.

4. **BUSCO evaluation**. BUSCO's rationale is to evaluate the presence in the predicted proteins of a set of single-copy orthologous genes considered universal in a given group. The more of those genes are retrieved, the better the gene predictions is thought to be.

Controls

In order to provide controls, we assessed the gene predictions metrics of the reference genomes such as they are deposited on GenBank and VEupathDB. The official metrics for these genomes are presented in Table 2.5. We also computed our own metrics on these

assemblies to ascertain their congruence.

	G. niphandrodes Unknown	*C. hominis 30976*	*P. falciparum 3D7*	*T. gondii ME49*
Nb of protein coding genes	6375	3949	5460	8322
Mean gene length (nt)	1375	1759	2267	2407

Table 2.5: **Official metrics of deposited genomes**

For each of these genomes, several gene models are also available: the two models natively implemented in Augustus (**PfAug model** for *P.falciparum* and **TgAug model** for *T. gondii*), as well as our own gene models: **Ch model** for *C. hominis*, **Gniph model** for *G. niphandrodes*, **Pf model** for *P. falciparum* and **Tg model** for *T. gondii*.

Re-predictions were made for these 4 genomes according to all of these gene models, and our metrics were compared to the official metrics; the tables 2.6 and 2.7 show the number of genes and their average length for all of these predictions (in bold).

The analysis of all these results showed that the predictions made with our own gene models are very similar to the gene predictions of the 4 reference genomes, and of higher quality than those made with the gene models implemented in Augustus (that we believe to be old, non-updated gene models).

Choosing the best gene model for *G. acridiorum* and *D. hatti*

Using these metrics from the official annotations, we were able to compare the set of predictions made for each of our assemblages, and define the most appropriate gene model for each. For the sake of conciseness, the whole set of calculated metrics is not represented, but the metrics **Number of predicted genes** and **Average length of predicted genes** for all performed predictions are reproduced in Table 2.6 and Table 2.7. An exemple of **Predicted fused genes** search is displayed in Figure 2.7. Finally, the **BUSCOs assessments** for all predicted proteomes for *G. acridiorum* and *D. hatti* are reproduced in Figure 2.8.

The gene predictions for *Gregarina acridiorum* and *Diplauxis hatti* that overall showed the best metrics were the ones made with the following models:

Genome	Pg model	Dhatti model	Gacri model	Gniph model	Ch model	PfAug model	Pf model	TgAug model	Tg model
P. cf. gigantea A	5270	1144	822	4232	5241	7851	7118	2291	3409
P. cf. gigantea B	5361	1211	852	4257	5297	7830	7065	2294	3444
G. acridiorum	3220	116	2423	8939	14373	23888	20746	1176	745
D. hatti	587	2386	819	4275	6260	10026	9231	106	37
G. niphandrodes	4809	477	267	**5040**	6172	9588	8509	2377	3140
C. hominis 30976	166	2720	68	53	**3807**	4101	3926	46	30
P. falciparum 3D7	403	30	27	3927	4879	6052	**5575**	211	67
T. gondii ME49	4943	1309	103	7362	29390	69156	56327	5419	*7191*

Table 2.6: **Number of predicted genes.** The numbers in **bold** correspond to the deposited genomes control predictions (predictions using the model created from the same genome)

.

	Pg model	Dhatti model	Gacri model	Gniph model	Ch model	PfAug model	Pf model	TgAug model	Tg model
P. cf. gigantea A	1527	5233	4471	1547	1311	999	1111	2352	1587
P. cf. gigantea B	1536	5021	4431	1556	1319	1013	1131	2377	1578
G. acridiorum	1447	39829	3609	1499	1225	807	929	1564	1147
D. hatti	1190	1805	2449	1256	1090	817	895	938	671
G. niphandrodes	1722	16039	15398	**1864**	1468	1194	1343	2654	1828
C. hominis 30976	804	41606	25152	1940	**1841**	1721	1801	607	609
P. falciparum 3D7	627	63331	30848	2430	2272	2067	**2231**	611	885
T. gondii ME49	1600	3185	3247	1607	922	671	807	2839	**2679**

Table 2.7: **Average gene length (CDS) in nucleotides.** The numbers in **bold** correspond to the deposited genomes control predictions (predictions using the model created from the same genome)

.

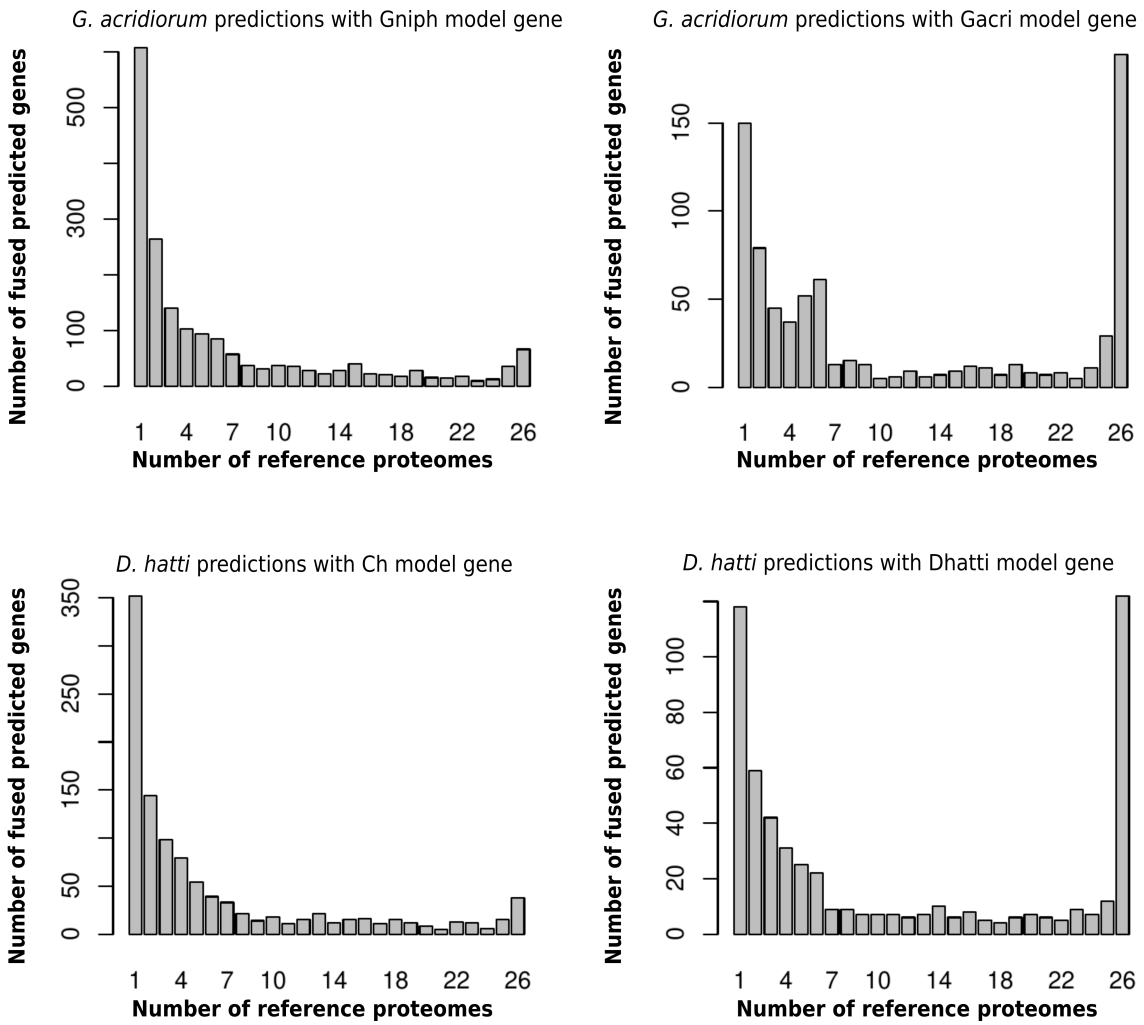


Figure 2.7: **Number of fused genes in *D. hatti* and *G. acridiorum* predictions**. The results of two gene predictions from two different gene models are presented for each species. For *G. acridiorum*, on overall 8939 predited genes with the Gniph model, 600 fused genes were found but only within a given proteome, which may correspond to a biological event. On the other hand, around 50 fused proteins only were identified at the scale of +20 proteomes. However, on only 2423 predicted genes with the Gacri model, more than 200 fused proteins were found at the +20 proteomes scale, which strongly suggests an artificial fusion during the prediction. Similarly for *D. hatti*, the predictions made with the Ch model found around 50 fused proteins at the +20 proteomes scale (on a overall of 6260 predicted proteins). In contrast, predictions made with the Dhatti model retrieved more than 150 fused proteins at the +20 proteomes scale (on an overall of only 2386 predicted proteins).

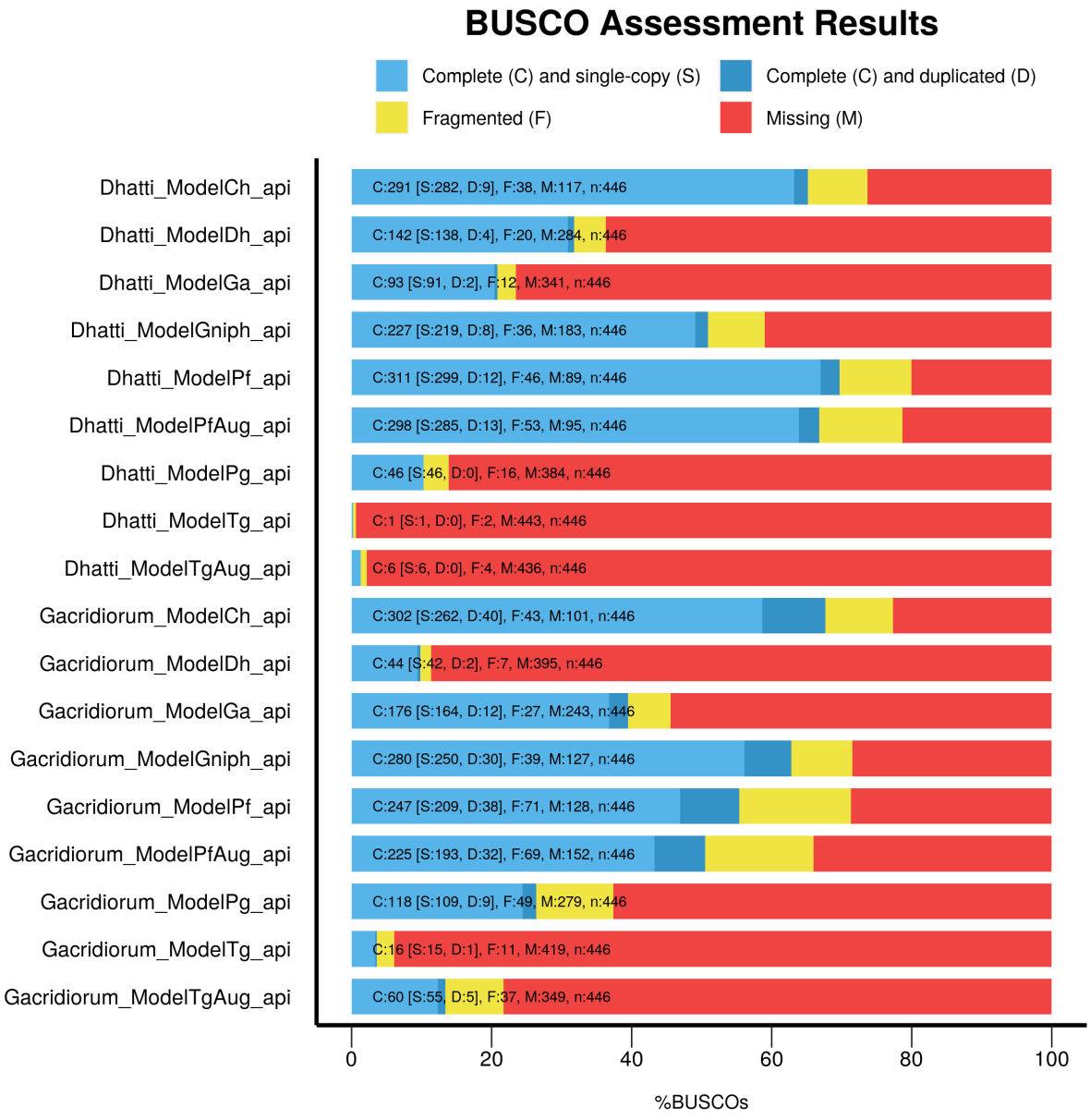


Figure 2.8: **BUSCOs assessment results for all the predicted proteomes of *G. acridiorum* and *D. hatti*** (geneset apicomplexa odb10).

for *Gregarina acridiorum*: the **Gniph model**, which we constructed from the deposited annotation of the genome of *Gregarina niphandrodes Unknown*. Indeed, this gene model is the one that results in ∼8939 predicted genes for a coding sequence of ∼1499nt in average, while displaying the lesser number of fused genes. BUSCOs assessment retrived around 65% of markers.

for *Diplauxis hatti*: the **Ch model**, which we constructed from the deposited annotation of the genome of *Cryptosporidium hominis 30976*. This gene model is the one that results in ∼6260 predicted genes for a coding sequence of ∼1090nt in average, while displaying the lesser number of fused genes. BUSCOs assessment retrived around 63% of markers.

2.5 Gregarine genomes are here and more are needed

The aim of this chapter was to outline the main challenges associated with analysing the highly divergent eukaryotic genomes of non-model, non-culturable species, and to suggest solutions to address them.

In this study, we have emphasized the importance of choosing an appropriate biological model. When collecting samples directly from the field, care must be taken to ensure that sufficient biological material can be gathered. Harvesting non-culturable samples of parasitic species involves collecting them directly from the host; therefore it is critical to guarantee easy access to the hosts, for example by using species that are themselves accessible in captivity or can be maintained in the laboratory.

Collecting directly from the host is similar to environmental sampling in that it is likely to involve picking up a bit of the host tissues as well as the surrounding bacteria or fungi. When a reference genome is not available, the removal of these environmental contaminants is particularly challenging and can be tackled through methods used in metagenomics, such as kmer composition or sequence similarity analysis. If you have a mixture of two closely related genomes in an environmental sample, the median coverage of the contigs by reads can help you discriminate between them.

For the prediction of genes of a new species without close reference, the optimal solution is currently to build a gene model using RNAseq data. However, this assumes that one is able to generate such data, which is not always obvious. In addition to the constraints outlined above, and which remain valid, preparations for RNA sequencing are more delicate to perform, especially when biological material is scarce.

As this situation is frequently encountered in our discipline, I aimed to highlight alternative solutions to the use of an RNA-based gene model. They mainly consist in generating gene models from closely related or sharing genomic characteristics species, that are summarized above. Attention should also be given to the many options of the prediction softwares, such as the introns length or the splice sites sequences - especially when working with non model and lesser known species.

Create a gene model from a set of proteins inferred by Exonerate For our two species *D.hatti* and *G.acridiorum*, the results are not conclusive. Indeed, these models failed everytime in predicting genes in our species. It is possible that the high divergence known to exist in these genomes at the nucleotide level is the cause of this setback.

Use the gene model of another species In this case, two parameters seem to predict the quality of the predictions:

- *Phylogenetic proximity* This is the prime criterion to consider. This applies to *Gregarina acridiorum*, an insect-infecting gregarine, for which the genetic model of *G. niphandrodes*, also an insect-infecting gregarine, seems to be the most suitable. Indeed, as far as we know, the phylogenetic relationships between insects infecting gregarines tend to follow those of their hosts (Florent et al., 2021).
- *GC%* In the case where a GC% bias is known, the gene model of a species with the same bias should be preferred. This is the case for *Diplauxis hatti*, a AT-rich genome, whose gene model with the best metrics is *Cryptosporidium hominis*, also AT-rich.

However, none of the methods discussed is sufficient to assess the quality of a gene prediction in the absence of a reference. It is through the confrontation of each of the metrics that sufficient information can be gathered. Despite this, such metrics may also contradict each other; for example, the BUSCOs assessment for *G. acridiorum* is better when the **Ch model** is used, but this model also creates many fragmented genes.

We believe that we have now characterized the genomes of *G. acridiorum* and *D. hatti* to the best of our ability, based on the available data. We have managed to define new references for the gregarines in terms of genome size and structure, as well as the number of proteins that can be expected. These gene predictions will only be effectively refined with the contribution of RNA data, but in the meantime we already have a better picture of the genomic diversity of the gregarines. For instance, one of the most striking results of this research concerns precisely the extent of this diversity. We expected that the gene model for *P. gigantea* would be a sufficiently close gene model to predict genes in other gregarines. However, we found that even for a marine gregarine, it is still too divergent due to a AT-rich bias in *D. hatti* sequences.

Sequencing techniques are evolving very quickly, and it is likely that these challenges will be overcome in the coming years. During this PhD, we have seen the effectiveness of Single Cell technology to produce transcriptomic data on trophozoites stages, as demonstrated by the recent papers presented in the first chapter (Mathur et al., 2019; Janouškovec et al., 2019; Salomaki et al., 2021; Mathur et al., 2021b). It is to be expected that more gregarine genomes will enrich the public databases, so that the full diversity of apicomplexan can be revealed.

This sampling effort will allow us to revise the taxonomy of the gregarines which is likely to be disrupted by the input of molecular information. Here we have highlighted the presence of two very close genomes where only one species, *P. gigantea*, was expected. The next chapter will be devoted to describe the integrative taxonomy approach that allowed us to decide, at least temporarily, on two species names for these two genomes. This chapter will also be dedicated to the presentation of a complementary example concerning *G. acridiorum* versus *G. garnhami* species discrimination.

Chapter 3

Investigate the diversity of gregarines using integrative taxonomy

3.1 Species delimitation in gregarines

Gregarines are a heterogeneous group of apicomplexan parasites infecting a very wide diversity of non-vertebrate hosts, in which they mostly occupy intestinal tracts and coelomic spaces (Desportes and Schrével, 2013). The biodiversity of gregarines currently corresponds to ∼1700 formally described species (Portman and Šlapeta, 2014), but according to experts in the field, this number may be vastly underestimated (Adl et al., 2019; Desportes and Schrével, 2013). Recent metagenomic surveys of terrestrial soils and marine environments further confirmed the high occurrence and abundance of gregarine-like sequences in these environments, that remain however poorly ascribed to formally described species (de Vargas et al., 2015; del Campo et al., 2019; Mahé et al., 2017).

Gregarine species assignations have been historically based on combinations of morphological and behavioral features including parasitic life traits (host and host range specificities), parasite locations in hosts (i.e. intestine or coelom, more rarely intracellular), life-cycle developmental stages descriptions (morphology measures, duration of phases,

SEM and TEM microscopy), gamonts pairing (frontal, lateral, caudo-frontal) and gametocysts dehiscence modes (Clopton, 2009; Desportes and Schrével, 2013; Levine, 1988). During the last decades, increasing consideration of molecular data enabled confirming but sometimes revising the taxonomic and phylogenetic vision we have of gregarines, revealing synonym for some species that were once considered distinct (Hussain et al., 2013) or conversely, allowing identifying novel cryptic species, i.e. morphologically indistinguishable but clearly distinct at the molecular level (Ninham, 1995). In some cases, species have been taxonomically relocated to other groups, after molecular markers were sequenced and phylogenetic analyses performed (Janouškovec et al., 2019; Mathur et al., 2019, 2021a).

In this chapter, two examples of gregarine species delimitation are presented: one concerns the marine gregarine *P. cf. gigantea* for which we assembled two genomes, as exposed in the previous chapter. The other one reassesses the taxonomy of the terrestrial gregarines species infecting two species of locusts: *G. garnhami* infecting *Schistocerca gregaria* and *G. acridiorum* infecting *Locusta migratoria*. Both studies demonstrate, in different contexts and by different means, the presence of cryptic species infecting both the European lobster *Homarus gammarus* and two locust species, *Locusta migratoria* and *Schistocerca gregaria*. The evidence of a co-infection by two cryptic species, on the one hand, and the existence of different gregarines where it was previously thought that only one species infested two locusts, on the other hand, was enabled by a thorough integrative taxonomic analysis, combining morphological and molecular data.

3.2 Methods

3.2.1 Morphological studies

Porospora cf. gigantea A and B, marine gregarine species

Collection of hosts, isolation of parasites and morphological studies on *P. cf. gigantea* was carried out by I. Florent. See detailled methods in Boisard et al. (submitted).

As exposed in the prevous chapter, lobster specimens *Homarus gammarus* were col-

lected in the British Channel at Roscoff (Britany, France), either directly from the field (Roscoff bay) or through lobster tanks facilities (lobster provenance : South of England), in which crustaceans are maintained in captivity several weeks to months before their commercialization.

All the lobsters dissected during the study, for both morphology and molecular extractions, are identified in Table 3.1, together with their respective parasite loads.

G. acridiorum and *G. garnhami*, terrestrial gregarines

Collection of hosts, isolation of parasites and morphological studies on *G. acridiorum* and *G. garnhami* was carried out by I. Florent. See detailled methods in Florent et al. (2021). The different locust hosts used in this study are listed in the Table 3.2.

Statistical tests. In order to compare the measurements' averages carried out for the gregarines infecting either *S. gregaria* or *L. migratoria* hosts, statistical tests were performed as follows. For the group of measurements with n = 18, a Shapiro-Wilk Test was used to assess the normality of the data, which established normality. For this sample and all the other groups of measurements tested with n > 30, we used parametric tests. First, a Fisher test was conducted to test the homoscedasticity of the variances within the groups. When homoscedasticity was retrieved, a Student's t-test was conducted to compare the means of each group. When homoscedasticity was not retrieved, a Welch's t-test was then performed. Analyses were performed using R software.

3.2.2 Molecular studies

Sequence data

Experimental reconstruction of 18S/28S loci of *P. cf. gigantea* A and B has been achieved by I. Florent, L. Duval and A. Labat. See detailled methods in Boisard *et al.* (submitted). The two complete ribosomal loci (5977bp) are available on NCBI GeneBank under the accession number PRJNA734792.

Partial 18S SSU rDNA gene amplification and sequencing for the *L. migratoria* in-

Lobster Specimen	Sampling date	Host from Tanks/Bay	Host sex	Host Lenght(cm)	Host Weight (g)	Cysts load in host rectal ampulla	Trophozoites Load in host gut lumen
#1	24/05/16	Tanks	male	25	355	10	none
#2	24/05/16	Tanks	male	29	645	10-100	none
#3	24/05/16	Tanks	female	26	450	10-100	none
#4	25/05/16	Tanks	female	29	620	10	none
#5	25/05/16	Tanks	male	25	420	10-100	none
#6	26/05/16	Tanks	male	29	745	10	10
#7	26/05/16	Tanks	male	25	375	10-100	10
#8	27/05/16	Tanks	female	27	445	10-100	none
#9	27/05/16	Tanks	male	26	490	10-100	none
#10	30/05/16	Tanks	male	26	470	none	none
#11	30/05/16	Bay	female	25	420	10-100	none
#12	31/05/16	Bay	male	24	465	100-1000	10
#13	31/05/16	Bay	female	24	435	100-1000	none
#14	18/10/16	Tanks	male	27	485	200	10
#15	19/10/16	Tanks	male	26	685	none	none
#16	19/10/16	Tanks	female	27	535	none	none
#17	20/10/16	Bay	male	23	455	100-300	10
#18	20/10/16	Bay	male	25	450	10-100	10
#19	24/10/16	Bay	female	25	510	10-100	none
#20	24/10/16	Tanks	male	23	405	10-100	10
#21	25/10/16	Tanks	male	27	550	none	none
#22	26/10/16	Tanks	female	30	895	10-100	10
#23	26/10/16	Tanks	male	29	510	10-100	none
#24	03/10/17	Tanks	female	27	505	10-100	10
#25	03/10/17	Tanks	female	28	580	10-100	10
#26	04/10/17	Bay	male	34	815	10-100	none
#27	05/10/17	Bay	male	26	515	100-500	200
#28	06/10/17	Tanks	female	30	635	10	none
#29	06/10/17	Tanks	male	26	560	10-100	none
#30	09/10/17	Tanks	male	27	655	none	none
#31	09/10/17	Tanks	male	28	710	none	none
#32	09/10/17	Tanks	female	26	450	10-100	none
#33	11/10/17	Tanks	male	26	470	10-100	10
#34	12/10/17	Tanks	male	27	510	10-100	none
#35	17/07/15	Bay				100-300	none

Table 3.1: **Sampling of the lobsters specimen and parasite load.** Data gathered by I. Florent in Boisard et al. (submitted).

Acrididae host/ designation in study	Source	Host status	Gregarines sampled
Schistocerca gregaria gregaria (2014)/SG-M	Long-standing laboratory strain from CNLA Agadir, Morocco	Sick	Young trophozoites in ceca, gamonts, syzygies and gametocysts in the midgut, occasionally gametocysts in feces; high infection level
Schistocerca gregaria gregaria (2014)/SG-B	Long-standing laboratory strain from KU Leuven, Belgium	Healthy	Young trophozoites in ceca, gamonts, syzygies and gametocysts in the midgut, occasionally gametocysts in feces; high infection level
Schistocerca gregaria flaviventris (2014)/SG-SA	Natural population from Tankwa Karoo National Park, South Africa	Sick	Young trophozoites in ceca, gamonts, syzygies and gametocysts in the midgut, occasionally gametocysts in feces; high infection level
Locusta migratoria (2012, 2014, 2015)/LM-M	Long-standing laboratory strain from MNHN Vivarium, France	Healthy	Gamonts, syzygies and gametocysts in the midgut, occasionally gametocysts in feces; mild infection level

Table 3.2: **Acridians hosts and sampled of gregarines.** 'Sick' hosts died rapidly (within days) in laboratory conditions in contrast to 'healthy' hosts that were maintained for weeks. Data gathered by I. Florent in Florent et al. (2021).

fecting gregarine *G. acridiorum* and the *S. gregaria* infecting gregarine *G. garnhami* was carried out by PCR and cloning by I. Florent and A. Labat. Their lengths are 1637bp and 1638bp long respectively, and cover the V1-V8 region of the 18S ribosomal locus. See detailled methods in Florent et al. (2021). All 43 sequences are available on NCBI GeneBank under the accession number PRJEB38991.

P. cf. gigantea 18S SSU rDNA phylogenetic analyses

Gregarines phylogeny. The 100 sequences phylogeny was built from the 18S SSU rDNA sequences of the 2 genotypes of *Porospora cf. gigantea*, which were aligned with 84 sequences from a diversity of gregarines, either marine or terrestrial, as well as 12 other sequences representative of the actual known apicomplexans lineages. Two chromerids sequences were used as outgroup (Woo et al., 2015). A total of 1614 sites were conserved after a selection of conserved blocks as defined by Gblocks (version 0.91b) (Castresana, 2000) (Parameters used: Minimum Number Of Sequences For A Conserved Position: 51; Minimum Number Of Sequences For A Flanking Position: 51; Maximum Number Of Contiguous Nonconserved Positions: 8; Minimum Length Of A Block: 3; Allowed Gap Positions: All). A General Time Reversible (GTR) substitution model with gamma-distributed rate variation across sites and a proportion of invariant sites was suggested as the best-fit model according to the Bayesian Information Criterion (BIC) and the Akaike Information Criterion (AIC) calculated by MEGA X (Kumar et al., 2018). Maximum likelihood analyses were performed using RAxML (version 8.2.12) (Stamatakis, 2014); bootstraps were estimated from 1,000 replicates. Bayesian phylogenetic tree was constructed with MrBayes (version 3.2.3) (Ronquist and Huelsenbeck, 2003) using the following parameters: lset nst=6 rates=invgamma; mcmc ngen=10000000 relburnin=yes burninfrac=0.25 samplefreq=1000 printfreq=10000 nchains=4 nruns=2 savebrlens=yes; sump burnin=2500000; sumt burnin=2500000 contype=allcompat. Trees were visualized and edited using FigTree (version 1.4.4) (Rambaut, 2018) and Inkscape (www.inkscape.org).

Environmental phylogeny focused on crustacean gregarines. The 189 sequences phylogeny was built from the 1694bp 18S SSU rDNA sequences of the genomes

A and B, which were aligned with 14 sequences from crustaceans' gregarines, as well as 154 environmental sequences from several projects described in Rueckert et al. (2011) or gathered from NCBI Genbank. The sequences from Gregarinoidae clade (n=19) were used as outgroup, as this group has been placed as a sister group to the crustacean gregarines clade in recent literature (Mathur et al., 2019; Janouškovec et al., 2019; Mathur et al., 2021a). A total of 1135 sites were conserved after a selection of conserved blocks as defined by Gblocks (Parameters used: Minimum Number Of Sequences For A Conserved Position: 95; Minimum Number Of Sequences For A Flanking Position: 95; Maximum Number Of Contiguous Non conserved Positions: 8; Minimum Length Of A Block: 3; Allowed Gap Positions: All). Maximum likelihood and bayesian analyses were performed following the same protocol and parameters as in the previous phylogeny.

G. acridiorum and *G. garnhami* 18S SSU rDNA phylogenetic analyses

Using maximum likelihood (ML) and Bayesian methods, phylogenetic trees were built from 69 sequences from gregarines infecting either *S. gregaria* (20 sequences), *L. migratoria* (23 sequences), a range of different insect hosts (22 sequences) or marine crustaceans, chosen as the gregarine outgroup specimen (4 sequences) (Clopton, 2009; Ninham, 1995; Schrével et al., 2016). Using a previously published alignment (Schrével et al., 2016), the new gregarine sequences were added manually to yield a confident alignment of 1433 positions, after selection of conserved blocks defined using Gblocks (version 0.91b) (Castresana, 2000) (Parameters used: Minimum Number Of Sequences For A Conserved Position: 35; Minimum Number Of Sequences For A Flanking Position: 58; Maximum Number Of Contiguous Nonconserved Positions: 8; Minimum Length Of A Block: 3; Allowed Gap Positions: With Half Use Similarity Matrices: Yes). Maximum likelihood and bayesian analyses were performed following the same protocol and parameters as in the previous phylogenies.

Estimates of genetic divergence between and within of *G. acridiorum* and *G. garnhami* 18S SSU rDNA sequences

The numbers of base differences per site from averaging over all sequence pairs between and within each group were calculated using MEGA X (Kumar et al., 2018). This analysis involved 44 nucleotide sequences: 20 from gregarines that infect *S. gregaria*, 23 from gregarines that infect *L. migratoria*, and the sequence of *G. caledia* that infects *C. captiva* (NCBI GeneBank accession number L31799). For each sequence pair, all ambiguous positions were removed (pairwise deletion option). There were a total of 1784 positions in the final dataset.

3.3 The discovery of co-infection of the European lobster by two species of genus *Porospora*

3.3.1 An unexpected species

As described in the previous chapter, we identified the presence of two genomes in the samples collected from European lobsters caught in Roscoff Bay. These two genomes, similar in size (∼9Mb) and highly syntenic, however display a divergence of about 10% at the nucleotide level.

In this section, we present the approach that allowed us to propose and argue the presence of two different species, as a result of an integrative taxonomy analysis involving both morphological and molecular data, including genomic data.

First, we conducted an analysis of microscopy images, with the joint goals of confirming the presence of morphological features associated with the taxonomic description of *P. gigantea* in the literature, and evaluating the presence of potential morphological characteristics which could discriminate the two taxa.

Then, we sought to reconstruct the evolutionary history of these two taxa, using several phylogenetic analyses based on the molecular marker 18S SS rDNA. A first phylogeny was devoted to place the two sequences of *P. cf. gigantea* A and B within the Apicomplexa. A

second analysis proposed a phylogeny centered on crustacean gregarines, with the addition of closely related environmental sequences available in public databases. The genomic data presented in Chapter 2 were evidently crucial for the delineation of these species.

3.3.2 Morphological and molecular descriptions' confrontation

Morphological description of *P. cf. gigantea*

Morphological description has been carried out by I. Florent; for the sake of clarity, the main results are reported below. See detailed results in Boisard et al., submitted.

In order to proceed to the biological and genomic study of *Porospora gigantea*, several specimens of the *Homarus gammarus* type host species were collected from two different sources (Table 3.1, page 76). Most specimens were acquired from lobster tank facilities in Roscoff (Britany, France; lobsters' origin: South of England), where lobsters are maintained in captivity for up to several months before their commercialization. But, as their infection levels by *Porospora* were very variable and often very low, we also turned in parallel to specimen freshly caught from the Roscoff bay (Britany, France), which were analyzed immediately after their capture (Table 3.1).

All 35 lobsters were dissected and their intestinal parasites were directly observed, first at the level of rectal ampulla, where cyst forms sometimes in very high numbers are easily seen (Figure 3.2, page 84 - highly infected rectal ampulla of Lobster #12), then further upstream along the intestinal tract where trophozoite forms, freely moving in the intestinal lumen could be rarely observed (Figure 3.1 and Figure 3.2). The infection levels were however highly variable depending on the host specimen. Out of the 35 (26 from Roscoff lobster tanks, 9 from Roscoff bay), 6 - all from lobster tanks - had no sign of infection by either cysts or trophozoites (Table 3.1). While cysts were found in the rectal ampulla of the remaining 29 examined lobsters, we found trophozoites in the digestive tract of only 11 of them. Globally, the infection levels were significantly much higher in the lobsters freshly captured from the Roscoff bay than that in the lobsters maintained in captivity in tanks (Table 3.1). Interestingly, a similar observation was done by Van

Beneden in its initial description of *Porospora (Gregarina) gigantea* in 1869:

> Je n'ai pas trouvé de traces de ces parasites sur les homards tenus pendant longtemps en captivité dans les parcs à Ostende. En serait-il des homards conservés dans les parcs comme des animaux de nos jardins zoologiques et des poissons de nos aquariums ? La perte de leurs parasites serait-elle la conséquence de leur captivité ?

that can be translated as:

> I have not found any traces of these parasites on lobsters kept for a long time in captivity in the parks in Ostend. Would it be the same for lobsters kept in parks as for animals in our zoos and fish in our aquariums? Would the loss of their parasites be the consequence of their captivity? (Van Beneden, 1869).

Morphological measurements were performed on these cysts and trophozoites, to be compared to the available descriptions on *Porospora gigantea* gathered in Desportes and Schrével (2013), in order to confirm whether the parasites observed in these lobsters from the Roscoff bay area did correspond to the previously described *P. gigantea* species. Cysts, that are mostly spherical but sometimes ovoid (Figure 3.1) had diameters ranging from 108 µm to 240 µm (mean 151.1± 45.3 µm), and enclose thousands of gymnospores (Figure 3.1), that are also mostly spherical, with diameters from less than 5 µm to almost 7 µm (mean 5.63± 0.08 µm). As already described in the literature, these gymnospores are composed of radially arranged zoites forming a monolayer with an optically void center (Figure 3.1). The observation of broken gymnospores (Figure 3.1) allowed measuring the length of their constitutive zoites (mean 1.04± 0.16 µm) as well as their apical width (mean 0.630± 0.129 µm).

Trophozoite stages, reported to be extremely long - up to 16mm in the initial report by Van Beneden (1869) - were indeed very thin and long in our hands, up to 2585 µm for a mean width 41.8± 10.4 µm (Figure 3.1). Moreover, and as described by several authors, their posterior end is slightly thinner, around 30 µm. The whole trophozoite

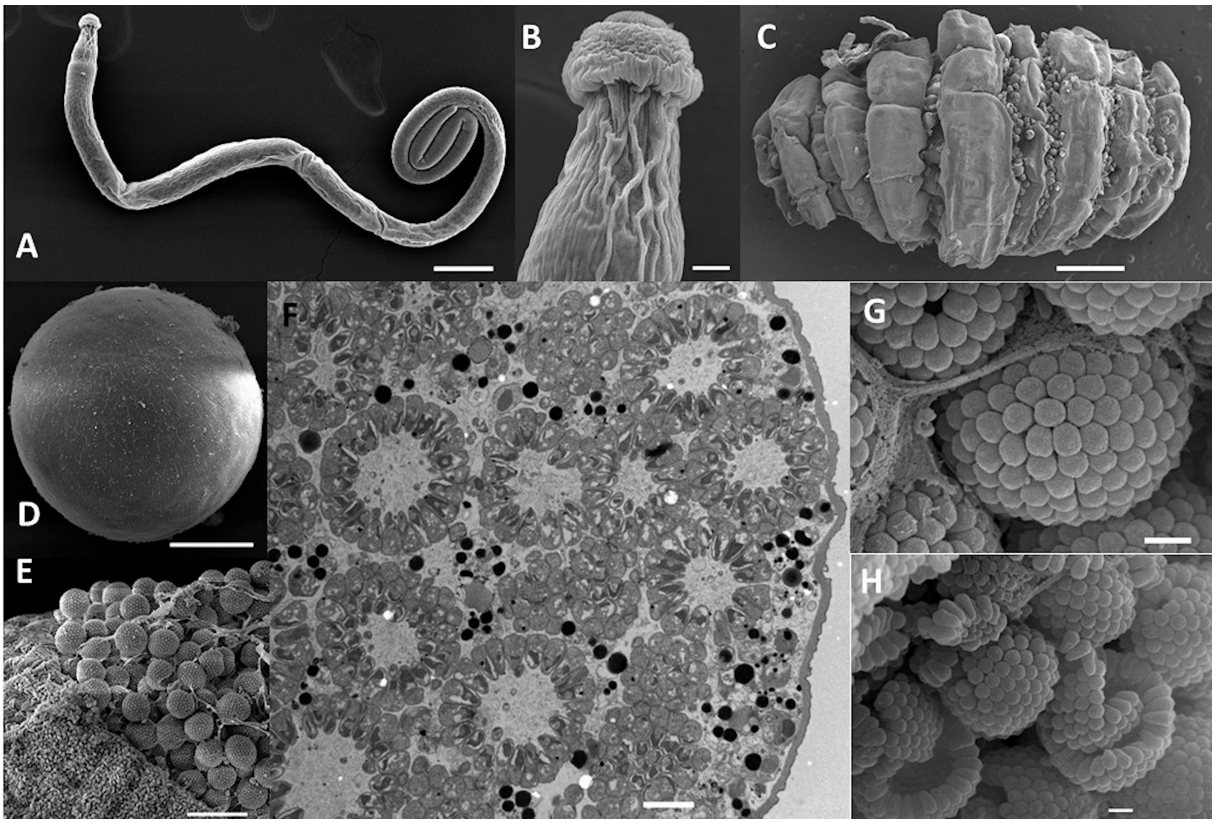


Figure 3.1: **Morphological characterization of *Porospora cf gigantea***. A. Trophozoite stage (Tropho #8, Lobster #12) (scale=100 µm). B. Zoom on A, showing trophozoite epimerite (scale=10 µm). C. Rectal ampulla showing cysts in folds (Lobster #4) (scale=1mm). D. Isolated cyst (Cyst #4, Lobster #12) (scale=50 µm). E. Broken cyst allowing to visualize enclosed, packed gymnospores (Lobster #4) (scale=10 µm). F. Section across a cyst illustrating radial arrangement of zoites in gymnospores (JS449=Lobster #35) (scale=2 µm). G., H. Zoom on intact and broken gymnospores allowing visualizing zoites (Lobster #4) (scale=1 µm). Scanning (A, B, C, D, E, G, H) and transmission (F) electronic microscopy. Microscopy analyses were performed by I. Florent, G. Prensier and S. Le Panse. In Boisard et al. (submitted).

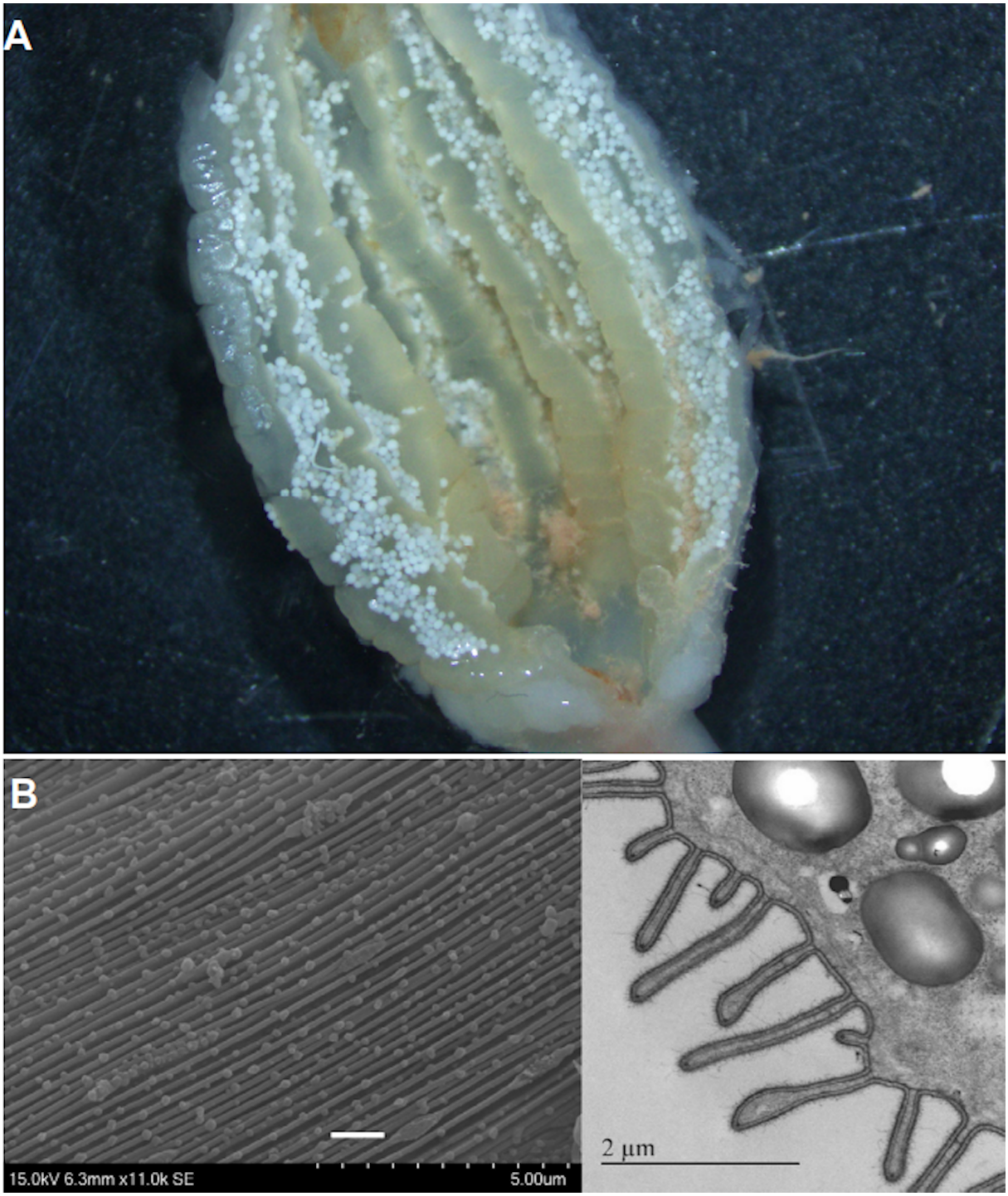


Figure 3.2: **Additional microscopy figures**. A. Photonic image of the rectal ampulla of Lobster #12, longitudinally opened and heavily packed with *Porospora cf. gigantea* cysts in chitinous folds. The length of the rectal ampulla is about 3 cm. B. Morphological evidence for epicytic folds. Zoom on epicytic folds for trophozoite #9, Lobster #12. Scale=1 µm. SEM imaging (left); TEM imaging (right). Microscopy analyses were performed by I. Florent, J. Schrével and S. Le Panse. In Boisard et al. (submitted).

surface is covered by longitudinal epicytic folds (Figure 3.2.B), that are reported necessary to allow eugregarine gliding movement (Valigurová et al., 2013). The sum of these morphological observations concerning tropohozoites, cysts, gymnospores and their zoites are all in perfect agreement with the species being *Porospora gigantea*, from the type host *Homarus gammarus* (Van Beneden, 1869; Desportes and Schrével, 2013). Although the microscopy images showed some morphological diversity in term of shape, color, size between individual parasite forms, it could either reflect different developmental stages or simply correspond to intraspecific morphological variation.

P. cf. gigantea A and B have different ribosomal loci

Ribosomal loci reconstruction has been carried out by I. Florent, L. Duval and A. Labat. See detailled results in Boisard et al. (submitted).

The complete ribosomal loci of the 2 genomes were unsuccessfully recovered in our assembled contigs but a small SSU rRNA was captured for JS-470 sample, the librairie containing only genome A. Using a combination of specific primers amplifications, initially based on Simdyanov et al. (2015) and Schrével et al. (2016) then in part redesigned, as well as *in silico* clusterings with the tool iSegWalker (Karadjian et al., 2016), we were able to fully reconstruct complete ribosomal loci covering: 18S-ITS1-5.8S-ITS2-28S (5977bp), for both A and B genomes. Both complete loci were experimentally confirmed totally (genome A) or partially (60%, genome B) by PCR. We found 30 polymorphic positions between A and B that were unevenly distributed i.e. only one position was polymorphic at the level of the 18S sequences, while the 29 remaining polymorphic sites were found at the level of the 28S.

The 18S SSU rDNA, which benefits from a largest taxonomic sampling for gregarines into the sequence databases, was used to construct two different phylogenies.

P. cf. gigantea A and B is the sister group of all other crustaceans gregarines

The first aim was to place *P. cf. gigantea* A and B within the gregarines and the other apicomplexans (Figure 3.3, page 87; detailed phylogeny is available in Figure 3.4, page 88). This phylogeny was constructed using 18S SSU rDNA as molecular marker and allowed us to assign them together to one clade (maximum support 100/1), and to position them as a sister group to all other crustacean gregarines (Cephaloidophoroidea, support 89/1) as established in Rueckert et al. (2011), with nevertheless shorter branch lengths, thus looking less derived than the other crustacean gregarine sequences.

We retrieved the superfamily Cephaloidophoroidea described by Rueckert et al. (2011), gathering all eugregarines isolated from the intestines of marine and freshwater crustaceans (*Cephaloidophora*, *Heliospora*, *Thiriotia*, and *Ganymedes* species). As the genus *Porospora* belongs to the family Porosporidae, the clade gathering the sequences of *P. cf. gigantea* A and B therefore represents the family Porosporidae. Thus, we consider that the sequences of *Thiriota* spp., which form a distinct clade in our phylogenetic study, belong to a new family, which we suggest to name Thiriotiidae, following the proposal in Desportes and Schrével (2013).

We have recovered, in this study, some of the previously published results for the main families of gregarines currently documented at the molecular level (Archigregarines, Actinocephaloidea, Cryptogregarinorida, Gregarinoidea, Lecudinoidea, Ancoroidea, (Schrével et al., 2016; Clopton, 2009; Simdyanov et al., 2017; Diakin et al., 2017)) as well as two *incertae sedis* clades gathering species previously described in Rueckert and Leander (2010); Iritani et al. (2018a,b), but as in these published studies, the relationships between these groups could not be resolved by using a single phylogenetic marker.

Contribution of environmental data

We also performed a phylogeny focused on crustacean gregarines including the sequence sampling published by Rueckert et al. (2011) and added environmental sequences that were deposited more recently in Genbank (Figure 3.5, page 89).

Thus, the vast majority of environmental clones are marine sediment-derived sequences

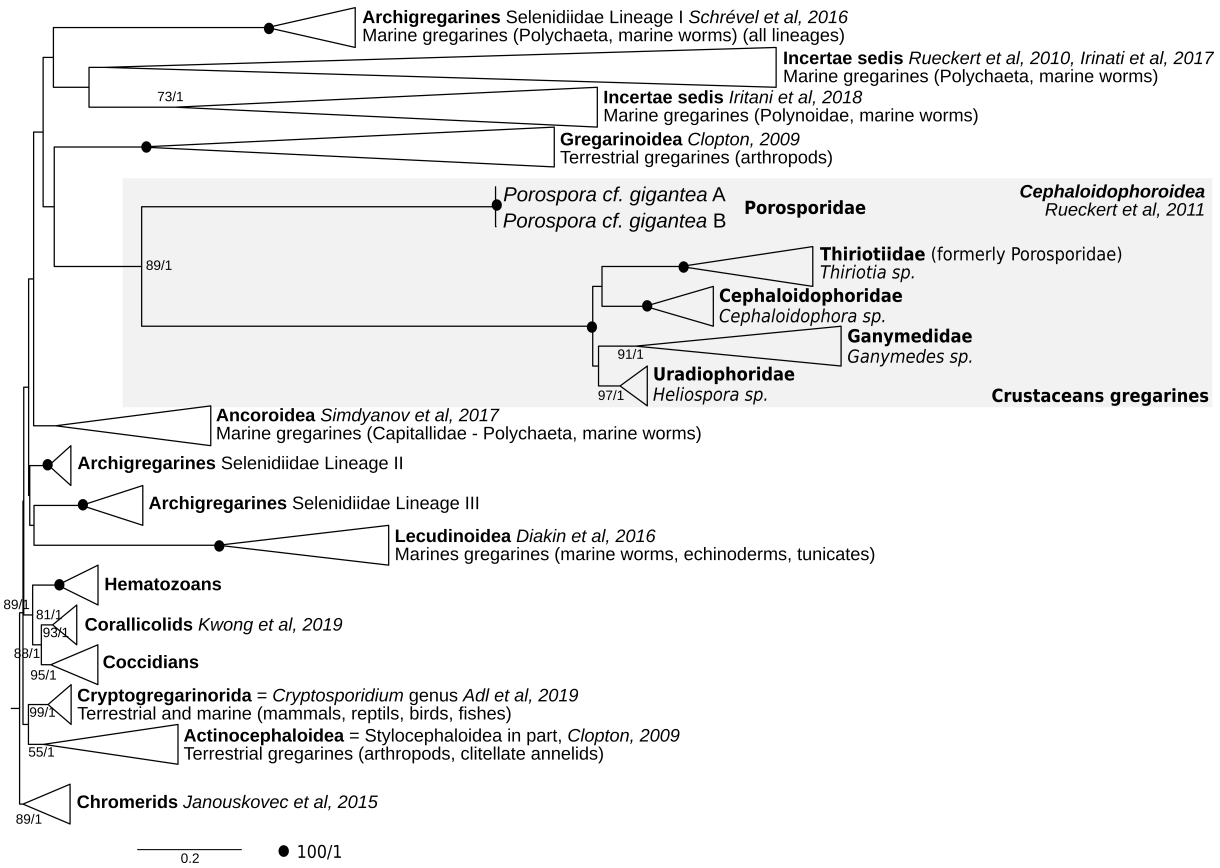


Figure 3.3: **Gregarines/apicomplexan phylogeny.** Phylogenetic tree built using 100 18S SSU rDNA sequences 1614 sites in order to situate the two *P. cf. gigantea* sequences among other known gregarines and apicomplexan clades. Chromerid sequences were used as outgroup, as they are considered the sister group of all other apicomplexans (Woo et al., 2015). Evolutionary history was inferred by Maximum Likelihood and Bayesian Inference using a GTR+G+I model. Topologies were identical according to both methods. Black spots indicate 100/1 supports. Supports <70/0.7 are not shown. Families and associated literature are indicated. Sub-trees have been collapsed at the family level in order to make the phylogeny more legible. The complete phylogeny is available in Figure 3.4, page 88. In Boisard et al. (submitted).

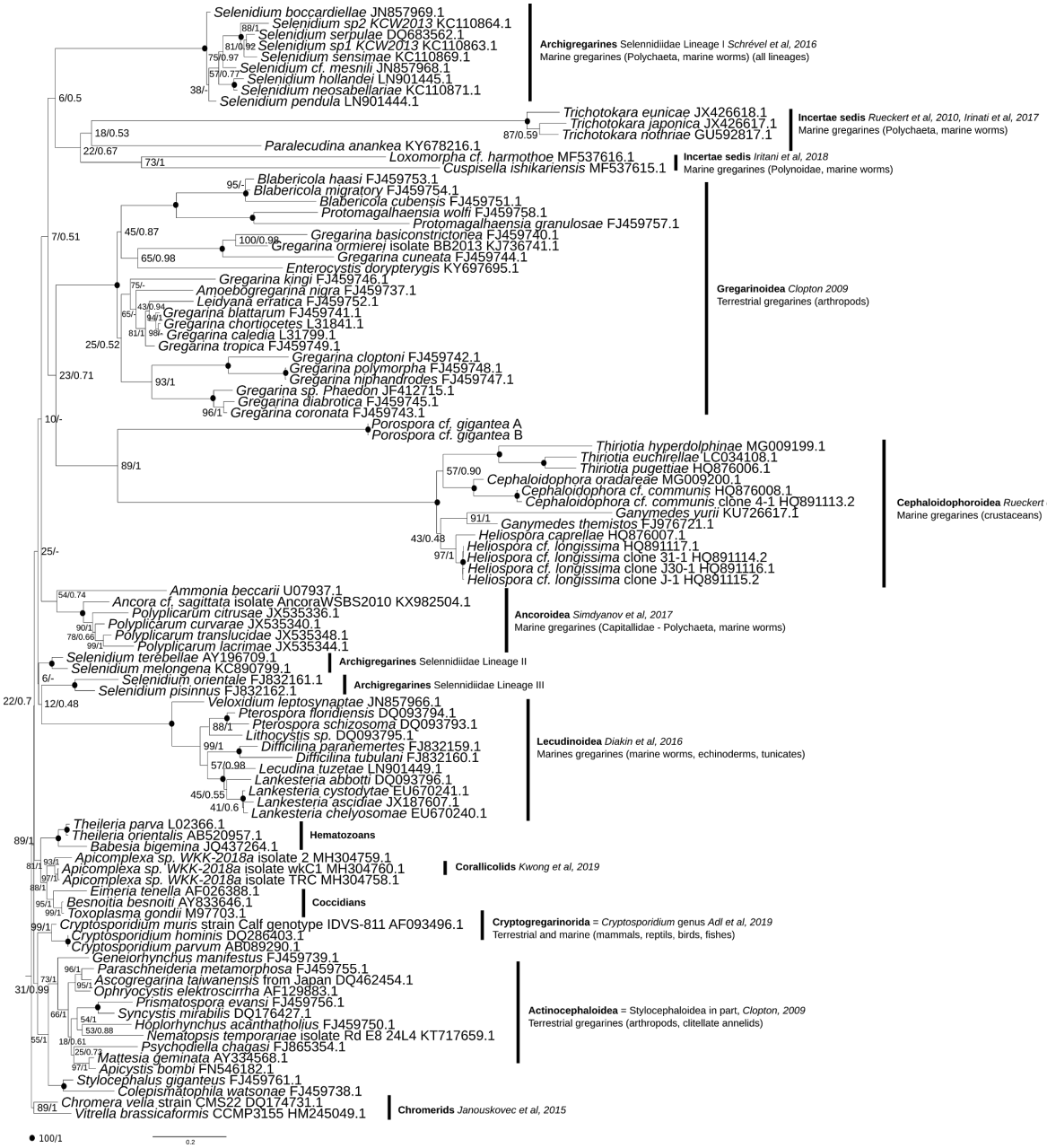


Figure 3.4: **Gregarines/apicomplexan phylogeny.** Phylogenetic tree built using 100 18S rDNA sequences 1614 sites in order to situate *P. cf. gigantea* A and B among other known gregarines and apicomplexan clades. Chromerid sequences were used as outgroup, as they are considered as the sister group of all other apicomplexans (Woo et al., 2015). Evolutionary history was inferred by Maximum Likelihood and Bayesian Inference using a GTR+G+I model. Topologies were identical according to both methods. Black spots indicate 100/1 supports. Supports <70/0.7 are not shown. Families and associated literature are indicated. In Boisard et al. (submitted).

Figure 3.5: **Cephaloiphoroidea environmental phylogeny.** Phylogenetic tree built using 189 18S rDNA sequences for 1135 sites in order to situate two *P. cf. gigantea* A and B among other crustacean gregarines and environmental sequences. Considering that Gregarinoidea sequences were placed as sister group of other crustaceans gregarines in the gregarines/apicomplexan phylogeny, as well as in recent literature (Mathur et al., 2019; Janouškovec et al., 2019; Mathur et al., 2021b), they were used as outgroup. Evolutionary history was inferred by Maximum Likelihood and Bayesian Inference using a GTR+G+I model. Topologies are identical according to both methods. Black spots indicate 100/1 supports. Supports <70/0.7 are not shown. Geographical provenance of all environmental sequences are indicated and their localization is highlighted in bold. In Boisard et al. (submitted).

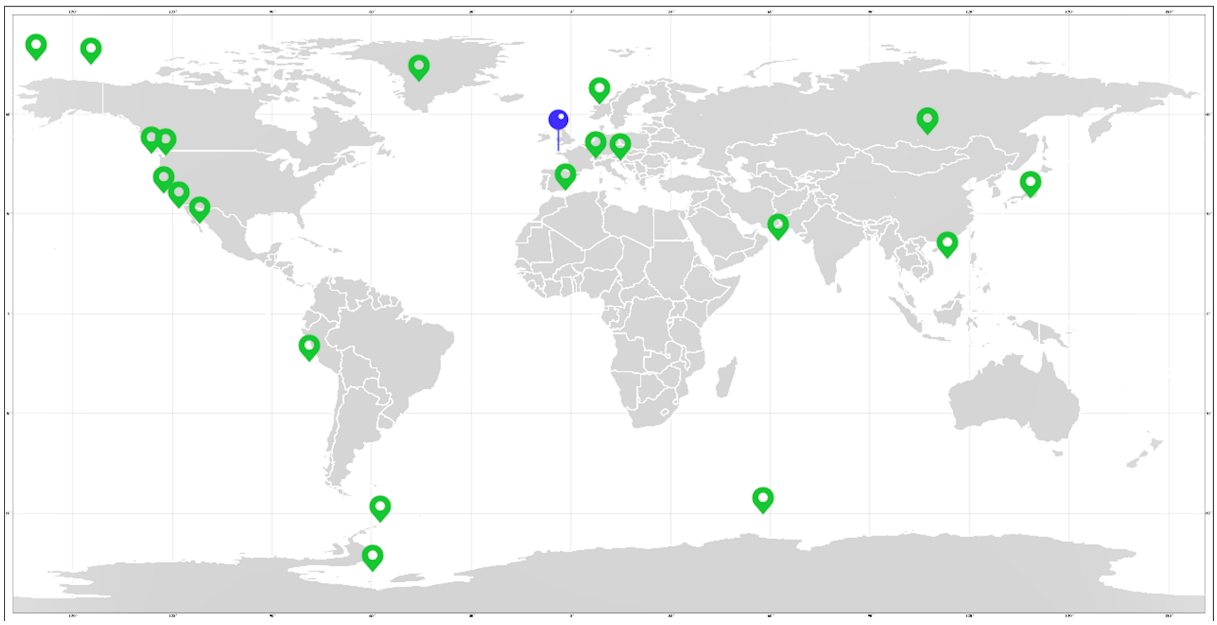


Figure 3.6: **Provenance of environmental sequences.** The sampling locations of the environmental sequences used in Figure 3.5 are shown in green on the map. The sampling location of the lobsters from which *P. gigantea* cysts and trophozoites were collected is shown in blue. Figure was made with the help of E. Duvernois-Berthet.

from a wide range of habitats, from intertidal to deep-water habitats; but we also recorded the presence of two sequences from freshwater river sediments. These environmental sequences are distributed in all five oceans including the Arctic Ocean, the Antarctic Ocean, the Pacific Ocean, the Indian Ocean, the Atlantic Ocean as well as the Mediterranean Sea and the China Sea. Concerning fresh waters, they are represented in Europe (Switzerland, Slovenia) and in Asia (Lake Baikal). However, we found only two sequences coming from the North Atlantic, corresponding to the distribution area of the European and the American lobsters, as displayed in Figure 3.6.

We found again within this environmental phylogeny the 5 main clades constituting the superfamily Cephaloiphoroidea, distributed as follows: on the one hand the 4 clades previously outlined in Rueckert et al. (2011) (redenominated as Ganymedidae, Cephalophoridae, Thiriotiidae and Uradiophoridae), and on the other hand, at their base, the clade Porosporidae, constituted of the two sequences of *P. cf. gigantea*, again displaying much shorter branch lengths. Finally, we noted the presence of a new putative clade formed by 5 environmental sequences from a Slovenian karst spring published by Mulec and Summers Engel (2019). This very well supported clade is placed as a sister group

to the 4 crustacean gregarines families Ganymedidae, Cephalophoridae, Thiriotiidae and Uradiophoridae, while the family Porosporidae retains its position as a sister group to all these other clades.

3.3.3 The demonstration and discussion of two *Porospora* species

Molecular data on *Porospora* show the presence of two species

We report in this study clear lobsters coinfection by two distinct organisms that we named *P. cf. gigantea* A and *P. cf. gigantea* B (*confer* indicates that both species correspond to the described species *P. gigantea*). At molecular level, these two organisms display highly similar genomes in terms of size, protein coding capacity, GC content and overall organization (86% synteny conservation), while displaying a 10.8% global divergence at nucleotide level.

The delineation of species now requires integrative morpho-molecular approaches, especially in protistology. Currently, the only molecular tool available for species discrimination in gregarines is the nucleotide sequence of the 18S SSU rDNA. At this molecular marker level, *P. cf. gigantea* A and *P. cf. gigantea* B differ by only a single nucleotide, a divergence level classically considered as indicative of organisms belonging to the same species.

Having only this molecular tool to discriminate taxonomically between *P. cf. gigantea* A and *P. cf. gigantea* B, we may have concluded to a single species, which we would have named *P. gigantea*, for its morphological similarities (at trophozoite, cysts, gymnospore, and zoite levels), host specificity (*H. gammarus*, the type species for *P. gigantea*) and localization of its developmental forms in the host (intestinal tract for trophozoites, rectal ampulla for cysts/gymnospores), accordingly to the available literature.

However, it is currently known that the 18S SSU rDNA marker alone is probably unable to reveal the real diversity of organisms, especially unicellular species, for which this highly conserved gene probably conceals many cryptic species (Piganeau et al., 2011).

Indeed, at the genomic level, both genomes show a nucleotide divergence of more than 10% which is incompatible with subspecies or strain definitions. By comparison, the same protocol applied for *P. falciparum* and *P. reichenowi* concluded at a divergence of only 3.2%. Moreover, a divergence of 3-5% has also been reported between the genomes of *Cryptosporidium parvum* and *C. hominis* (Guo et al., 2015).

This large overall divergence at genomic level indicates that *P. cf. gigantea* A and *P. cf. gigantea* B are probably not inter-fertile, and thus should be considered as different species, as the 'grey zone' of speciation spans from 0.5% to 2% of divergence, according to Roux et al. (2016) - although it must be kept in mind that this estimation is primarly relevant to metazoan.

Contribution of morphological and behavioral characteristics to the species delimitation

We could not evidenced significant morphological differences between *P. cf. gigantea* A and *P. cf. gigantea* B. Our morphological observations, if all pointing to the published features of *P. gigantea* (type host infected, host-compartments infected, morphology of developmental stages, including speed of gliding), remain insufficient to point to distinctive morphological features between *P. cf. gigantea A* and *P. cf. gigantea B*.

Genomic analyses demonstrated that Lobster #7 (isolate JS-470) is mainly infected by *P. cf. gigantea* A, while that Lobster #11 (isolate JS-482) and Lobster #12 (isolates JS-488 and JS-489) are co-infected by *P. cf. gigantea* A and *P. cf. gigantea* B. Currently, we have not been able to discriminate these two species at the morphological level. It is likely that this is partly due to the limited number of images we currently have for these lobsters' parasites. It would be profitable to carry out a more advanced imaging campaign in association with molecular analyses. For instance, using single cell genomics would allow to differentially sequence samples taken from developmental stages with morphological characters subject to variation, in order to evaluate if these variations are due to intra-specific polymorphism or if they can qualify as species discrimination markers.

We must also consider the possibility that despite these complementary analyses, we

may not be able to identify clearly discriminating morphological features. Indeed, it is a complex task to control the impact of parasite development and host environment in order to clearly identify morphological characters that would allow the unambiguous identification of *P. cf. gigantea* A or *P. cf. gigantea* B.

Geographical distribution and host living conditions

Although we do have collected, over this sampling campaign, additional images for 24 lobsters infected by *Porospora*, we can only speculate about the parasite species they harbor (i.e. *P. cf. gigantea* A and/or *P. cf. gigantea* B). We have two different hypothesis to tentatively explain why "Lobster bay" specimens (n=2) would be coinfected while a single "Lobster tank" specimen was infected only with by *P. cf. gigantea* A.

Indeed two parameters differentiate these two lobster sources. "Lobster bay" specimens are coming from Roscoff bay and have been living in the wild. "Lobster tank" specimen are in fact coming from South England, and have been raised in captivity. A first hypothese would be that both *P. cf. gigantea* A and *P. cf. gigantea* B are found in the wild (and both in Brittany and South of England) and *P. cf. gigantea* A only survives a long captivity of its lobster host. This hypothesis would be in agreement with Van Beneden's observations on the lesser infection of lobsters held in captivity (see page 82). Lower odds of infection by parasites are also documented and summarized in Milotic et al. (2020).

An other hypotheses consists in an uneven distribution of *P. cf. gigantea* A and *P. cf. gigantea* B in the wild. Both species would co-occur near Roscoff while only or at least mostly *P. cf. gigantea* A, would remain in the wild in the South of England.

To resolve these hypotheses, lobsters should be collected from the wild at different locations in the English Channel and through the Atlantic North Ocean to assess the geographic distribution of *P. cf. gigantea* A and *P. cf. gigantea* B. In parallel, further sampling of farmed lobsters from various tank facilities could confirm the presence of a single taxon under captive conditions.

For now, we maintain the proposal of *P. cf. gigantea* A and *P. cf. gigantea* B to name

the two organisms we found in *H. gammarus*, pending a more integrated morpho-molecular definition of their taxonomy, as well as a better documentation of Cephaloidophoroidea species.

Revision of the crustacean gregarines clade taxonomy

In their study about gregarines from crustacean hosts, Rueckert et al. (2011) described and named a new species, *Thiriotia pugettiae*, confirming the existence of the genus *Thiriotia*, following the proposal of Desportes et al. (1977) to rename *Porospora pisae* to *Thiriotia pisae* based on morphological criteria. Desportes et al. (1977) also suggested moving the genus *Thiriotia*, originally assigned to the family Porosporidae, to the family Ganymedidae.

However, this proposal was discarded by Rueckert et al. (2011), in the absence of sufficiently strong arguments to endorse this taxonomical change. Thus, the original taxonomy, that is to place the genus *Thiriotia* in the family Porosporidae Léger and Dubosq, 1911, was retained and thus confirmed Grassé's gregarine classification (Grassé, 1953), i.e. the existence of only four subgroups of gregarines infecting crustaceans.

Sequences from *Thiriotia*, as well as several environmental sequences, isolated from marine and estuarine sediments - thus suggesting decapod hosts - are brought together into a very well supported clade in Rueckert et al. (2011). However, in the absence of data for other decapod gregarines (including the genus *Porospora*), these sequences, while fully justifying not relocating *Thiriotia* to the family Ganymedidae, were also insufficient to definitively rule on their attribution to the family Porosporidae. In this sense, it was suggested to name a new family devolved to these sequences, named Thiriotiidae (Desportes and Schrével, 2013).

Indeed, our phylogeny retrieves the sequences of *P. cf. gigantea*, members of the Porosporidae (based on the morphological argument previously stated), as a sister group of the crustacean gregarines, including the clade containing the genus *Thiriotia*, still assigned to the family Porosporidae (Figure 3.3, page 87 and Figure 3.4, page 88). However, this attribution is rendered obsolete by our analysis, and again questions the family to

which the genus *Thiriotia* should be assigned. We propose to maintain the family name Porosporidae for the clade including the two species of *P. cf. gigantea*, as proposed by Labbé, 1899.

We find in our analysis, in agreement with Rueckert et al. (2011), the presence of a very well supported clade gathering the sequences of the genus *Thiriotia* and placing itself as a sister group of the family Cephaloidophoridae. Also, in light of this new evidence, we propose a new classification for the Cephaloidophoroidea superfamily, which would then consist of not four but five families: Cephaloidophoridae, Uradiophoridae, Ganymedidae, Porosporidae (consisting of sequences for *P. cf. gigantea*) and a fifth clade named Thiriotiidae, gathering sequences of the genus *Thiriotia*.

Perspectives on improving knowledge of crustacean gregarines

It should be noted that this distribution into five major subfamilies is also mirrored in our environmental phylogeny (Figure 3.5, page 89), which includes the environmental sequences currently associated with crustacean gregarines. We also found within this environmental phylogeny, the presence of a new clade consisting of the 5 environmental sequences published by Mulec and Summers Engel (2019), from a Slovenian Spring Karst.

These 5 sequences form a very well supported clade at the base of the clade gathering the families Cephaloidophoridae, Uradiophoridae, Ganymedidae and Thiriotiidae; the family Porosporidae remains basal to all the other clades, including this new environmental clade. Nevertheless, this clade being only constituted by environmental sequences, and in the absence of associated morphological description, it seems premature to make a new family of it. This should however draw our attention to the taxonomy of crustacean gregarines, which could evolve further, as more species are documented. Indeed, so far they are very few sequences from the North Atlantic Ocean, which corresponds to the area of distribution of European and American lobsters, both susceptible to be infected by gregarines of the genus *Porospora* (Figure 3.6, page 90).

Thus, new sampling campaigns would be necessary to better describe Porosporidae, for which we currently have no nearby environmental sequence. We must also recall

that the life cycle of *P. gigantea* is still very poorly known, and if the existence of sexual reproduction in a second mollusk host has been presumed (Desportes and Schrével, 2013), it is currently assumed that most of its reproduction is asexual and carried out entirely in conjunction with the life cycle of the lobster, re-infecting itself at each molt (Desportes and Schrével, 2013). The presence of cystic forms in marine sediments would consequently be much rarer.

Finally, compared to other crustacean-infecting gregarines for which 18S SSU rDNA data are available, the two species seem to present a more limited evolutionary distance (Figure 3.3, page 87, Figure 3.4, page 88 and Figure 3.5, page 89); nevertheless, these results need to be completed by a multi-protein analysis at the scale of the superfamily Cephaloidophoroidea, for instance by harnessing recently published transcriptomic data for Cephaloidophoroidea species (Mathur et al., 2019; Janouškovec et al., 2019; Mathur et al., 2021b).

3.4 The example of two locust-infecting gregarines

3.4.1 A conflictual historical taxonomy

Orthoptera (Ensifera - crickets and katydids and Caelifera - grasshoppers, groundhoppers and pygmy mole crickets) are reported to be parasitized by about 60 species assigned to the genus *Gregarina* Dufour, 1828 (see Desportes and Schrével (2013) for a recent, extensive review of the literature). Based on morphological descriptions, some gregarine species have been found to be restricted to one host family or superfamily, while others seem to have the capacity to infect a large range of hosts, widely distributed all over the world (Corbel, 1968b; Desportes and Schrével, 2013; Semans, 1941; Song et al., 2018).

Problems of identification based on morphological characters likely arose from phenotypic plasticity in response to wide-range host species and/or other contrasted environmental conditions. As a result, species delimitation within the genus *Gregarina* has been the subject of debate, with confusion, descriptions and synonymies, in particular for

gregarines that infect the Caelifera suborder, as illustrated below. Species delimitation is, however, a global and recurrent issue in protistology (Boenigk et al., 2012).

Léger (1893) described *Clepsidrina acridiorum*, which, a few years later, was termed *Gregarina* by Labbé (1899). This parasite was found in Acridoidea collected in Algeria (Léger, 1893). As the infected specimens belonged to different genera of Caelifera (*Truxalis, Pamphagus, Sphingonotus*), Léger concluded that "other acridians from Africa should be investigated for potential *G. acridiorum* infections" (Léger, 1893).

Interestingly, he noticed that *G. acridiorum* was not found in the desert locust, *Schistocerca gregaria* (Léger, 1893). Later in 1956, Canning described a gregarine she named *Gregarina garnhami*, sampled from this *S. gregaria* host (Canning, 1956). Interestingly, *G. garnhami* was also reported by the same author in both the migratory locust, *L. migratoria* and in the Egyptian locust, *Anacridium aegyptium* (Canning, 1956).

According to data in the literature, *G. acridiorum* and *G. garnhami* share common morphological and behavioral characteristics, such as their development in the midgut of their hosts, a small globular epimerite, stout bodied gamonts, and barrel-shaped (or dolioform) oocysts (Canning, 1956; Léger, 1892; Lipa and Santiago-Alvarez, 1996). Lipa and Santiago-Alvarez (1996) concluded that the species described in 1956 by Canning in *S. gregaria* was in fact *G. acridiorum*.

This interpretation was supported by the fact that in 1956, Canning had not been aware of the existence of *G. acridiorum* (Lipa and Santiago-Alvarez, 1996). *Gregarina acridiorum* has been reported in a range of Orthoptera hosts (Ensifera and Caelifera: Acrididae, Tetrigidae) including *L. migratoria* and *A. aegyptium* (Corbel, 1967; Lipa and Santiago-Alvarez, 1996), two species also described as hosts of *G. garnhami* (Canning, 1956). Consequently, the two acridian species could be infected by the two gregarines species.

Gregarina acridiorum and *Gregarina garnhami* also closely resemble *Gregarina rigida* (Hall, 1907) Ellis, 1913, described in a broad range of widely distributed orthopteran hosts (Desportes and Schrével, 2013) and also similar to *Gregarina ronderosi*, a parasite of the argentine grasshopper *Dichroplus elongatus* (Lange and Wittenstein, 2002). The devel-

opmental and morphological similarities of these four gregarines question their species definition as well as their hosts specificities and indeed, based on morphological considerations, Corbel (1967) even proposed that *G. rigida* and *G. acridiorum* were synonym species. For a very detailed illustration of those differences and how tenuous they can be among five gregarines of acridians, see Table 1 in Florent et al. (2021), that lists the main biological and morphological characters of these four very similar gregarines of acridians, including also the data concerning a fifth species, *Gregarina caledia* (*nomen nudum*), a parasite of the Australian grasshopper *Caledia captiva*, described in the PhD Thesis of Jennifer Ann Ninham (1995) and reported to be very similar to *G. garnhami*.

The limited availability of DNA sequences corresponding to these species is an obstacle to the resolution of these controversies (only partial SSU rDNA sequences (1210 bp) for *G. caledia* (L31799) and *Gregarina chortiocetes* (L31841)). This latter species, intestinal parasite of *Chortiocetes terminifera* is however poorly described at morphological level (Ninham, 1995).

In 2002, Lange and Wittenstein (2002) indicated that: "given the great similarity of *Gregarina* spp. associated to Acrididae, it would be very informative to study at the molecular level, most species as possible". In this purpose, we combined morphological and molecular data to better explore species boundaries of gregarines infecting two orthoptera acridians hosts, *S. gregaria* (Forskål, 1775) and *L. migratoria* (Linné, 1758). These two hosts are locusts, i.e. grasshoppers that can form dense migrating swarms, often destructive to agriculture, through an extreme form of density-dependent phenotypic plasticity, known as phase polyphenism (Ayali, 2019; Uvarov, 1977). Here we sought to determine whether they are infected by the same or distinct gregarine species, as the information in the current literature is not congruent (Canning, 1956; Corbel, 1967, 1968a,b; Labbé, 1899; Léger, 1893; Lipa and Santiago-Alvarez, 1996).

Morphological observations of developmental stages of gregarines from *L. migratoria* and two subspecies of *S. gregaria* were performed and completed with the sequencing of their SSU rDNA loci. The results herein reported reveal clear molecular differences at this marker's level, despite highly similar morphological features, strongly supporting

that these two acridian hosts are not infected by the same gregarine species. Some subtle differences could also be established between these two species at morphological level.

3.4.2 Morphological and molecular descriptions' confrontation

Morphological description of *G. acridiorum* and *G. garnhami*

Morphological description of G. acridiorum and G. garnhami has been carried out by I. Florent, while statistical comparison of morphological measurements have been realized by myself; for the sake of clarity, the main results of morphological description are reported below. See detailed results in Florent et al. (2021).

Gregarines isolated from the intestinal tracts of various acridian *S. gregaria* and *L. migratoria* host specimens (Table 3.2, page 77) were mostly located between the host intestine epithelial cells and digested food material. In addition, in all *S. gregaria* specimens, young trophozoite stages were invariably observed in the host's ceca, whereas this was never observed in *L. migratoria*. Occasionally, gametocysts were also isolated from insect feces and kept at room temperature to observe dehiscence.

Gregarines infecting *Schistocerca gregaria*. The observed stages were trophozoites, solitary gamonts, gamonts associated in caudo-frontal syzygies, and gametocysts enclosing oocysts or emitting them as chains through sporoducts (Figure 3.7, page 100).

Young trophozoite stages (also referred to as cephalonts in historical publications (Canning, 1956; Desportes and Schrével, 2013; Ninham, 1995)) (Figure 3.7.A) were observed in the two subspecies, regardless of the geographical location/raising facilities (Table 3.2). The globular epimerite with a short neck was visible in their anterior part (Figure 3.7.A).

The density of infections could be very high, as shown by the number of trophozoites attached to the gut epithelium of an *S. g. gregaria* host from Morocco (Figure 3.7.B). The epimerite of attached trophozoites was enclosed in the host epithelial cell (Figure 3.7.C). High densities of trophozoites were also found in the ceca (data not shown) and midgut (solitary gamonts and syzygies (Figure 3.7.D)). The protomerite of trophozoites and gamonts was oval or slightly conical (Figure 3.7.A-D); in syzygies, it appeared to be flattened at the top of the satellite with a ridge formed during pairing with the primite

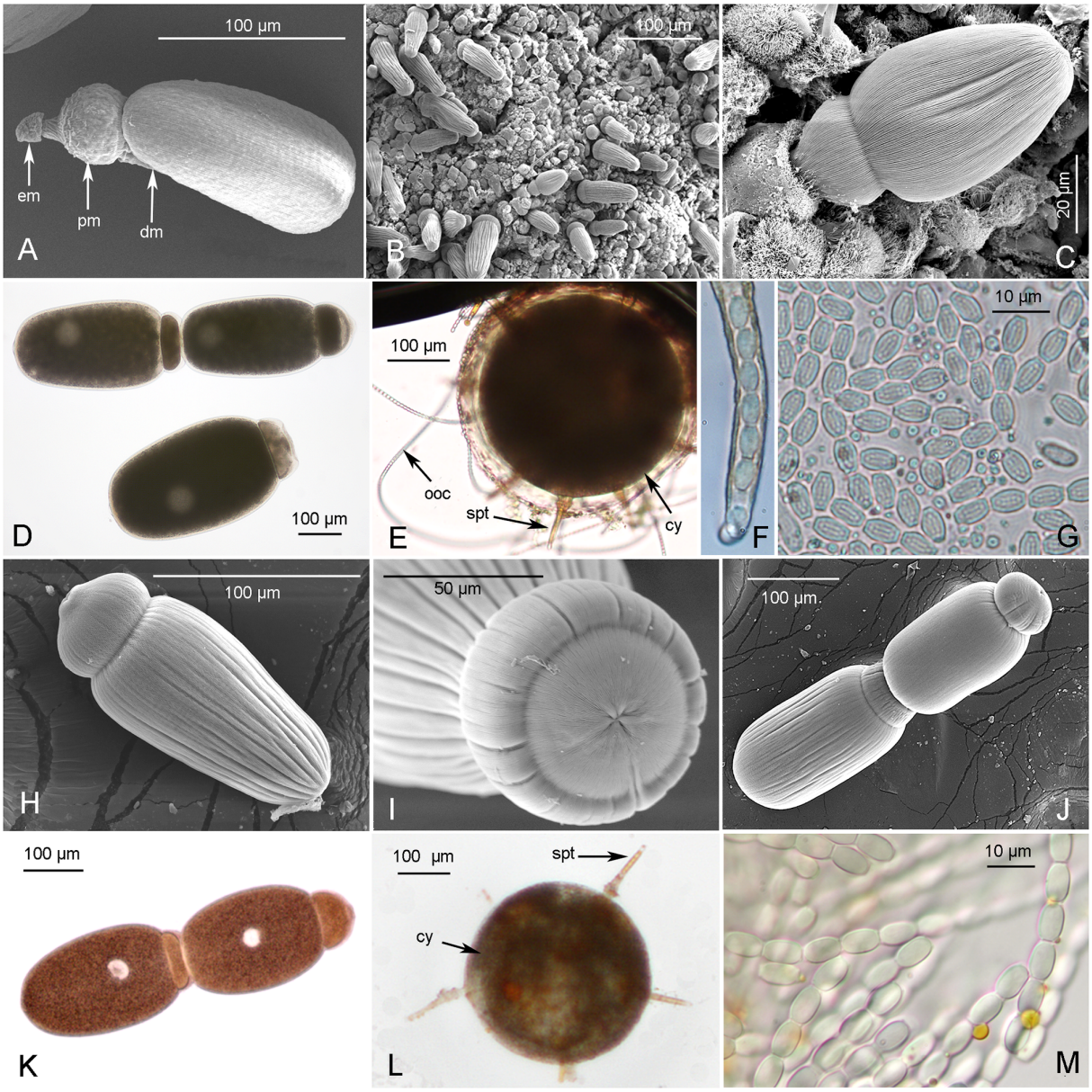


Figure 3.7: **Scanning Electron Microscopy (A–C, H–J) and photonic imaging (D–G, K–M) of gregarines infecting *S. gregaria* (A–G) and *L. migratoria* (H–M).** *S. gregaria* gregarines: A, young trophozoite (epimerite (em) protomerite (pm) and deutomerite (dm)), (South Africa); B, intestinal tract infected by numerous gregarines (Morocco); C, gregarine encased in an intestinal host cell, enlargement of B (Morocco); D. Solitary gamont and syzygy (Belgium); E. Gametocyst form (cy) with developed sporoducts (spt) releasing oocyst chains (ooc); F, zoom on sporoduct extremity showing enclosed oocysts; G. released oocysts. *L. migratoria* gregarines: H, solitary gamont detached from intestinal host cell; I. zoom on gamont protomerite; J–K, gamonts associated in syzygies; L, Gametocyst form (cy) with developed sporoducts (spt); M. released oocysts. Scales are given for each figure. Microscopy images were performed by I.Florent and B. Michel, in Florent et al. (2021).

(Figure 3.7.D). Scanning electron microscopy revealed a similar ridge at the top of the satellite in *G. garnhami* syzygies (Valigurová and Koudela, 2008).

The deutomerite was cylindrical or ovoid, and quite stocky in older trophozoites and syzygies (Figure 3.7.D). A constriction of the septum was visible between the posterior part of the protomerite and the anterior part of the deutomerite (Figure 3.7.D). The nucleus was seen in the opaque endocyte of the deutomerite. Longitudinal epicytic folds were visible at the surface of these trophozoite/gamont stages (Figure 3.7.A-C). Thickening of the ectocyte was visible above the endocyte at the apex of the primite protomerite (Figure 3.7.D).

The gamonts in *S. g. flaviventris* from South Africa (length, L = 402 ± 79 µm; width, W = 172 ± 42 µm, n = 27) were very similar in size to gamonts in *S. g. gregaria* from Belgium (L = 410 ± 53 µm, W = 200 ± 39 µm, n = 34), but slightly smaller in *S. g. gregaria* from Morocco (L = 332 ± 43 µm, W = 96 ± 16 µm, n = 4). Moreover, smaller and much thinner trophozoites were observed in the latter (L = 192 ± 15 µm, W = 34 ± 4 µm, n = 12) (Figure 3.7.A). Also interestingly, gamonts in *S. g. gregaria* from Belgium were much stockier (L/W = 2.1 ± 0.2 µm) than gamonts in *S. g. flaviventris* from South Africa (L/W = 2.4 ± 0.3 µm) and gamonts (L/W = 3.5 ± 0.2 µm) and trophozoites (L/W = 5.8 ± 1.0 µm) in *S. g. gregaria* from Morocco.

The ratios of protomerite (P) to deutomerite (D) lengths were, however, similar for gamonts, regardless of the infected hosts (P/D = 0.25 ± 0.04 µm (South Africa, n = 27); P/D = 0.23 ± 0.06 µm (Belgium, n = 34); P/D = 0.23 ± 0.07 µm (Morocco, n = 4), and also for the thinner trophozoites found in Moroccan *S. g. gregaria* specimens (P/D = 0.26 ± 0.04 µm, n = 12). Overall, for trophozoites and gamonts infecting these hosts, regardless of the subspecies and their geographical location, the values were: L = 370 ± 98 µm; W = 159 ± 69 µm; L/W = 2.83 ± 0.38 µm (n = 77).

Gametocysts in dehiscence were observed, producing ∼ 8 (but sometimes more) pale orange basal discs, circular cellular structures with a central opening that eventually developed across the mucilaginous layer (ectocyst) into sporoducts with swollen bases (Figure 3.7.E). Their length was ∼ 1/3 that of the diameter of the gametocyst (Figure

3.7.E). Gametocysts diameters were 350 ± 56 µm, n = 36 (from 210 to 420 µm). Oocysts extruding as chains through these sporoducts (Figure 3.7.F) were barrel-shaped with a thick wall enclosing eight sporozoites (Figure 3.7.G). Their size was quite uniform (L = 6.54 ± 0.32 µm, W = 4.32 ± 0.23 µm, n = 89) (Figure 3.7.G).

Gregarines infecting *Locusta migratoria*. Trophozoite stages attached to the gut epithelium of hosts were not seen, but a scar remained visible where the epimerite had been present at the top of the protomerite of detached gamonts (Figure 3.7.H-I). These gamonts were rather cylindrical with a sub-globular protomerite, flattened and slightly constricted at the proto-deutomerite septum (Figure 3.7.I-K).

The deutomerite was much longer and more slender towards the posterior end (Figure 3.7.H). The size of the gamonts varied but the mean size (L = 219 ± 48 µm, W = 93 ± 30 µm, n = 37) was smaller than the mean size observed in *S. gregaria* specimens (see above). Gamonts were also quite stocky (L/W = 2.5 ± 0.6 µm, n = 37).

In caudo-frontal syzygies, the protomerite was sub-globular in the primite, but shorter and flattened with a circular anterior edge in the satellite (Figure 3.7.J-K). The deutomerite was cylindrical, slightly wider in the anterior part (Figure 3.7.J), ovoid in syzygies (Figure 3.7.K), with a rounded posterior end. The spherical nucleus could be seen in the opaque cytoplasm (endocyte) of the deutomerite (Figure 3.7.K).

Longitudinal epicytic folds were seen at the surface of these stages (Figure 3.7.H-J). The length of syzygies was (L = 456 ± 73 µm, W = 93 ± 30 µm, n = 16) in our studies. The ratio of protomerite (P) deuteromerite (D) lengths was 1/4 (P/D = 0.25 ± 0.05 µm, n = 21).

Gametocysts were spherical with a mucilaginous layer (ectocyst). Under this layer, and as observed in gregarines that infect *S. gregaria*, basal discs of the future sporoducts differentiated at the surface of encysted gametocysts. These basal discs were also orange with a central white aperture, but were fewer in number (< 8, n = 15). Like in the case of gregarines that infect *S. gregaria*, chains of oocysts were extruded through sporoducts (Figure 3.7.L-M) whose length in gregarines of *L. migratoria* is longer and represents ∼ 1/2 the diameter of the gametocyst (Figure 3.7.L).

Gametocysts diameters were 227 ± 35 µm, n = 18 (from 190 to 296 µm). Oocysts, that were also emitted as chains from sporoducts, were also barrel-shaped with a double wall but were slightly longer and slimmer (L = 6.83 ± 0.27 µm, W = 3.99 ± 0.19 µm, n = 40, Figure 3.7.M) than the oocysts emitted by gregarines that infect *S. gregaria* (Figure 3.7.F-G).

Statistical comparison of morphological measurements. For the gamonts, the means of the lengths (p-value = 2.2e-16; df = 111.97) and of the widths (p-value = 8.574e-11; df = 111.13) were significantly different between the gregarines infecting *S. gregaria* and *L. migratoria*. However, there were no significant differences between the length/width ratios between these two groups. Concerning the gametocysts diameters, the mean was significantly different (p-value = 1.986e-13; df = 49.386). Finally, for the oocysts, both mean length (p-value = 6.664e-07; df = 89.407) and mean width (p-value = 5.722e-13; df = 88.967) were significantly different.

G. acridiorum and *G. garnhami* are distinct species

SSU rDNA sequences. To further characterize these gregarines, a molecular study was designed to sequence most of the SSU rDNA locus from gamonts and gametocysts, isolated from several host specimens belonging to *L. migratoria* and two subspecies of *S. gregaria*. A total of 23 sequences were generated from gregarines found in 7 specimens of *L. migratoria* on three collection dates, and 20 sequences were generated from gregarines found in five specimens of *S. gregaria* from a total of three geographical origins and/or raising facilities. Regardless of the subspecies and the geographical location of hosts and their maintenance facilities, all the gregarines isolated from *S. gregaria* specimens shared the same *type 1* sequence (1638bp long), presumably corresponding to *G. garnhami*, whereas all the gregarines isolated from *L. migratoria* specimens presented a clearly distinct *type 2* sequence (1637bp long), presumably corresponding to *G. acridiorum*. Multiple sequence alignment and distance analyses were performed to qualify intra-species and inter-species variations, and clearly revealed two distinct clusters.

Within the sequence group of gregarines from the host *S. gregaria*, the mean level of

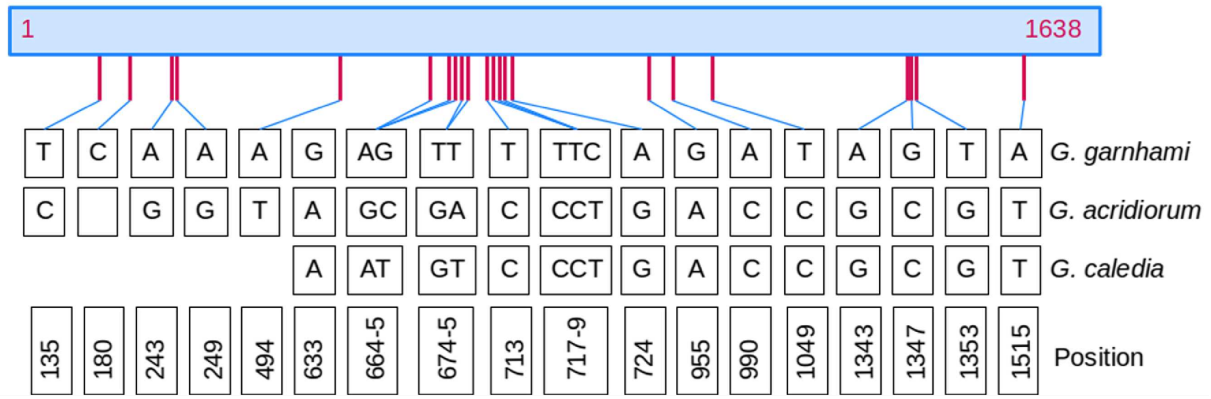


Figure 3.8: **Assessment of acridian SSU rDNA sequences divergence.** Distribution of the 22 polymorphic positions in SSU rDNA locus regions V1-V8 (1638bp), between *type 1* (presumably *G. garnhami*) (n = 20) and *type 2* (presumably *G. acridiorum*) (n = 23) sequences, amplified from gregarines parasitizing respectively *S. gregaria* and *L. migratoria*. The corresponding positions are also given for *G. caledia* (L31799, 1210 bp) parasitizing *Caledia captiva*. Eleven additional positions, otherwise conserved between *G. garnhami* and *G. acridiorum* sequences, are modified in *G. caledia* sequence: site 1059, G deletion; sites 1161-1164: GAGC substituted by AG-G; site 1181: G substituted for C; site 1187: G substituted for A; sites 1231 and 1240: T substituted for C; site 1493: T insertion; site 1584: G substituted for A. In Florent et al. (2021).

divergence was 0.2%, whereas within the sequence group of gregarines from the host *L. migratoria*, the mean level of divergence was 0.3%. The mean level of genetic distance between gregarine sequences from *S. gregaria* and those from *L. migratoria* was 1.5%, whereas the genetic divergence from *G. caledia*, parasite of *C. captiva*, was 1.1% with the gregarine group from *L. migratoria*, but 2.2% with the gregarine group from *S. gregaria*. In all, 22 conserved polymorphic positions, rather evenly distributed along the SSU rDNA locus, were identified between *type 1* and *type 2* sequences (assumed to be *G. garnhami* and *G. acridiorum*, respectively), as schematized in Figure 3.8.

Phylogenetic analysis. A phylogenetic approach, using partial SSU rDNA sequences and both maximum likelihood and Bayesian inference reconstructions, indicated that gregarine sequences from the two different host species studied clustered with sequences from other Gregarinoidea species (as described in Cavalier-Smith (2014); Clopton (2009); Schrével et al. (2016)) with a high ML bootstrap value and Bayesian posterior probability (Figure 3.9, page 106).

These novel gregarine sequences form two clearly distinct clades according to their

host species, and it thus appears that all *S. gregaria* hosts, regardless of their subspecies and the geographical location at which they were maintained, were infected by the same species (based on their SSU rDNA sequence) that was clearly distinct from the parasitic species infecting *L. migratoria*.

The SSU rDNA sequence from *G. caledia* showed closer affinity to gregarine sequences from the host *L. migratoria* than from the host *S. gregaria* (see also Figure 3.8). Furthermore, we observed that hosts of the *type 2* (presumably *G. acridiorum*) and *G. caledia* sequences, i.e. *L. migratoria* and *C. captiva*, belong to the same clade B of the acridian phylogeny as defined by Song et al. (2018), while *S. gregaria*, infested with *G. garnhami* (*type 1* sequences), belongs to a distinct clade D, as indicated in Figure 3.9. Thus, gregarine distribution appears to parallel the taxonomy of these three hosts. This observation will however need to be confirmed, as the ML bootstrap support remains low (55), despite high Bayesian posterior probability (Figure 3.9).

3.4.3 Two different gregarines for two different locusts

Molecular data are able to discriminate gregarine taxa more accurately

To determine whether the acridian orthopterans *S. gregaria* and *L. migratoria* are infected by the same gregarine species, their parasites were isolated and morphological and molecular analyses were performed using a series of host specimens of both species collected from a range of different locations and insect raising facilities (Table 3.2, page 77).

While morphological investigations confirmed highly similar parasites with only tenuous behavioral and quantitative morphological differences, molecular investigations yielded unambiguous results strongly supporting different gregarine species in these *S. gregaria* and *L. migratoria* hosts.

Molecular characterization, based on the partial SSU rDNA marker (V1–V8 region (Hadziavdic et al., 2014)) of all gregarines studied, unambiguously demonstrated that all *S. gregaria* hosts – regardless of their subspecies and raising facilities – are infected by the same gregarine species (presumably *G. garnhami*), whereas all *L. migratoria* hosts

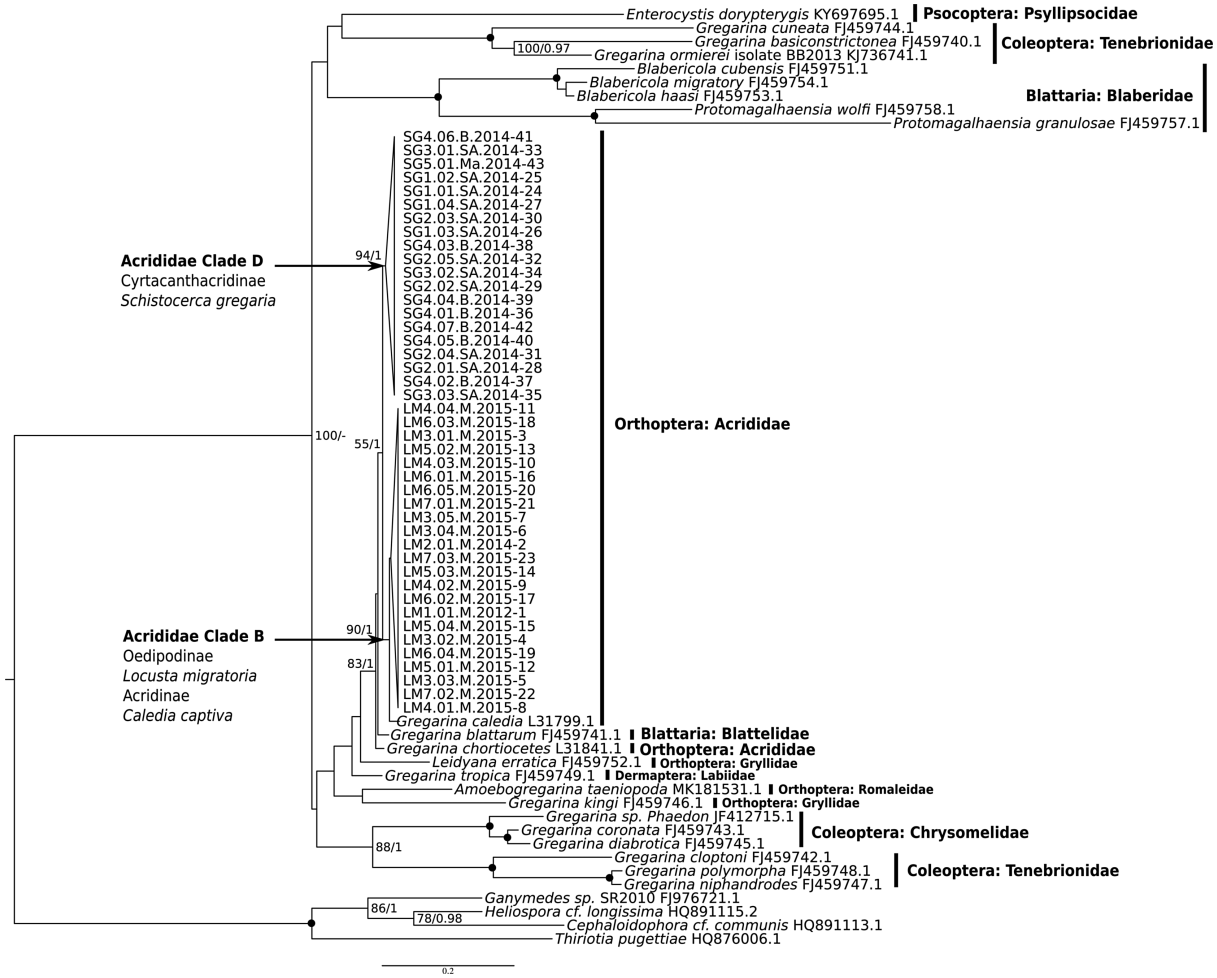


Figure 3.9: **Gregarinoidea phylogeny**. Phylogenetic tree built using 69 SSU rDNA sequences for 1,433 sites in order to zoom in on the clade Gregarinoidea including gregarines parasites of Orthoptera (Clopton, 2009). Outgroup consists of 4 sequences from Cephaloidophoroidea species that infect crustaceans, currently considered as the sister group of Gregarinoidea (Medina-Durán et al., 2019). Evolutionary history is inferred by maximum likelihood and Bayesian inference using a GTR substitution model with gamma-distributed rate variation across sites plus invariant sites. Maximum likelihood topology is shown, with supports from both methods. Bootstrap $< 75\%$ and posterior probabilities < 0.95 are not shown. Black spots indicate 100/1 supports. The gregarines infecting *L. migratoria* clustered with *G. caledia*, isolated from the grasshopper *Caledia captiva* (Ninham, 1995), the gregarines infecting *S. gregaria* forming a distinct independent clade. *G. chortiocetes*, infecting the locust *Chortiocetes terminifera* (Ninham, 1995), and *Gregarina blattarum*, infecting the cockroach *Blatella germanica* (Clopton, 2009) form sister branches to this group. The taxonomy of locust hosts is indicated, as established by Song et al. (2018). In Florent et al. (2021).

are infected by a distinct species (presumably *G. acridiorum*).

Both gregarine sequences clustered in the previously reported Gregarinoidea clade (Clopton, 2009; Medina-Durán et al., 2019; Nocciolini et al., 2018). Overall, 22 different bases were identified in this 1638bp region that could be used to delimit the species. The 1.5% genetic distance between the two sequences is in agreement with previously described inter-specific levels of genetic divergence that, for example, distinguish *Gregarina niphandrodes* from *Gregarina polymorpha* (1.44%) (Nocciolini et al., 2018). However concerning those latter species, it should be noted that, according to the same authors, such low genetic divergence could also correspond to intra-specific variability (Nocciolini et al., 2018).

Based on these molecular results and on data in the literature, notably the identification of their hosts, we propose that the *type 1* sequence found in gregarines infecting *S. gregaria* hosts may correspond to the species named *G. garnhami*, reported by several authors and collected from *S. gregaria* (Canning, 1956) (Valigurová and Koudela, 2008). The gregarine species found in *L. migratoria* likely corresponds to *G. acridiorum*, in agreement with Léger (1893), but not with the proposal of Lipa and Santiago-Alvarez (1996).

Acridian's gregarines remaining to be discovered

The gregarine developmental stages described in *S. gregaria* and *L. migratoria* hosts are very similar morphologically, and share many characteristics including the thick mucilaginous ectocyst of the gametocyst, orange basal discs associated with great variability of size parameters.

As these morphological features have also been observed in other species, particularly in *G. rigida*, *G. ronderosi* and *G. caledia* collected from different (and sometimes from identical) orthopteran hosts (Table 3.2, page 77), these species need to be further characterized at the molecular level to solve their phylogenetic relationships. The only molecular sequence available (*G. caledia*, L31799) although rather small (1210bp) strongly suggests a third distinct species, closely related phylogenetically to the proposed *G. acridiorum*

but still with some observed genetic distance (1.1%). *Gregarina caledia* is also potentially morphologically distinguishable by the larger size of its oocysts and its ability to infect host ceca (Table 3.2, page 77) (Ninham, 1995). Importantly though, in the first morphological reports, this species was said to be closely related to *G. garnhami* with which it also shares the ability to infect host ceca (Ninham, 1995).

Gregarina rigida (Hall, 1907) Ellis, 1913, has also been reported in a range of orthopterans. When describing this species, the authors did not cite any literature on *G. acridiorum*, so, Corbel (1968b) concluded that *G. rigida* was a junior synonym of *G. acridiorum*. To be confirmed, the status of this species (e.g. synonym of *G. acridiorum*) therefore requires molecular data, even though available measurements of oocysts and the fact that it has also been found in host ceca favor a distinct species.

Importantly, in 2002, *G. ronderosi*, which was found in the Argentine grasshopper *Dichroplus elongatus*, was named a novel species by Lange and Wittenstein (2002) due to the lack of infection in specimens of *L. migratoria* experimentally infected with this gregarine. It thus also possibly corresponds to a fifth distinct species, also awaiting molecular characterization. Lange and Wittenstein (2002) even suggested that *G. ronderosi* could be synonymous with *G. garnhami*, but that molecular data were required as morphometric differences did not enable conclusive delimitation of the species.

3.5 Integrative taxonomy is essential for assessing gregarine diversity

Assigning protist species can no longer rely on morphological information alone, but should include molecular data in an integrated taxonomic approach (Bernays, 1981; Boenigk et al., 2012). The data presented here confirm that most morphological and morphometric differences cannot conclusively delimit closely related species, while molecular data can reveal clearly measurable differences.

In our first example, concerning *P. cf. gigantea* A and B, we could not evidence morphological characters able to discriminate unambiguously the two species, although

they are clearly distinguishable at the molecular level. Indeed, we demonstrated that despite an almost identical SSU rDNA sequence (1bp difference for 1702 positions, i.e. ∼0.05% divergence), they displayed ∼10% overall nucleic acid divergence at the genomic level, potentially preventing genetic crossing, thus arguing for different species.

On the other hand, our second study demonstrates that *S. gregaria* is infected by *G. garnhami*, whereas *L. migratoria* is infected with *G. acridiorum*. In this case, we have been able to highlight two discriminating features: the respective size of the oocysts of *G. garnhami* and *G. acridiorum*, but also the location of their trophozoite forms in their respective host's gut. However, these characteristics retain some ambiguity (potential phenotypic plasticity), and while they support the molecular data, it is the latter that allow us to highlight precise and measurable differences at the level of the 18S molecular marker.

In this respect, additional molecular studies are crucial to determine the diversity of gregarine species that infect acridians, beyond the establishment of morphological specificities. A major challenge concerns the precise diversity of the species *G. acridiorum* that has been described in over 60 orthopteran hosts, from both the Caelifera and Ensifera orders, as is also the case for *G. rigida*. It is likely that these two species correspond to a much greater diversity of probably cryptic species that remain to be described by this type of integrative taxonomical approach, in the diversity of their currently described hosts.

Nevertheless, we must keep in mind that these studies cannot be satisfied with the search for a single molecular marker, such as the 18S marker, the only one currently available for a range of gregarines. We are now aware that this marker, in addition to being poorly suited for the delimitation of species within protists (Pawlowski et al., 2012, 2016), is not sufficient to capture the extent of genetic divergence between two species. Here, our study of the genomes of *P. cf. gigantea* demonstrates the inadequacy of this single marker to rule on species delimitation, and that a genomic scale approach is likely required to discriminate cryptic species, which appear to be numerous within gregarines, as shown by the example of Orthoptera gregarines.

Chapter 4

P. cf. gigantea genomes' new insights on apicomplexan evolution

4.1 Looking through *P. cf. gigantea* genomes hidden knowledge

After reporting the origins of the two genomes *P. cf. gigantea* A and *P. cf. gigantea* B in Chapter 2, and described in Chapter 3 the integrative taxonomy analysis that allowed us to describe the gregarine species corresponding to them in the most accurate way possible at the present time, this chapter is devoted to the detailed description and discussion of these two marine gregarine genomes and how they contribute to enrich our knowledge of Apicomplexa.

In a first part of results, the structural characteristics of the two genomes are presented, with reference to a selection of apicomplexan genomes from a variety of species. The main issue here is to show the specificities of these two marine gregarine genomes with respect to what is currently known about apicomplexan genomes, and particularly by comparing them to the genome of *G. niphandrodes*, the genome of a terrestrial gregarine whose unpublished data are however deposited in VEupathDB, CryptoDB section (Aurrecoechea et al., 2017). These two marine gregarine genomes, the first to be actually deciphered and hopefully soon published for gregarines as a whole, will be able

to serve as a reference for this group in all future analyses focused on the comparative genomics of apicomplexan. Indeed, while current knowledge on apicomplexan genomes is heavily biased towards pathogenic *genera* and species (mainly *T. gondii*, *Plasmodium* spp., *Cryptosporidium* spp.), it is essential to document the neglected basal apicomplexan lineages.

As demonstrated in the previous chapter, the biodiversity of gregarines is likely to be greatly underestimated, as the large number of morphologically described species (Desportes and Schrével, 2013; Portman and Šlapeta, 2014) likely contains many cryptic species, detectable only at the molecular level, including genomic level. It is therefore probable that the study of gregarine genomes also reveals an unsuspected diversity, capable of shedding light on the evolutionary history of apicomplexes, but also to better understand the mechanisms at work at the molecular level that give them the necessary adaptive capacities for the variety of their parasitic lifestyle.

One of the main characteristics of apicomplexans is their ability to travel through the tissues of their hosts until reaching their targeted cell, either extra or epi-cellularly attaching to it or completely invading it intracellularly. This is why we decided to dedicate the first functional analysis of a key structure of apicomplexan parasites to the data mining of glideosome proteins, presented in a second part. The glideosome model is a complex molecular structure sustaining a movement called gliding, the emblematic motility of the Apicomplexa group, essential to the manifestation of their pathogenicity. While this structure is currently well described in *T. gondii* and *P. falciparum*, we know nothing about the proteins involved in gliding in gregarines, which in addition have the peculiarity of exhibiting other types of movements like bending or rolling. Thus, we decided to study in detail the proteins involved in this molecular structure, in order to evaluate their conservation at the apicomplexan scale on the one hand, but also and especially to question the relevance of the glideosome model applied to the gregarines, and in default, to imagine possible alternatives in relation to what we know about the specific biology of gregarines.

4.2 Methods

4.2.1 Automatic annotation of coding and non-coding genes

The completeness of the gene predictions performed in Chapter 2, section 2.2.5, page 49 was assessed by using BUSCO version 4.0.6 (Seppey et al., 2019).

The predicted proteins, have been automatically annotated by using i) the best hit of a BLASTP search against VEupathdb version 2019-20-01 (Aurrecoechea et al., 2017), ii) the results of KoFamScam against the KEGG pathway database version 2019-05-11 (Aramaki et al., 2020) and iii) the signature domains obtained with Interproscan version 5.39-77.0 (Jones et al., 2014)).

The Infernal sofware version 1.3.3 (Nawrocki and Eddy, 2013) and the Rfam database version 14.2 (Kalvari et al., 2018) were used together to search for transfer RNAs, spliceosomal RNAs and ribosomal RNAs. The snoReport software version 2 (de Araujo Oliveira et al., 2016) was used to search C/D and H/ACA small nucleolar RNAs. Predictions of non coding genes were done by L. Ponger.

All assembly metrics were assessed using QUAST (version 5.0) (Gurevich et al., 2013).

4.2.2 Orthology and dating

The ortholog groups were identified by using orthoMCL version 2.0.9 with default parameters (Li et al., 2003)) (analyse done by L.Ponger) applied to the proteome of a selection of representative organisms available on VEuPathDB (see Chapter 1 , Table 1.1, page 32).

Orthogroups were visualized using the UpSetR package (Conway et al., 2017).

The divergence time of genome A and genome B were calculated by L. Ponger following methods described in Boisard et al. (submitted).

4.2.3 Expert annotation for glideosome proteins

A 37 reference apicomplexan glideosome proteins dataset was elaborated based on glideosome protein repertoires described in the literature mainly for *T. gondii* and *P.*

falciparum (Boucher and Bosch, 2015; Jacot et al., 2016; Frénal et al., 2017).

This reference dataset was used as a seed for parsing the orthogroups made for 25 reference proteomes (see Chapter 1, Table 1.1, page 32) and the proteomes of the two *P. cf. gigantea* genomes.

For each orthogroup containing at least one of the reference proteins, the list of proteins was extracted and the protein sequences were recovered, as well as their respective coding nucleic sequences for both *P. cf. gigantea* genomes. A BLASTP was performed for extracted proteins against the proteomes of *P. cf. gigantea*, as well as a BLASTP of the candidate proteins for each *P. cf. gigantea* genome against the 25 species reference proteomes. A BLASTN was performed against NCBI NR for the coding sequences of the candidate proteins of both *P. cf. gigantea* genomes.

The sequences thus collected for each described protein were aligned with mafft (Katoh and Standley, 2013). Maximum likelihood molecular phylogeny was performed on each alignment using RAxML software (Stamatakis, 2014). 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 (see 4.2.1, page 113) and structure analyzed with SMART (Letunic et al., 2021).

For each protein, all the performed analyses were then manually examined to validate the candidate proteins within the proteomes of the two *P. cf. gigantea* genomes. Presence/absence table of glideosome proteins was visualized using the tidyverse R package (Wickham, 2009) and edited using Inkscape (www.inkscape.org).

4.2.4 Search for TRAP like proteins

The identification of TRAP-like proteins was done by searching for the TSP1 molecular domain (IPR000884) within the two *P. cf. gigantea* proteomes. The structure of each candidate protein was then carefully studied. If necessary, partially predicted proteins were re-edited with Genewise (Birney, 2004). Schematic representation of TRAP-like proteins was done using BioRender (biorender.com).

4.3 First investigations into *P. cf. gigantea* A and B genomes

4.3.1 *P. cf. gigantea* genomes characteristics and completeness assessment

A total of 10,631 putative genes were predicted on the raw assembly, which could be splitted into two sets of similar size: 5270 genes in the genome A vs. 5361 genes in the genome B (see Chapter 2, Figure 2.3, page 56). All relevant metrics are presented in Table 4.1 and compared with 6 others species, including two "proto-apicomplexan" chromerids: *G. niphandrodes* Unknown, *C. parvum* Iowa II, *T. gondii* ME49, *P. falciparum* 3D7, *C. velia* CCMP2878 and *V. brassicaformis* CCMP3155.

In addition to having a similar number of coding genes, the two *P. cf. gigantea* genomes also have nearly identical %GC, average coding sequence length, number of tRNAs and rRNAs, and intron profile (Table 4.1).

The proportion of coding sequences (84%) in *P. cf. gigantea* A and B genomes is particularly high compared to other reference species (from 25% to 76%; Table 4.1).

The completeness of both A and B genomes/proteomes was addressed by using the BUSCO software. BUSCO searches for "core" genes that should be conserved in all species belonging to a specific taxon. We used the geneset Apicomplexa (n=446). Genomes A and B respectively showed a completeness score of 70% (n=312) and 67.7% (n=302). Theses percentages are lower than those found for the genome of the gregarine *G. niphandrodes* (83%) and the 24 genomes of other representative species we evaluated (from 76.9% for *C. suis* to 100% for *P. falciparum*; all BUSCOs assessments are available in Figure 4.1, page 116).

4.3.2 Comparison of orthogroups within apicomplexan

Orthologues were searched between both A and B genomes. The proteins of *P. cf. gigantea* A and B were splitted into 5656 orthogroups including 4443 (88%) groups with

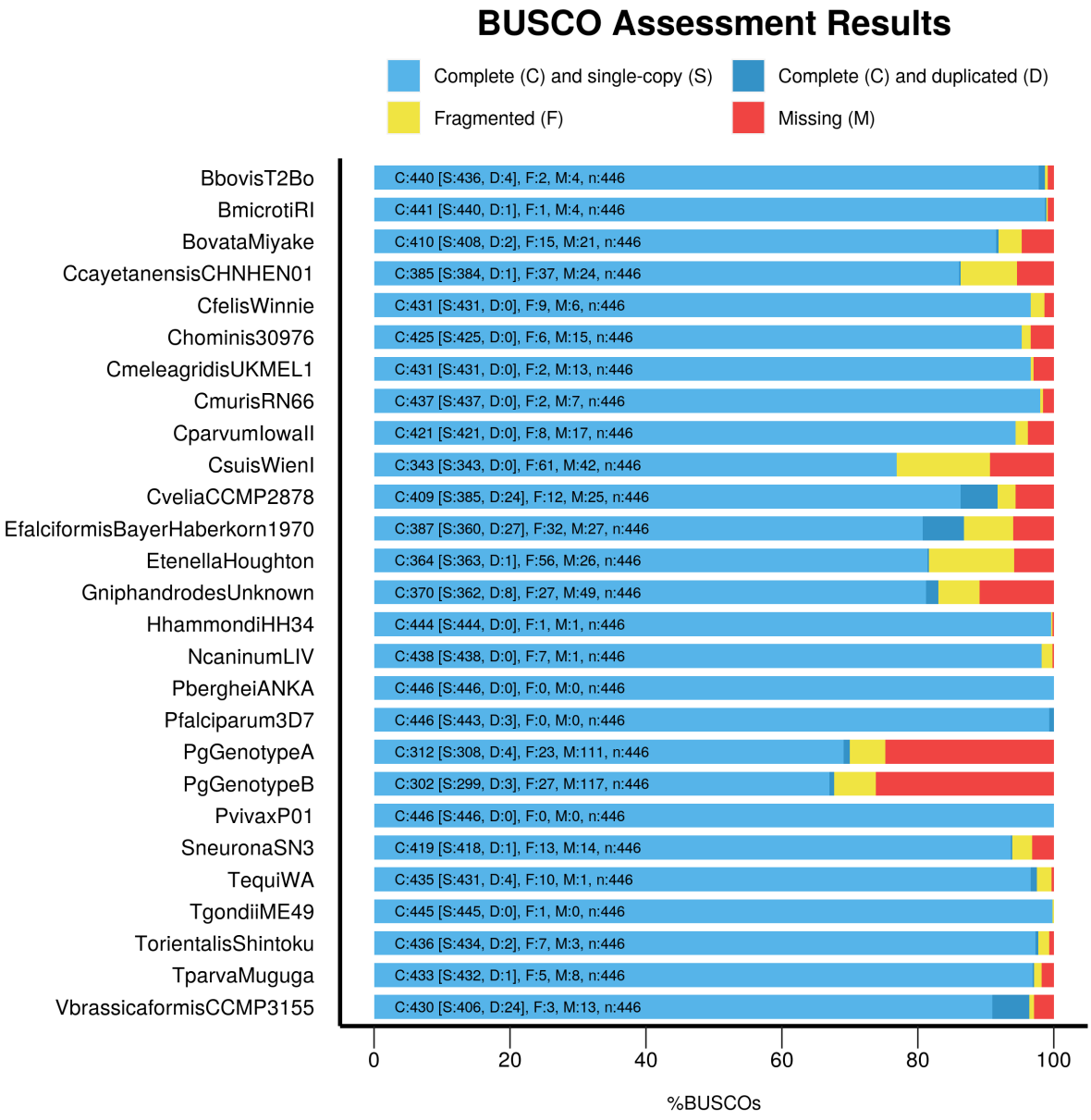


Figure 4.1: **BUSCOs assessment results for the proteomes of both *P. cf. gigantea* and a selection of 25 reference species** (geneset apicomplexa odb10). In Boisard et al. (submitted).

at least one orthologous gene for both A and B. As a comparison, this percentage is larger than the percentage observed between all the 6 reference species which are evolutionary divergent species (maximum of shared genes: 22% between *T. gondii* and *P. falciparum*). Less divergent species can be found within the genus *Plasmodium* and can be used for comparison. Indeed, the percentage of common orthogroups observed between genomes A and B is higher than the percentage of common orthogroups observed between *P. falciparum* and *P. berghei* (70%) which are documented to have diverged around 33 Mya ago (TimeTree, (Kumar et al., 2017)) and is similar to the percentage observed for the comparison of *P. falciparum* and *P. reichenowi* (86%, 3.3 - 7.7 Mya (TimeTree)). Using the hypothesis of similar substitution rates in gregarines and in *Plasmodium* species, we dated the split of genomes A and B between 15.5 Mya and 37.7 Mya (analyses done by L. Ponger, in Boisard et al., submitted). As a comparison, this dating is very distant from the emergence of Nephropidae (lobster group) which is estimated to ∼180 Mya (Crandall et al., 2009; Bracken-Grissom et al., 2014); however, this order of magnitude is similar to the basal split estimation for the mammal *Plasmodium* in Ricklefs and Outlaw (2010) (12.8 Mya) or all *Plasmodium* in Hayashida et al. (2012) (21.0–29.3 Mya).

The genomes A and B were also compared with those of the 6 reference species. The percentages of shared orthogroups between *P. cf. gigantea* genomes with each of the reference apicomplexan genomes are similar despite a highly variable divergence (*C. parvum*, 18%; *G. niphandrodes*, 17%; *P. falciparum*, 14%; *T. gondii*, 14%) but is higher than the percentages observed with chromerid species (*C. velia*, 8%; *V. brassicaformis*, 10%). We can underline from this result that the *P. cf. gigantea* genomes don't share significantly more orthogroups with *G. niphandrodes*, the only other gregarine for which a genome is available.

A global distribution of orthogroups between the genomes of *P. cf. gigantea* A and B and 4 reference apicomplexan species is also shown on Figure 4.2, page 118. A particularly striking result is the small number of orthogroups thus corresponding to the "apicomplexan core"; only 881 orthogroups bring together the 6 species. Thus, each species has a majority of unique genes that are not found in other apicomplexans.

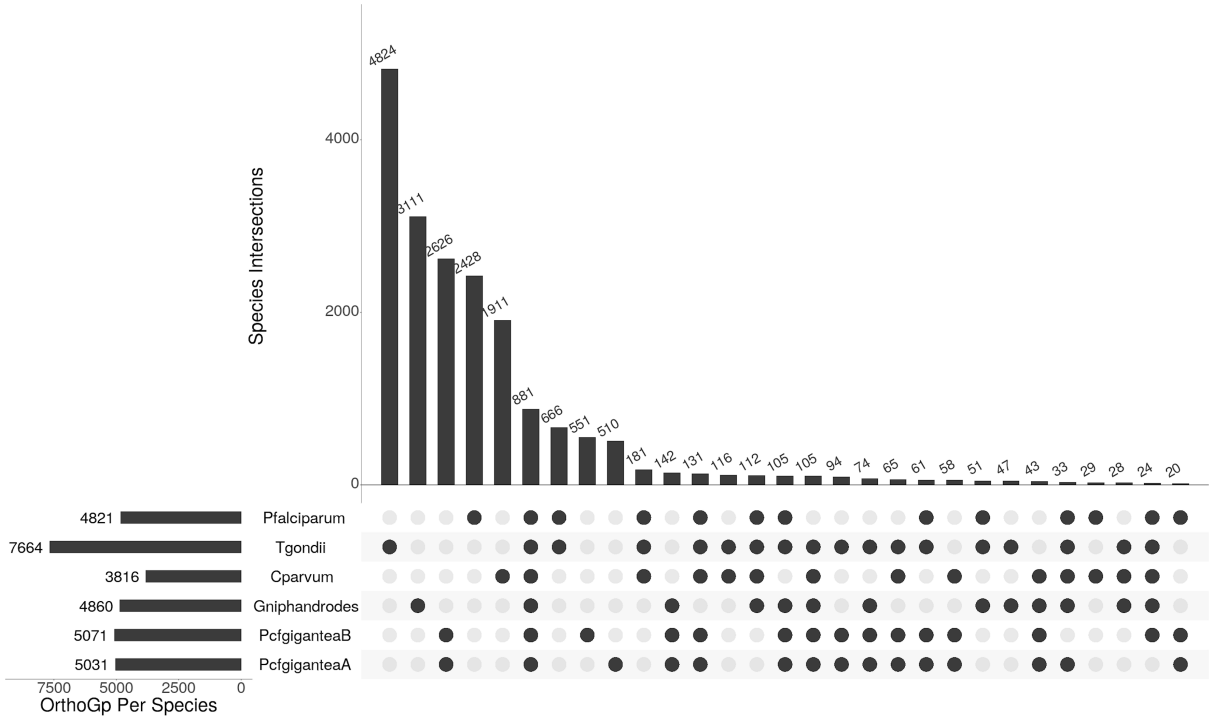


Figure 4.2: **Shared apicomplexan proteins.** Distribution of the orthogroups among *P. cf. gigantea* A and B and 4 species of apicomplexans: the gregarine *G. niphandrodes*, the cryptosporidian *C. parvum*, the coccidian *T. gondii* and the hematozoan *P. falciparum*. Only bars with more than 20 orthogroups are shown. Upset plot was made with UpSetR. In Boisard et al. (submitted).

Moreover, where one would have expected a greater number of genes common to the three available gregarine species, there are only 142 orthogroups shared only by them. In contrast, *G. niphandrodes* alone has a majority of unique genes (3111/4860), as do the two species of *P. cf. gigantea* (AB: 2626/∼5000). It is also noted that despite their close proximity, *P. cf. gigantea* A and *P. cf. gigantea* B each have about 500 unique genes.

4.3.3 The conservation of glideosome proteins in Apicomplexan

Gliding description in *P. cf. gigantea*

Dynamic recording of isolated trophozoites performing gliding were done by I. Florent and J. Schrével and were analyzed with I. Florent as detailed in Boisard et al. (submitted).

These recordings allowed us to confirm that they move uni-directionally, protomerite ahead, following straight or curved lines, depending on the observed individuals. The

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

Table 4.1: **Metrics of the genomes of *P. cf. gigantea* and a selection of 6 reference species.** In Boisard et al. (submitted).

whole body (deutomerite) follows the trace initiated by the apical part of the trophozoite (protomerite).

The speed of trophozoites displacement has been estimated to be ∼60 µm per second, as initially observed by King and Sleep (2005) and up to more than 100 µm per second in some recordings.

Glideosome description

In apicomplexan parasites, the glideosome refers to a set of proteins involved in gliding (sliding movement generated by the action of an actin-myosin motor) necessary for the motility of the zoite and its invasion and egress of/from the host cell and has been mostly described for species displaying intracellular life modes (Opitz and Soldati, 2002; Keeley and Soldati, 2004). Recently several articles have brought together the latest results of studies about the different components of this machinery and have updated the different understandings of this molecular architecture (Boucher and Bosch, 2015; Tardieux and Baum, 2016; Frénal et al., 2017).

As it is currently understood, the glideosome machinery relies on the mechanical forces generated by the actin-myosin motor that allows the parasite to move and then interact with the host cell, to which it attaches via a variety of adhesins proteins that are brought to the surface of the parasite by micronemes secretions. This approach to motility and invasion is known as the "capping model", as gliding is thought to occurs through the "capping" of membrane or transmembrane proteins during this process (Russell, 1983; King, 1988; Sibley et al., 1998).

I have conducted an inventory of the presence/absence of glideosome proteins currently described in the literature and mainly depicted in *T. gondii* and *P. falciparum*, which I have grouped according to their function, following the classification established by Frénal et al. (2017): actin dynamics, glideosome core proteins, adhesins, mobile junction proteins and host interaction regulators.

These references proteins from *T. gondii* and *P. falciparum* have been searched in both *Porospora cf. gigantea* genomes as well as in a selection of species including *Chromera*

velia, *Vitrella brassicaformis*, *Gregarina niphandrodes*, *Cryptosporidium parvum*, *Hammondia hammondi*, *Emeiria falciformis*, *Theileria parva* and *Babesia bovis* (Figure 4.3, page 122 and Table 4.2, page 123).

Actin and associated factors

Actin in apicomplexan is characterized by a globular monomeric form (G-actin) which polymerizes as needed into short unstable filaments (F-actin) (Skillman et al., 2011) using various supporting regulators such as profilin (Plattner et al., 2008; Pino et al., 2012; Skillman et al., 2012), ADF cofilin (Mehta and Sibley, 2011), formin (Tosetti et al., 2019; Daher et al., 2010; Baum et al., 2008) and Cpβ (Ganter et al., 2009).

It has been shown in *T. gondii* and *P. falciparum* that inactivation of actin or its associated regulators compromises motility, invasion and egress, although motility may persist in an altered form for a few days which raises the question of an alternative mechanism involved in parasite motility (Valigurová et al., 2013; Drewry and Sibley, 2015; Egarter et al., 2014; Whitelaw et al., 2017). With the exception of profilin in *G. niphandrodes* and Cpβ in piroplasmas, all the described proteins were found in all of the examined species.

Glideosome apicomplexan-specific proteins

The glideosome machinery itself is composed of specialized apicomplexan-specific proteins that have been described in Frénal et al. (2017) to which we refer the reader for more detailed information. The single-headed short heavy chain myosin class XIV, myosin A, acts as a motor generating the rearward traction required for motility, invasion and egress, as evidenced by various conditional depletion protocols in *T. gondii* and *Plasmodium* species (Meissner et al., 2002; Frénal et al., 2014; Siden-Kiamos et al., 2011).

The glideosome itself takes place between the plasma membrane and the apicomplexan-specific trilamellar membrane (IMC: Inner Membrane Complex), in which Myosin A is associated with a light chain (myosin light chain 1 – MLC1 in *T. gondii* (Herm-Gotz, 2002) and MyoA tail domain-interacting protein - MTIP in *P. falciparum*

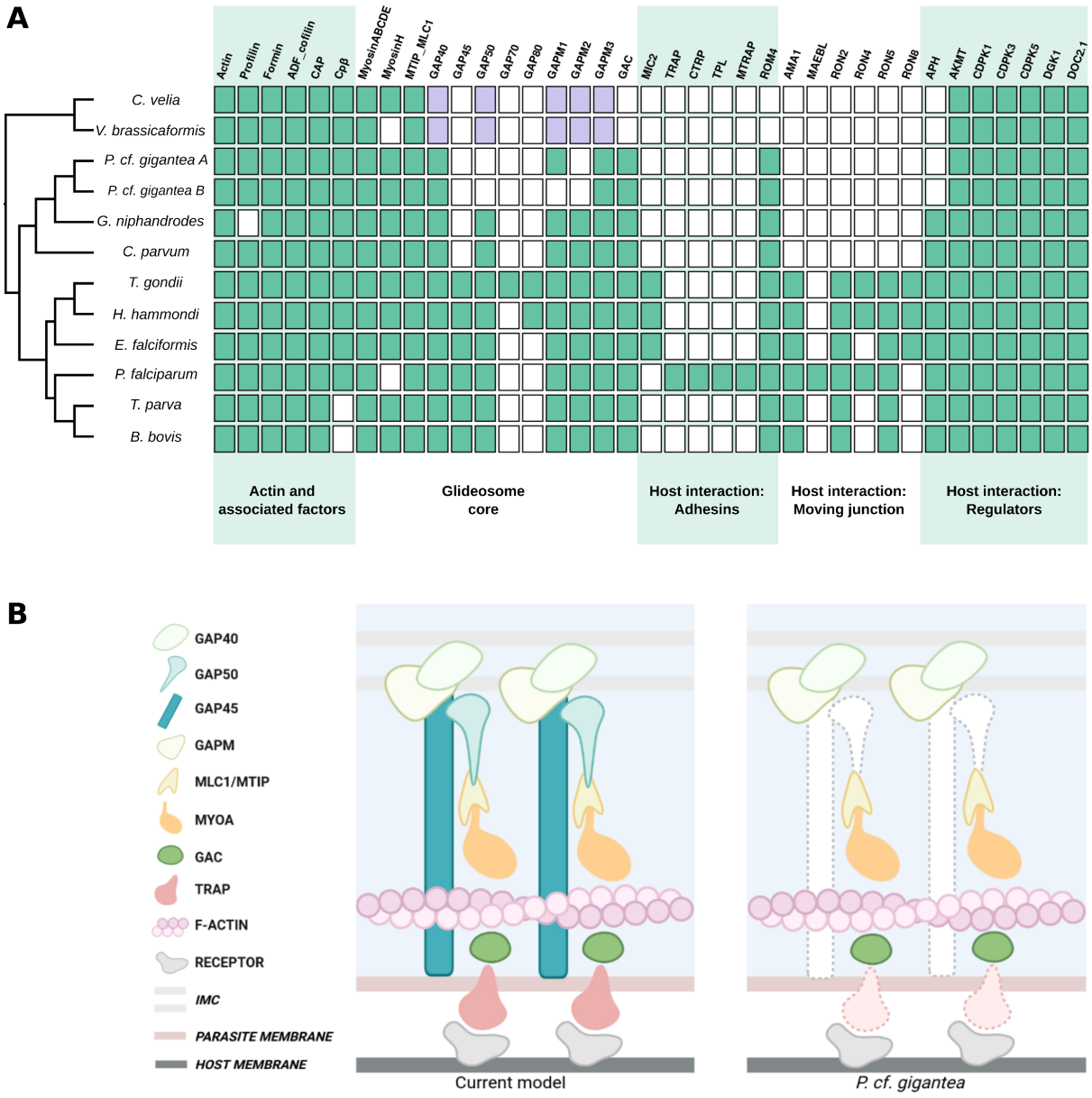


Figure 4.3: **Comparative analysis of glideosome components. A.** Table of presence/absence of glideosome proteins, distributed into functional groups. Glideosome components have been described mainly in *T. gondii* and *P. falciparum*. Proteins were searched for in both *Porospora* genomes as well as in a selection of representative species. Green indicates the presence, while white indicates the absence of a protein. Light green refers to the cases where one-to-one orthologous relationships have not been conclusively identified in *C. velia* and *V. brassicaformis*, but related protein expansions have been observed (Woo et al., 2015). All *P. cf. gigantea* ortholog proteins are detailled in Table 4.2, page 123. **B.** Schematic comparison of the canonical model of the glideosome and the elements found in *P. cf. gigantea* A and B. Missing proteins are shown in dotted line. In Boisard et al. (submitted).

Protein name	P. cf. gigantea A	P. cf. gigantea B
Actin	gene148 gene3475 gene9783 gene7381	gene1189 gene176 gene7820
Profilin	gene4246	gene2168
Formin	gene1045 gene6583	gene2548 gene2999
ADF_cofilin	gene2779	gene8323
CAP	gene8064	gene1376
Cp F-actin capping protein -subunit	gene5850	gene2019
MyosinABCDE ClassXIV	gene7213 gene5479 gene4089 gene9485 gene5566	gene1418 gene3446 gene8205 gene4288 gene1172
MyosinH ClassXIV	gene5024	gene6924
MTIP_MLC1	gene3722	gene9046
GAP40	gene2452	gene4035
GAPM1	GAPM3 gene8241 GAPMx gene7900 GAPMx gene4892	GAPM3 gene1427 GAPMx gene10037
GAPM2		
GAPM3		
GAC	gene9350	gene3183
ROM4	gene7177 gene6011	gene2979 gene5358
AKMT	gene241	gene2337
CDPK1(Tg)/CDPK4(Pf)	gene9265	gene7530
CDPK3(Tg)/CDPK1(Pf)	gene8870	gene2741
CDPK5(Pf)/CDPK5(Tg)	gene4113	gene8961
DGK1	gene3462	gene10271
DOC2.1	gene3776	gene6607
TSP-1 (a)	gene7210	gene1404
TSP-2 (a)	gene4371	gene4603
TSP2 (a)	gene9608	gene6110-6121
TSP_EGF-1 (a)	gene2135	gene5987
TSP_EGF-2 (a)	gene951	gene3952

Table 4.2: ***P. cf. gigantea A* and B glideosome and TRAP-like proteins identifiers.** (a) TRAP like candidates.

(Bergman, 2003)) as well as glideosome associated proteins (GAP): GAP40, GAP45, GAP50 (Gaskins et al., 2004; Baum et al., 2006; Frénal et al., 2010) - GAP70 and GAP80 being only described in *T. gondii* (Frénal et al., 2014). GAP45 binds the glideosome to the plasma membrane by recruiting Myosin A that acts as a bridge (Frénal et al., 2010); GAP40 and GAP50 are thought to help anchoring myosin A to the parasite cytoskeleton (Tardieux and Baum, 2016), whereas another set of glideosome-associated proteins with multiple-membrane spans (GAPM) are believed to interact with the alveolin and subpellicular microtubules network, suggesting an indirect interaction with the IMC (Bullen et al., 2009; Frénal et al., 2017). Finally, the conoid-associated myosin H has been proven necessary for initiating gliding motility in *T. gondii* (Graindorge et al., 2016).

Myosins ABCDE and its associated light chain were found in all species. Myosin H is also widely distributed, although it is missing in *P. falciparum* and *V. brassicaformis*.

The situation is more complex for glideosome associated proteins. GAP40 is the only one found in all species, including probable homologues in Chromerids. Surprisingly, given the central role attributed to GAP45 in the glideosome model, we found no orthologues in either gregarines, *Cryptosporidium* or Chromerids. We found GAP50 in all species except in the two *P. cf. gigantea* genomes, whereas an orthologue exists in the only other known gregarine genome, *G. niphandrodes*. As expected, GAP70 and GAP80, only identified in *T. gondii*, were not found in other species, with the exception of an orthologue for GAP80 in the coccidia *H. hammondi*.

Concerning the GAPMs, we have found orthologues for all species. More precisely regarding the two *P. cf. gigantea* genomes, 2 orthologous proteins for GAPM3 have been identified but none for GAPM2, while one orthologue for GAPM1 has been uniquely identified in genome A. Finally, GAC was found in all species with the exception of the Chromerids, confirming its occurrence in Apicomplexa only.

Adhesins and TRAP-like candidates

The glideosome machinery anchored in the cytoskeleton of the parasite needs to interact with the extracellular receptors of the host cell, in order to propel the parasite forward

on its surface; this is made possible by the presence of extracellular adhesins secreted by the micronemes (Paing and Tolia, 2014; Boucher and Bosch, 2015) and connected to the glideosome through the glideosome associated connector (GAC) protein (Jacot et al., 2016).

One adhesin described in particular in *Plasmodium* is required for gliding: PfTRAP, for Thrombospondin Adhesive Protein (Sultan et al., 1997) whose homologue in *T. gondii* is TgMIC2 (Huynh and Carruthers, 2006); PfTRAP is stored in the micronemes and released on the cell surface at the anterior end in contact with a host cell and translocated towards the posterior end of the sporozoite (Morahan et al., 2009). At the end of the gliding process, rhomboid protease 4 (ROM4) attaches to the adhesins, disengaging them from receptors and, for intracellular parasites, allowing them to enter the host cell (Buguliskis et al., 2010; Shen et al., 2014; Rugarabamu et al., 2015).

TRAP-like proteins constitute a family of functionally homologous proteins involved in parasite gliding motility and cell penetration (Kappe et al., 1999; Morahan et al., 2009; Templeton and Pain, 2016). TRAP-like or TRAP-related proteins have been described in various stages of *Plasmodium* (CTRP, MTRAP, (Dessens et al., 1999; Bargieri, 2016); TLP, (Lacroix and Ménard, 2008)); TRAP-like proteins have also been described *in silico* in *Cryptosporidium* (TRAPCs, CpTSPs, (Deng et al., 2002; Putignani et al., 2008; Templeton and Pain, 2016)) as well as in several *Babesia* and *Theileria* species (Gaffar et al., 2004; Zhou et al., 2006; Yu et al., 2018; Montenegro et al., 2020), in *Neospora caninum* (Lovett, 2000) and in *Eimeria* (Clarke et al., 1990; Witcombe et al., 2003).

For our analysis, we first looked for the TRAP proteins whose implication in gliding have been described by experimental studies (that are TgMIC2, PfTRAP, PfTPL, PfCTRP and PfMTRAP), as well as the ROM4 protein involved in adhesin cleavage. Unsurprisingly, the currently described TRAP proteins seem to be genus- or even species-specific; indeed, MIC2 is only found in Coccidia, while no orthologue to the proteins described in *Plasmodium* could be identified. On the other hand, we found orthologues for ROM4 in all species, except for Chromerids.

The TRAP proteins described in the literature to date have the following characteris-

tics: an extracellular region containing one or more TSP1 domains and/or one or more von Willebrand factor A (vWA) domains (Kappe et al., 1999; Morahan et al., 2009; Templeton and Pain, 2016). They are also characterized by the presence of a single transmembrane domain, a signal peptide, as well as, in some cases, a juxtaposed putative rhomboid protease cleavage site (AGGxxGG), and a short and charged C-terminal cytoplasmic domain, together with aromatic residues. The presence of a YXXϕ tyrosine sorting pattern has also been described (X meaning any amino acid, and ϕ a hydrophobic amino acid - isoleucine, leucine, methionine, phenylalanine, or valine) (Morahan et al., 2009). In order to evaluate the presence of TRAP-like proteins in *P. cf. gigantea* genomes, we inventoried all proteins containing at least one TSP1 domain. We have not identified any orthologs to PfTRAP/TgMIC2 proteins in *P. cf. gigantea* genomes, nor have we identified proteins with all known structural characteristics. Nevertheless, we have identified potential candidates for the TRAP-like family, that are illustrated in Figure 4.4, page 127. We were able to identify a CpTSP2 orthologue within the two *P. cf. gigantea genomes*; like CpTSP2, it is a large protein (∼2800aa) composed of Notch, TSP1, and Sushi domains. It has an addressing signal, a transmembrane domain and a short and charged basic cytoplasmic tail. This protein also has orthologues in *G. niphandrodes*, as well as in Chromerids and Coccidia. We also demonstrated the presence of four other protein pairs present in both A and B genomes, most of which appear to be specific to *P. cf. gigantea*. The first, PgTSP1, has a TSP1 domain, a peptide signal, a transmembrane domain and a short acidic and charged cytoplasmic tail; the YXXϕ motif is also conserved. Despite the absence of an aromatic residue and a vWA domain, this protein appears to be a prime candidate for a TRAP-like protein. The second candidate, PgTSP1a, very similar in structure to the first, also has a TSP1 domain, a peptide signal, a transmembrane domain and a short, charged but basic cytoplasmic tail. The YXXϕ motif and the aromatic residue are not preserved. PgTSP1a seems to have orthologues in Coccidia, that are however very divergent, some of them being predicted as alpha-tubulin suppressor proteins. We also identified in A and B a protein with two TSP1 domains, a peptide signal, a transmembrane domain and a short acidic charged cytoplasmic tail, PgTSP1-EGF. In addition, these candidates possess

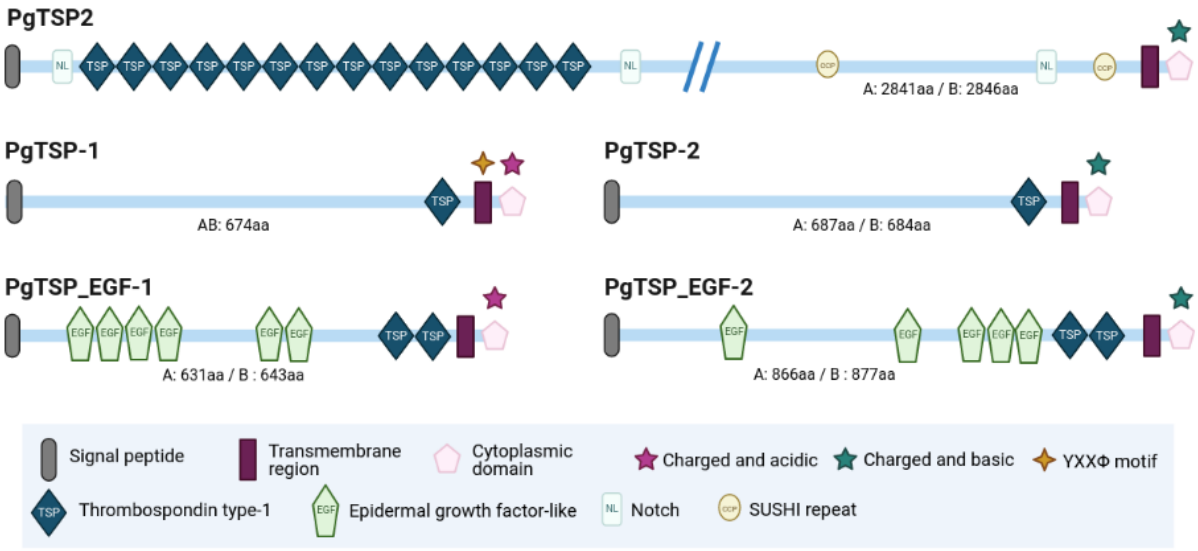


Figure 4.4: **Structures and molecular domains of candidate TRAP-like proteins in *P. cf. gigantea* A and B.** In Boisard et al. (submitted).

several EGF or EGF-like domains in their extracellular portion, as also described in *C. parvum* (CpTSP7, CpTSP8 and CpTSP9 (Deng et al., 2002)), making these two proteins very promising candidates, although the YXXϕ motif and the aromatic residue are not conserved. We also identified another protein very similar in structure, PgTSP1-EGF2, but lacking a peptide signal, and having a basic cytoplasmic tail.

Finally, we inventoried about 30 proteins with a vWA domain, although none of them has a TSP1 domain. Although not corresponding to the canonical structure of TRAP-like proteins, these proteins with adhesion domains would be worth studying.

Moving junction associated proteins

In intracellular apicomplexan such as *T. gondii*, invasion occurs as the tachyzoite initiates a pivotal movement known as reorientation, while the mobile junction settles into the host cell membrane, allowing the parasite to enter the host cell; gliding forces are also involved in this process (Bichet et al., 2014), as well as an active role of the host cell (Portes et al., 2020; Bichet et al., 2014).

A micronemal protein, AMA1, combines with rhoptries neck proteins (RON2, RON4, RON5 and RON8) to firmly maintain the parasite's attachment to the host cell. In

P. falciparum, another AMA-like protein, merozoite apical erythrocyte-binding ligand (MAEBL) is known to have an important role in invasion alongside with AMA1 (Yang et al., 2017).

Unsurprisingly, we have not identified any orthologue to the moving junction proteins nor in gregarines neither in *Cryptosporodium*; indeed, these groups are known to remain mainly extra- (gregarines) or epi-cellular (*Cryptosporium*). We also searched for proteins described in *Cryptosporidium* as implicated in adherence and invasion, such as GP15/40, GP900 and mucins (O'Hara and Chen, 2011; Singh et al., 2015), but found no equivalent in gregarines.

Regulatory factors and signaling pathways

The last important point concerning the study of the glideosome is the signaling pathways involved in the regulation of its activity. The increase of intracellular calcium in the parasite, by activating calcium-dependent protein kinases (CDPK), is involved in the regulation of motility, microneme secretion, invasion and egress (Lourido and Moreno, 2015; Ghartey-Kwansah et al., 2020).

Other proteins known in such signaling pathways include diacylglycerol kinase 1 (DGK1), acylated pleckstrin homology domain-containing protein (APH) which are involved in microneme secretion regulation (Bullen et al., 2016; Darvill et al., 2018), the C2 domains-containing protein DOC2.1 which mediates apical microneme exocytosis (Farrell et al., 2012) ; finally, the apical lysine methyltransferase (AKMT), which is involved in gliding motility, invasion and egress in *T. gondii* (Heaslip et al., 2011).

With the exception of the APH that we were unable to identify in *P. cf. gigantea* or Chromerids, all the regulatory factors appeared to be largely conserved.

4.4 *P. cf. gigantea* genomes reveal unexpected genomic diversity in gregarines

4.4.1 Two genomes highly divergent from all apicomplexan

Compact genomes with similar coding capacities

These two marine gregarine genomes, the first deciphered genomes for any gregarine, highlight several important findings: both exhibit the features of a reduced genome, with a small genome size compared to other apicomplexan, and an especially high gene density (e.g., at similar genome size, *Cryptosporidium* sp. display a number of protein-coding genes of about 3900 only). This result could be partially explained by the absence of certain non-coding sequences into the assembly such as centromeres, telomeres and repeated sequences which are particularly difficult to sequence and assemble, notably in *de novo* assembled genomes. However, we have also evidenced the existence of particularly short introns within these two genomes, which also explain their compaction (analyses done by L. Ponger, in Boisard et al., submitted)). Small introns with similar consensus sequences have been described in *Babesia microti* (Cornillot et al., 2012) - the selective pressure at the origin of this specific class of intron being unknown as of today. These highly synthenic genomes also show a nucleotide divergence of more than 10%, despite a high similarity in terms of size and gene content.

Organellar genomes

So far, we have not identified any organellar genome (mitochondrion and apicoplast); but this absence needs to be investigated more precisely, especially concerning the mitochondrial genome; indeed, the cystic stages from which genomic DNA has been collected are unlikely to have many copies of it. To address this issue, it would be particularly suitable to investigate the trophozoite stages, via single-cell genomics for example. Furthermore, mitochondrial genomes seem to have disappeared from eugregarines according to the recent study by Salomaki et al. (2021). Instead of a mitochondrial genome, the

differentially conserved mitochondrial proteins among the gregarine lineages are encoded in the nuclear genome. It would be interesting to identify the proportion of these proteins conserved within the genomes of *P. cf. gigantea*. As for the apicoplast genome, a recent study stated that it is probably lost in all eugregarines, while archigregarines may have conserved a highly reduced plastid genome (Janouškovec et al., 2019).

An hidden diversity

Despite *de novo* assembled genomes with no close reference available, we were able to evaluate the completeness of these two assembled genomes against available apicomplexan genomes, although very divergent. The two genomes of *P. cf. gigantea* show a completeness score of about 70%; this score, which is not unusual for non-model species (Seppey et al., 2019), is lower than that of the other apicomplexan species we evaluated. This result is however not totally unexpected in this context of lack of sufficiently close reference genomes for effective comparison; since the currently available "apicomplexa" markers cannot possibly reflect the possibility that some of them are lost during evolution in apicomplexan lineages whose genomes are unknown. The gregarine genomes may have lost some genes considered specific to other apicomplexans, while others, absent in other lineages, may have been retained.

Indeed, by analyzing orthologs within apicomplexan genomes, we have observed that a large majority of proteins are differentially conserved between every genera, or even species. A similar analysis has recently been published in Derilus et al. (2021) about several *Plasmodium* species. Here, we evidenced that even within the gregarines, the vast majority of proteins shared by *P. cf. gigantea* A and *P. cf. gigantea* B are absent from the genome of *G. niphandrodes*.

While it was known that currently documented apicomplexan genomes already show great divergence, these two new marine gregarine genomes reveal even greater divergence within the gregarines, and should encourage the scientific community to take hold of these unknown lineages. Their study will allow a better understanding of the evolutionary history of apicomplexan species by highlighting an astonishing protein diversity and a

complex differential inheritance from the common ancestor. Through comparative analyses, we will be able to understand how this inheritance has allowed such a wide range of documented adaptations to parasitism in apicomplexans, which have been able to establish themselves in most animal lineages, vertebrate or not, marine or terrestrial, in one or more hosts, intra- or extracellularly.

4.4.2 A partially conserved gliding machinery

An incomplete machinery in spite of a impressive gliding capacity

Gliding motility in gregarines and other apicomplexans share features in common (IMC, actomyosin motor and the polymerisation of actin during gliding), involving more than 35 proteins that appear well conserved between apicomplexan species and which make together the structure currently known as the glideosome. This structure requires coordinated interactions between surface adhesins and proteins of the parasite cytoskeleton; the actomyosin motor inserts itself between the plasma membrane and the IMC, while the myosin A head moves along the actin filament connected to a cell adhesion molecule. This structure consisting of adhesins and actin filaments can then move towards the posterior end of the cell; this is how the actomyosin motor generates a glide on the cell surface, moving the transmembrane adhesins across the plasma membrane.

This process is well described in *T. gondii* in particular and is actually considered as the mechanism of gliding in Apicomplexa (Boucher and Bosch, 2015; Frénal et al., 2017; Mueller et al., 2017). However, several models of the detailed mecanisms co-exists, as the machinery itself has never been directly observed. This model remains incomplete in its ability to fully explain the mechanics of gliding and invasion, as well as the exact involvement of the proteins described (Tardieux and Baum, 2016). Furthermore, it is unclear if the mecanisms are the same in deep branching apicomplexans such as gregarines, that exhibit alternative modes of motility (Valigurová et al., 2013; Desportes and Schrével, 2013).

Our molecular analysis of the glideosome components shows that the currently known mechanisms described in *T. gondii* and *P. falciparum* cannot fully account for gliding in all

apicomplexans, as it has previously been anticipated (Boucher and Bosch, 2015; Tardieux and Baum, 2016; Frénal et al., 2017). Indeed, some key molecular components such as canonical adhesins or GAP45 are missing, implying an only partially retrieved glideosome in gregarines as well as in *Cryptosporidium* species. The molecular mechanisms that allow *P. cf. gigantea* to move by gliding in the absence of adhesins or GAP45 remain unknown, even though we have observed this gliding movement in *P. cf. gigantea* trophozoites at an impressive rate - asking the question of how *Porospora* trophozoites manage to perform this movement so well in the absence of a complete dedicated machine.

A classical machinery but partially compensated by alternative proteins ?

The TRAP adhesin in *T. gondii*, named TgMIC2, has been demonstrated to be an important but non-essential protein. The reduction in invasion and motility observed experimentally upon inactivation or deletion of TgMIC2 seems to be the result of a deficit in the initiation of motility rather than its total prevention; indeed, some parasites remain able to initiate motility despite the absence of TgMIC2, and are able to sustain it afterwards (Gras et al., 2017). This suggests that TRAP proteins may not be the only proteins involved in host surface adhesion. As we have seen, in the genomes of *P. cf. gigantea*, as well as in other apicomplexans, there are proteins with a structure close to TRAPs, called TRAP-like, that could replace the canonical TRAP proteins. TRAP homologs through apicomplexans (such as TgMIC2, PfTRAP, PfMTRAP, PfCTRP, PfTPL) share structural and functional domains, which suggests that they constitute a family of functionally homologous proteins, playing central roles in the ability of parasites to recognize, adhere to and invade host cells (Templeton and Pain, 2016; Deng et al., 2002). While homologous, these proteins remain species- and stage-specific, allowing the parasites to use the same molecular mechanisms while adjusting them to different host cells (Mohamed et al., 2018).

This is why understanding the evolution of TRAPs proteins involves experimental validation of predicted adhesion proteins in gregarines and *Cryptosporidium* - especially since the presence of these domains in Alveolata does not correlate with gliding motility

(Templeton and Pain, 2016). Indeed, chromerids, which do not glide, have many predicted extracellular proteins with TSP1 and vWA domains. Similarly, the vWA domains, which are found in the canonical TRAPs, appear to be absent from the *Cryptosporidium* proteomes; however, since gliding is observed in these species, it can be assumed that, if the TRAP-like proteins described in *Cryptosporidium* are indeed involved in gliding, then the vWA domains are not essential for this process; it is also possible that the TSP1 domain in 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* (Mohamed et al., 2018; Deng et al., 2002).

As to GAP45, it is thought to maintain the interaction between the IMC and the parasite plasma membrane, and to act as an essential bridge between the two structures (Harding et al., 2019). Likewise, the absence of GAP45 in gregarines and *Cryptosporidium* may be compensated by other GAP-like proteins or even not be a problem at all ; indeed, it has been proposed that a motor architecture could be organized in a much looser manner, in which actin-myosin motors push in a general backward direction, but without necessarily being guided by GAP proteins (Tardieux and Baum, 2016). Furthermore, while TgMLC1 binding to TgGAP45 is considered a key component of the parasite's force transduction mechanism, it has recently been shown that loss of TgMLC1 binding to TgGAP45 has little effect on their ability to initiate or maintain movement (Rompikuntal et al., 2021), questionning again the real role of GAP45 and suggesting our comprehension of the glideosome's proteins' intrication is still incomplete.

A completely different structure taking advantage of the other forms of motility known in gregarines?

Gregarines have other means of motility, presumably governed by other molecular mechanisms. Yet questions have been raised about the relevance of the glideosome concept as applied to gregarines (Valigurová et al., 2013, 2017). In particular, it is known that archigregarines use several modes of movement such as rolling and bending (Desportes and Schrével, 2013).

Furthermore, some gregarines in microscopy studies display the ultrastructural components of gliding and yet do not seem to move by gliding. For example, despite the presence of gliding components such as the three-layer apicomplexan pellicle, actin, myosin, micronemes and a glycocalyx layer where adhesins might be located, the blastogregarine *Siedleckia nematoides* show no sign of gliding motility (Valigurová et al., 2017).

For their part, coelomic and intestinal eugregarines like crustacean gregarines have longitudinal, drapery-like surface structures called epicytic folds that represent the most noticeable feature that differentiates eugregarine trophozoites and gamonts from other apicomplexans, and are considered to be involved in eugregarines' gliding, by increasing the surface area and facilitating actomyosin-based gliding motility (reviewed in Valigurová et al. (2013)). Indeed, actin and myosins A, B and F have been localized in epicytic folds in *Gregarina polymorpha* (Heintzelman, 2004; Heintzelman and Mateer, 2008). Epicytic folds, together with the mucus, which refers to the material often observed in the trace left by gliding eugregarines (Valigurová et al., 2013; Desportes and Schrével, 2013), are definitely key structures to investigate in order to understand their exact composition and thus be able to propose an alternative model to the glideosome one, suited to the motility of eugregarines.

A particularly interesting study concerning the crustacean gregarine *Cephaloidophora cf. communis* reports on the specific structures of its attachment apparatus (Kováčiková et al., 2017). 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). *Cephaloidophora cf. communis* also has a septum, a structure that separates the protomerite from the deutomerite at the cell apex, consisting of tubulin-rich filamentous structures. Together with microneme-like structures, these features suggest a production of adhesion proteins which would be sent through the membrane by the numerous pores visible on the epimerite (Kováčiková et al., 2017).

We were unable to identify alternative movements to gliding motility in *P. cf. gigantea*, like peristaltic movement described in other coelomic eugregarines (Desportes

and Schrével, 2013; Diakin et al., 2017)), and we believe that additional observations are needed to fully document the range of potential motilities in these species, especially since the closely related species *C. cf. communis* is capable of jumping or jerking movements during discontinuous gliding (Kováčiková et al., 2017). The different described structures, or their absence must be evidenced as well; indeed, in eugregarines, subpellicular microtubules have never been observed, whereas they are supposed to be involved in gliding motility in other apicomplexan (Valigurová et al., 2013; Kováčiková et al., 2017).

In light of these hypotheses, involving alternative proteins compensating for canonical glideosome machinery or suggesting the implication of other motility mechanisms altogether, it is likely that the molecular mechanisms leading to gliding motility in *P. cf. gigantea* reveal a unique molecular structure, consecutive to the specific evolutionary path of gregarines, and which differs from what is currently documented in other apicomplexan lineages.

4.5 Perspectives on gregarine genomics

Our investigations from the first genomes of marine gregarines highlight, on the one hand, a molecular diversity of unsuspected magnitude, and on the other hand, offers us the possibility to question the knowledge acquired on apicomplexan by introducing data from a lineage that has been neglected until now.

Although the great diversity of apicomplexan genome structures was already known, whether in terms of genome size, number of coding genes or proportion of non-coding genes (see Chapter 1, Table 1.1, page 32 and this chapter, Table 4.1), the gregarines were until recently considered as a taxonomic group containing closely related organisms and of little interest. However, their ability to parasitize most if not all invertebrate metazoans should have enabled us to envisage a substantial molecular diversity, which would provide the resources for adaptations to extremely varied contexts.

Thus, these first studies allow us to start measuring the extent of this diversity. We currently know genomes for only three gregarines, two marine and one terrestrial, and we already see that their genetic inheritance is very divergent. We have to recall that

these gregarines are part of the same group, the Eugregarines. But what about the genomic diversity of the other two groups of gregarines (see Chapter 1 Figure 1.2, page 14): the archigregarines, which include species considered to be basal to all gregarines, and probably polyphyletic (Wakeman and Leander, 2012; Rueckert and Leander, 2009), and the neogregarines, which gather species considered to be more recently derived and capable of intracelullar life mode (Desportes and Schrével, 2013)?

Beyond the structural characteristics of these genomes and the diversity of their coding capacity, it is essential to document the molecular architectures of key functions or structures in the biology of Apicomplexa. We began this investigation with the proteins involved in gliding, and highlighted their differential conservation at the apicomplexan scale. What about apical complex proteins, and the arsenal of proteins involved in host cell invasion and egress in intracellular parasites? Are these proteins conserved in gregarines? Are they refunctionalized?

If the search for proteins already identified in well-documented apicomplexan lineages appears to be the priority, there is the much more complex question of all the proteins specific to gregarines, and their involvement in structures or behaviors that are unique to them, such as the alternative movements to gliding that we have presented in this chapter.

The abyss of knowledge to be discovered fully justifies the deciphering of other gregarine genomes, because these studies could not be carried out without reference data. The two genomes of the marine gregarine *P. cf. gigantea*, if they are a first step, could never account for the diversity of gregarines alone. Various reference genomes, at the scale of the many documented groups of gregarines, are needed.

Conclusive remarks

At the end of this thesis, the diversity of gregarine genomes is now better documented with the acquisition of assembled genomes for 4 gregarines: *Porospora cf. gigantea* A and *P. cf. gigantea* B, *Diplauxis hatti* and *Gregarina acridiorum*. By taking advantage of the access to the MNHN locust breeding farm for *G. acridiorum*, or by exploiting the specific biological characteristics of *P. gigantea* and *D. hatti*, this study enabled to overcome the major obstacle that prevented the elaboration of such data for gregarines: their non-cultivability. However assembling a genome is only the begining. Gene prediction, which is necessary to get access to functional understanding of these genomes, remains an obstacle in the absence of RNA data - though single cell RNAseq performed on trophozoites stages will help to address this issue in the future. We hope to be able to quickly provide more complete gene predictions for *D. hatti* and *G. acridiorum* so that these two other genomes can also be published.

This research has also confirmed the extent of future taxonomy revisions concerning the gregarines. In fact, the two examples of species delimitation that we examined led in both cases to the identification of cryptic species by molecular data, whereas morphological data did not allow us to discriminate them. Given that it is expected that the diversity known by the described species of gregarines is largely underestimated, the evidence of cryptic species in the very same described species augurs an even greater taxonomical effort to come.

As we have demonstrated with the *P. cf. giganta* genomes, it is no longer possible to rely solely on the 18S SSU rDNA marker to delimit species from a molecular point of view. At the very least, several markers are needed; and possibly genome/transcriptome-

wide analyses will be necessary to properly address this issue. However, an immense sampling effort is required to explore the full diversity of gregarines. The dozen or so species currently documented at the -omic scale are not sufficient to elucidate the evolutionary history of apicomplexan. There is a need to provide genomes/transcriptomes for at least every described gregarine family. It is also essential to take advantage of recent environmental metagenomic methods, since it is likely to discover new gregarines species, and probably also whole new lineages of apicomplexan parasites.

In the light of these results, the veil that hides the whole diversity of gregarines has hardly been lifted yet we can only realize the vastness of the knowledge that still remains to be discovered to understand apicomplexan evolution. As the first described and hopefully soon published reference genomic data for gregarines, *P. cf. gigantea* genomes will allow to begin to fill the knowledge gap at the molecular level for the gregarines, and open the way to genome-scale comparative analyses. The description of two marine gregarine genomes and their deduced proteomes enables to get a more precise overview of the many adaptive mechanisms developed by apicomplexan parasites during their evolution, especially since they have been able to establish themselves in such a wide variety of hosts.

We outlined the structural characteristics of these genomes, and performed the first analyses of their deduced proteomes by comparing them with a selection of apicomplexan proteomes. These first results highlight the unsuspected diversity of the gregarine gene pool; indeed, there is almost as much divergence between the proteome of our marine gregarines and that of the terrestrial gregarine *G. niphandrodes*, as there is with those of *T. gondii* or *P. falciparum*.

Moreover, our study of the glideosome reveals a differential conservation of its constituting proteins between apicomplexan lineages. In particular, our data for *P. cf. gigantea* A and B but also the analysis of the *Cryptosporidium parvum* proteome suggests that the molecular architecture at the origin of the gliding movement is different in gregarines and Cryptosporidia from the one described in *T. gondii* and *P. falciparum*. As gliding is a key component of host cell interaction and/or invasion, these results need to be extended to the recently published transcriptomic data for other gregarines and complemented by ex-

perimental studies to better understand the mechanisms at work. Certainly the in-depth study of such signature behaviors of apicomplexan will allow, by documenting their differential conservation at the proteic level, to determine what are the molecular structures that allow intracellular parasites to manifest such pathogenecity.

In the light of these first results, so many questions arise about the evolutionary history of gregarines, and of apicomplexan as a whole. Highly pathogenic parasites such as *T. gondii* or *Plasmodium* species have been very well documented by the scientific community at the molecular level, but we must now recognize that these species contain only a small part of the apicomplexan diversity. Furthermore, we need to question the knowledge gathered on these well known parasites, and be careful not to generalize it too hastily to all apicomplexan. Even if it is legitimate from a medical point of view to have concentrated the scientific efforts on highly pathogenic parasites, it is now necessary to avoid keeping in the shadow the knowledge that the gregarines are likely to bring us on Apicomplexa. We can't expect to elucidate the major evolutionary questions implied by adaptation to parasitism, by only considering them through a narrow fraction of all the diversity that is enclosed in Apicomplexa. The knowledge generated on the unknown groups will surely not only disrupt our assumptions on apicomplexan, but will also lead to new research concerns, which could answer crucial questions on adaptation to parasitism.

Aknowledgments

I would first like to deeply thank my supervisors, Isabelle Florent and Loïc Ponger, for their constant support during these years, from master to PhD. Many thanks for always being available and attentive; for the great guidance and advice, and for the freedom and confidence I was given during my research. I hope in the future to honor the extensive knowledge I have received from both of you.

I also wish to express special thanks to Evelyne Duvernois-Berthet for her unfailing help and support, and even more for her friendship.

I want to convey my gratitude to Isabelle Desportes and Joseph Schrével for always enthusiastically sharing their invaluable knowledge about gregarines.

Many thanks to my two host labs, MCAM and STRING, especially the PPL and ARCHE teams; it was an extremely enriching experience to work with you all; thanks to all lab members for the inspiring exchanges, the scientific seminars and the knowledge acquired through working by your side.

I would like to thank all my collaborators, especially Linda Duval and Amandine Labat from the PPL team, as well as Laure Guillou from the Roscoff Marine Station.

Thanks also to the PCIA team for the access to the MNHN computing cluster, and to Joël Pothier and Sophie Brouillet from ABI for sharing their equipment.

Thanks to the National Center for Scientific Research CNRS for funding this PhD, as well as to the funders involved in this project : EMBRC France 2016-2018 and Federative Project MNHN AVIV 2018-2019.

Finally, thank you to all those, family and friends, who encouraged me and believed in me, and who have been there for all these years.
Thanks to my parents and especially to my two sisters Marion and Gaëlle,
Thanks to Maxime, Servin and Caroline,
Thanks to all my Spectrum friends.

Thanks to Loïc, for everything.

Bibliography

Abrahamsen, M. S., Templeton, T. J., Enomoto, S., Abrahante, J. E., Zhu, G., Lancto, C. A., Deng, M., Liu, C., Widmer, G., Tzipori, S., Buck, G. A., Xu, P., Bankier, A. T., Dear, P. H., Konfortov, B. A., Spriggs, H. F., Iyer, L., Anantharaman, V., Aravind, L., and Kapur, V. (2004). Complete Genome Sequence of the Apicomplexan *Cryptosporidium parvum*. 304:6.

Adl, S. M., Bass, D., Lane, C. E., Lukeš, J., Schoch, C. L., Smirnov, A., Agatha, S., Berney, C., Brown, M. W., Burki, F., Cárdenas, P., Čepička, I., Chistyakova, L., Campo, J., Dunthorn, M., Edvardsen, B., Eglit, Y., Guillou, L., Hampl, V., Heiss, A. A., Hoppenrath, M., James, T. Y., Karnkowska, A., Karpov, S., Kim, E., Kolisko, M., Kudryavtsev, A., Lahr, D. J., Lara, E., Le Gall, L., Lynn, D. H., Mann, D. G., Massana, R., Mitchell, E. A., Morrow, C., Park, J. S., Pawlowski, J. W., Powell, M. J., Richter, D. J., Rueckert, S., Shadwick, L., Shimano, S., Spiegel, F. W., Torruella, G., Youssef, N., Zlatogursky, V., and Zhang, Q. (2019). Revisions to the Classification, Nomenclature, and Diversity of Eukaryotes. *Journal of Eukaryotic Microbiology*, 66(1):4–119.

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(3):403–410.

Aly, A. S., Vaughan, A. M., and Kappe, S. H. (2009). Malaria Parasite Development in the Mosquito and Infection of the Mammalian Host. *Annual Review of Microbiology*, 63(1):195–221.

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(7):2251–2252.

Auburn, S., Böhme, U., Steinbiss, S., Trimarsanto, H., Hostetler, J., Sanders, M., Gao, Q., Nosten, F., Newbold, C. I., Berriman, M., Price, R. N., and Otto, T. D. (2016). A new *Plasmodium vivax* reference sequence with improved assembly of the subtelomeres reveals an abundance of pir genes. *Wellcome Open Research*, 1:4.

Aurrecoechea, C., Barreto, A., Basenko, E. Y., Brestelli, J., Brunk, B. P., Cade, S., Crouch, K., Doherty, R., Falke, D., Fischer, S., Gajria, B., Harb, O. S., Heiges, M., Hertz-Fowler, C., Hu, S., Iodice, J., Kissinger, J. C., Lawrence, C., Li, W., Pinney, D. F., Pulman, J. A., Roos, D. S., Shanmugasundram, A., Silva-Franco, F., Steinbiss, S., Stoeckert, C. J., Spruill, D., Wang, H., Warrenfeltz, S., and Zheng, J. (2017). EuPathDB the eukaryotic pathogen genomics database resource. *Nucleic Acids Research*, 45(D1):D581–D591.

Ayali, A. (2019). The puzzle of locust density-dependent phase polyphenism. *Current Opinion in Insect Science*, 35:41–47.

Bahrndorff, S., Alemu, T., Alemneh, T., and Lund Nielsen, J. (2016). The Microbiome of Animals: Implications for Conservation Biology. *International Journal of Genomics*, 2016:1–7.

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., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., and Pevzner, P. A. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*, 19(5):455–477.

Bargieri, D. Y. (2016). *Plasmodium* Merozoite TRAP Family Protein Is Essential for Vacuole Membrane Disruption and Gamete Egress from Erythrocytes. *Cell Host Microbe*, 20(5):618–630.

Barta, J. R. and Thompson, R. A. (2006). What is *Cryptosporidium*? Reappraising its biology and phylogenetic affinities. *Trends in Parasitology*, 22(10):463–468.

Bartošová-Sojková, P., Oppenheim, R. D., Soldati-Favre, D., and Lukeš, J. (2015). Epicellular Apicomplexans: Parasites "On the Way In". *PLOS Pathogens*, 11(9):e1005080.

Baum, J., Richard, D., Healer, J., Rug, M., Krnajski, Z., Gilberger, T.-W., Green, J. L., Holder, A. A., and Cowman, A. F. (2006). A Conserved Molecular Motor Drives Cell Invasion and Gliding Motility across Malaria Life Cycle Stages and Other Apicomplexan Parasites. *Journal of Biological Chemistry*, 281(8):5197–5208.

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(3):188–198.

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(1):39–49.

Bernays, E. A. (1981). A specialized region of the gastric caeca in the locust, *Schistocerca gregaria*. *Physiological Entomology*, 6(1):1–6.

Bertiaux, E., Balestra, A. C., Bournonville, L., Louvel, V., Maco, B., Soldati-Favre, D., Brochet, M., Guichard, P., and Hamel, V. (2021). Expansion microscopy provides new insights into the cytoskeleton of malaria parasites including the conservation of a conoid. *PLOS Biology*, 19(3):e3001020.

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 Biology*, 12(1):773.

Birney, E. (2004). GeneWise and Genomewise. *Genome Research*, 14(5):988–995.

Boenigk, J., Ereshefsky, M., Hoef-Emden, K., Mallet, J., and Bass, D. (2012). Concepts in protistology: Species definitions and boundaries. *European Journal of Protistology*, 48(2):96–102.

Boisard, J. and Florent, I. (2020). Why the –omic future of Apicomplexa should include gregarines. *Biology of the Cell*, 112(6):173–185.

Boucher, L. E. and Bosch, J. (2015). The apicomplexan glideosome and adhesins – Structures and function. *Journal of Structural Biology*, 190(2):93–114.

Bouzid, M., Hunter, P. R., Chalmers, R. M., and Tyler, K. M. (2013). *Cryptosporidium* Pathogenicity and Virulence. *Clinical Microbiology Reviews*, 26(1):115–134.

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., Robles, R., Chu, K.-H., Tsang, L.-M., Kim, D., Martin, J. W., and Crandall, K. A. (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(4):457–479.

Brayton, K. A., Lau, A. O. T., Herndon, D. R., Hannick, L., Kappmeyer, L. S., Berens, S. J., Bidwell, S. L., Brown, W. C., Crabtree, J., Fadrosh, D., Feldblum, T., Forberger, H. A., Haas, B. J., Howell, J. M., Khouri, H., Koo, H., Mann, D. J., Norimine, J., Paulsen, I. T., Radune, D., Ren, Q., Smith, R. K., Suarez, C. E., White, O., Wortman, J. R., Knowles, D. P., McElwain, T. F., and Nene, V. M. (2007). Genome Sequence of *Babesia bovis* and Comparative Analysis of Apicomplexan Hemoprotozoa. *PLoS Pathogens*, 3(10):e148.

Buchmann, K. (2014). Evolution of Innate Immunity: Clues from Invertebrates via Fish to Mammals. *Frontiers in Immunology*, 5.

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 Pathogens*, 6(4).

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(3):349–360.

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. *Journal of Biological Chemistry*, 284(37):25353–25363.

Burki, F., Roger, A. J., Brown, M. W., and Simpson, A. G. (2019). The New Tree of Eukaryotes. *Trends in Ecology & Evolution*.

Canning, E. U. (1956). A New Eugregarine of Locusts, *Gregarina garnhami* n.sp., parasitic in *Schistocerca gregaria* Forsk. *The Journal of Protozoology*, 3(2):50–62.

Castresana, J. (2000). Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Molecular Biology and Evolution*, 17(4):540–552.

Cavalier-Smith, T. (2014). Gregarine site-heterogeneous 18S rDNA trees, revision of gregarine higher classification, and the evolutionary diversification of Sporozoa. *European Journal of Protistology*, 50(5):472–495.

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. *Molecular and Biochemical Parasitology*, 41(2):269–279.

Clopton, R. E. (2009). Phylogenetic Relationships, Evolution, and Systematic Revision of the Septate Gregarines (Apicomplexa: Eugregarinorida: Septatorina). *Comparative Parasitology*, 76(2):167–190.

Clopton, R. E., Janovy, J. J., and Percival, T. (1992). Host stadium specificity in the gregarine assemblage parasitizing *Tenebrio molitor*. *J. Parasitol.*, (78):334–337.

Conway, J. R., Lex, A., and Gehlenborg, N. (2017). UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics*, 33(18):2938–2940.

Corbel, J. (1967). Intensité, fréquence et facteurs des infections grégariniennes chez les Insectes Orthoptères. *Annales de Parasitologie Humaine et Comparée*, (42):373–385.

Corbel, J. C. (1968a). New or poorly known gregarines parasitic on orthopterus insects. *Annales de Parasitologie Humaine et Comparée*, 43(3):291–320.

Corbel, J. C. (1968b). Parasitic specificity of gregarines of Orthoptera. *Annales De Parasitologie Humaine Et Comparee*, 43(1):25–32.

Cornillot, E., Hadj-Kaddour, K., Dassouli, A., Noel, B., Ranwez, V., Vacherie, B., Augagneur, Y., Brès, V., Duclos, A., Randazzo, S., Carcy, B., Debierre-Grockiego, F., Delbecq, S., Moubri-Ménage, K., Shams-Eldin, H., Usmani-Brown, S., Bringaud, F., Wincker, P., Vivarès, C. P., Schwarz, R. T., Schetters, T. P., Krause, P. J., Gorenflot, A., Berry, V., Barbe, V., and Ben Mamoun, C. (2012). Sequencing of the smallest Apicomplexan genome from the human pathogen *Babesia microti*. *Nucleic Acids Research*, 40(18):9102–9114.

Counihan, N. A., Kalanon, M., Coppel, R. L., and de Koning-Ward, T. F. (2013). *Plasmodium* rhoptry proteins: why order is important. *Trends in Parasitology*, 29(5):228–236.

Crandall, K. A., Pérez-Losada, M., and Porter, M. L. (2009). Crabs, shrimps, and lobsters (Decapoda). In *The Timetree Of Life*, page 551. Oxford University Press, New York.

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 Pathogens*, 6(10):e1001132.

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(8):1059–1071.

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(S18):464.

De Bauchamp, P. (1910). Sur une grégarine nouvelle du genre *Porospora*. *Comptes rendus de l'Académie des Sciences de Paris*, 151(151):997–999.

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahe, F., Logares, R., Lara, E., Berney, C., Le Bescot, N., Probert, I., Carmichael, M., Poulain, J., Romac, S., Colin, S., Aury, J.-M., Bittner, L., Chaffron, S., Dunthorn, M., Engelen, S., Flegontova, O., Guidi, L., Horak, A., Jaillon, O., Lima-Mendez, G., Luke, J., Malviya, S., Morard, R., Mulot, M., Scalco, E., Siano, R., Vincent, F., Zingone, A., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Tara Oceans Coordinators, Acinas, S. G., Bork, P., Bowler, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Not, F., Ogata, H., Pesant, S.,

Raes, J., Sieracki, M. E., Speich, S., Stemmann, L., Sunagawa, S., Weissenbach, J., Wincker, P., Karsenti, E., Boss, E., Follows, M., Karp-Boss, L., Krzic, U., Reynaud, E. G., Sardet, C., Sullivan, M. B., and Velayoudon, D. (2015). Eukaryotic plankton diversity in the sunlit ocean. *Science*, 348(6237):1261605–1261605.

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. *Environmental Microbiology*, 21(10):3855–3861.

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. *Infection and Immunity*, 70(12):6987–6995.

Derilus, D., Rahman, M., Serrano, A., and Massey, S. (2021). Proteome size reduction in Apicomplexans is linked with loss of DNA repair and host redundant pathways. *Infection, Genetics and Evolution*, 87:104642.

Desportes, I. and Schrével, J., editors (2013). *Treatise on zoology–anatomy, taxonomy, biology: The Gregarines. The early branching Apicomplexa.* Brill, Leiden.

Desportes, I., Vivarès, C., and Théodoridès, J. (1977). Intérêt taxonomique de l'ultrastructure épicytaire chez *Ganymedes* Huxley, *Porospora* Schneider et *Thiriotia* n. g., eugrégarines parasites de crustacés. *Ann Sci Nat Zool*, 19:261–277.

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. *The EMBO Journal*, 18(22):6221–6227.

Diakin, A., Paskerova, G. G., Simdyanov, T. G., Aleoshin, V. V., and Valigurová, A. (2016). Morphology and Molecular Phylogeny of Coelomic Gregarines (Apicomplexa) with Different Types of Motility: *Urospora ovalis* and *U. travisiae* from the Polychaete *Travisia forbesii*. *Protist*, 167(3):279–301.

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(1):56–66.

Drewry, L. L. and Sibley, L. D. (2015). *Toxoplasma* Actin Is Required for Efficient Host Cell Invasion. *mBio*, 6(3):e00557–15.

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(3):e91819.

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(6065):218–221.

Florent, I., Chapuis, M. P., Labat, A., Boisard, J., Leménager, N., Michel, B., and Desportes-Livage, I. (2021). Integrative taxonomy confirms that *Gregarina garnhami* and *G. acridiorum* (Apicomplexa, Gregarinidae), parasites of *Schistocerca gregaria* and *Locusta migratoria* (Insecta, Orthoptera), are distinct species. *Parasite*, 28:12.

Frénal, K., Dubremetz, J.-F., Lebrun, M., and Soldati-Favre, D. (2017). Gliding motility powers invasion and egress in Apicomplexa. *Nature Reviews Microbiology*, 15(11):645–660.

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(11):e1004504.

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(4):343–357.

Füssy, Z. and Oborník, M. (2017). Reductive Evolution of Apicomplexan Parasites from Phototrophic Ancestors. In Pontarotti, P., editor, *Evolutionary Biology: Self/Nonself Evolution, Species and Complex Traits Evolution, Methods and Concepts*, pages 217–236. Springer International Publishing, Cham.

Gaffar, F. R., Yatsuda, A. P., Franssen, F. F., and Vries, E. d. (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(1):25–34.

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(6):1356–1367.

Gardner, M. J. (2005). Genome Sequence of *Theileria parva*, a Bovine Pathogen That Transforms Lymphocytes. *Science*, 309(5731):134–137.

Gardner, M. J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R. W., Carlton, J. M., Pain, A., Nelson, K. E., Bowman, S., Paulsen, I. T., James, K., Eisen, J. A.,

Rutherford, K., Salzberg, S. L., Craig, A., Kyes, S., Chan, M.-S., Nene, V., Shallom, S. J., Suh, B., Peterson, J., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., Haft, D., Mather, M. W., Vaidya, A. B., Martin, D. M. A., Fairlamb, A. H., Fraunholz, M. J., Roos, D. S., Ralph, S. A., McFadden, G. I., Cummings, L. M., Subramanian, G. M., Mungall, C., Venter, J. C., Carucci, D. J., Hoffman, S. L., Newbold, C., Davis, R. W., Fraser, C. M., and Barrell, B. (2002). Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*, 419(6906):498–511.

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(3):383–393.

Gawad, C., Koh, W., and Quake, S. R. (2016). Single-cell genome sequencing: current state of the science. *Nature Reviews Genetics*, 17(3):175–188.

Ghartey-Kwansah, G., Yin, Q., Li, Z., Gumpper, K., Sun, Y., Yang, R., Wang, D., Jones, O., Zhou, X., Wang, L., Bryant, J., Ma, J., Boampong, J. N., and Xu, X. (2020). Calcium-dependent Protein Kinases in Malaria Parasite Development and Infection. *Cell Transplantation*, 29:096368971988488.

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(1):e1005388.

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 Research*, 2:32.

Grassé, P.-P. (1953). Sous-embranchement des Sporozoaires. In *Traité de Zoologie*, volume 1, fasc. 2, pages 546–633.

Gubbels, M.-J. and Duraisingh, M. T. (2012). Evolution of apicomplexan secretory organelles. *International Journal for Parasitology*, 42(12):1071–1081.

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(1).

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–1075.

Haas, B. J. (2003). Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research*, 31(19):5654–5666.

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., MacManes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C. N., Henschel, R., LeDuc, R. D., Friedman, N., and Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, 8(8):1494–1512.

Hadziavdic, K., Lekang, K., Lanzen, A., Jonassen, I., Thompson, E. M., and Troedsson, C. (2014). Characterization of the 18S rRNA Gene for Designing Universal Eukaryote Specific Primers. *PLoS ONE*, 9(2):e87624.

Hakimi, M.-A., Olias, P., and Sibley, L. D. (2017). *Toxoplasma* Effectors Targeting Host Signaling and Transcription. *Clinical Microbiology Reviews*, 30(3):615–645.

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(1).

Hatt, P. (1931). L'évolution des Porosporides chez les mollusques. *Archives de zoologie expérimentale et générale*, 72:341–415.

Hayashida, K., Hara, Y., Abe, T., Yamasaki, C., Toyoda, A., Kosuge, T., Suzuki, Y., Sato, Y., Kawashima, S., Katayama, T., Wakaguri, H., Inoue, N., Homma, K., Tada-Umezaki, M., Yagi, Y., Fujii, Y., Habara, T., Kanehisa, M., Watanabe, H., Ito, K., Gojobori, T., Sugawara, H., Imanishi, T., Weir, W., Gardner, M., Pain, A., Shiels, B., Hattori, M., Nene, V., and Sugimoto, C. (2012). Comparative Genome Analysis of Three Eukaryotic Parasites with Differing Abilities To Transform Leukocytes Reveals Key Mediators of *Theileria*-Induced Leukocyte Transformation. *mBio*, 3(5).

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 Pathogens*, 7(9):e1002201.

Heintzelman, M. B. (2004). Actin and myosin in *Gregarina polymorpha*. *Cell Motility and the Cytoskeleton*, 58(2):83–95.

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(1):158–168.

Heitlinger, E., Spork, S., Lucius, R., and Dieterich, C. (2014). The genome of *Eimeria falciformis*, reduction and specialization in a single host apicomplexan parasite. *BMC Genomics*, 15(1):696.

Herm-Gotz, A. (2002). *Toxoplasma gondii* myosin A and its light chain: a fast, single-headed, plus-end-directed motor. *The EMBO Journal*, 21(9):2149–2158.

Hussain, K. J., Krishnan, S. M., Johny, S., and Whitman, D. W. (2013). Phenotypic Plasticity in a Gregarine Parasite (Apicomplexa: Eugregarinordia) Infecting Grasshoppers. *Comparative Parasitology*, 80(2):233–239.

Huynh, M.-H. and Carruthers, V. B. (2006). *Toxoplasma* MIC2 Is a Major Determinant of Invasion and Virulence. *PLoS Pathogens*, 2(8):e84.

Ifeonu, O. O., Chibucos, M. C., Orvis, J., Su, Q., Elwin, K., Guo, F., Zhang, H., Xiao, L., Sun, M., Chalmers, R. M., Fraser, C. M., Zhu, G., Kissinger, J. C., Widmer, G., and Silva, J. C. (2016). Annotated draft genome sequences of three species of *Cryptosporidium* : *Cryptosporidium meleagridis* isolate UKMEL1, *C. baileyi* isolate TAMU-09Q1 and *C. hominis* isolates TU502_2012 and UKH1. *Pathogens and Disease*, 74(7):ftw080.

Iritani, D., Horiguchi, T., and Wakeman, K. C. (2018a). Molecular Phylogenetic Positions and Ultrastructure of Marine Gregarines (Apicomplexa) *Cuspisella ishikariensis* n. gen., n. sp. and *Loxomorpha* cf. *harmothoe* from Western Pacific scaleworms (Polynoidae). *Journal of Eukaryotic Microbiology*, 65(5):637–647.

Iritani, D., Wakeman, K. C., and Leander, B. S. (2018b). Molecular Phylogenetic Positions of Two New Marine Gregarines (Apicomplexa)- *Paralecudina anankea* n. sp. and *Lecudina caspera* n. sp.-from the Intestine of *Lumbrineris inflata* (Polychaeta) Show Patterns of Co-evolution. *Journal of Eukaryotic Microbiology*, 65(2):211–219.

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(1):56–70.

Jalovecka, M., Hajdusek, O., Sojka, D., Kopacek, P., and Malandrin, L. (2018). The Complexity of Piroplasms Life Cycles. *Frontiers in Cellular and Infection Microbiology*, 8:248.

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.

Janouškovec, J., Tikhonenkov, D. V., Burki, F., Howe, A. T., Kolísko, M., Mylnikov, A. P., and Keeling, P. J. (2015). Factors mediating plastid dependency and the origins of parasitism in apicomplexans and their close relatives. *Proceedings of the National Academy of Sciences*, 112(33):10200–10207.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez, R., and Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics*, 30(9):1236–1240.

Kalvari, I., Nawrocki, E. P., Argasinska, J., Quinones-Olvera, N., Finn, R. D., Bateman, A., and Petrov, A. I. (2018). Non-Coding RNA Analysis Using the Rfam Database. *Current Protocols in Bioinformatics*, 62(1):e51.

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(5):937–944.

Kappmeyer, L. S., Thiagarajan, M., Herndon, D. R., Ramsay, J. D., Caler, E., Djikeng, A., Gillespie, J. J., Lau, A. O., Roalson, E. H., Silva, J. C., Silva, M. G., Suarez, C. E., Ueti, M. W., Nene, V. M., Mealey, R. H., Knowles, D. P., and Brayton, K. A. (2012). Comparative genomic analysis and phylogenetic position of *Theileria equi*. *BMC Genomics*, 13(1):603.

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(35):9834–9839.

Katoh, K. and Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, 30(4):772–780.

Keeley, A. and Soldati, D. (2004). The glideosome: a molecular machine powering motility and host-cell invasion by Apicomplexa. *Trends in Cell Biology*, 14(10):528–532.

King, C. (1988). Cell motility of sporozoan protozoa. *Parasitology Today*, 4(11):315–319.

King, C. and Sleep, J. (2005). Modelling the mechanism of gregarine gliding using bead translocation. *The Journal of Eukaryotic Microbiology*, 52(2):7S–27S.

Kořený, L., Zeeshan, M., Barylyuk, K., Tromer, E. C., van Hooff, J. J. E., Brady, D., Ke, H., Chelaghma, S., Ferguson, D. J. P., Eme, L., Tewari, R., and Waller, R. F. (2021).

Molecular characterization of the conoid complex in *Toxoplasma* reveals its conservation in all apicomplexans, including *Plasmodium* species. *PLOS Biology*, 19(3):e3001081.

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics*, 5(1):59.

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.

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*.

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(6):1547–1549.

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(7):1812–1819.

Kuriyama, R., Besse, C., Gèze, M., Omoto, C. K., and Schrével, J. (2005). Dynamic organization of microtubules and microtubule-organizing centers during the sexual phase of a parasitic protozoan, *Lecudina tuzetae* (Gregarine, Apicomplexa). *Cell Motility and the Cytoskeleton*, 62(4):195–209.

Kwong, W. K., del Campo, J., Mathur, V., Vermeij, M. J. A., and Keeling, P. J. (2019). A widespread coral-infecting apicomplexan with chlorophyll biosynthesis genes. *Nature*, 568(7750):103–107.

Labbé, A. (1899). *Sporozoa, Das Tierreich: Eine Zusammenstellung und Kennzeichnung der rezenten Tierformen*. Berlin, Germany, r. friedlander und sohn edition.

Lacroix, C. and Ménard, R. (2008). TRAP-like protein of *Plasmodium* sporozoites: linking gliding motility to host-cell traversal. *Trends in Parasitology*, 24(10):431–434.

Lange, C. E. and Wittenstein, E. (2002). The life cycle of *Gregarina ronderosi* n. sp. (Apicomplexa: Gregarinidae) in the Argentine grasshopper *Dichroplus elongatus* (Orthoptera: Acrididae). *Journal of Invertebrate Pathology*, 79(1):27–36.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359.

Léger, L. (1892). Recherches sur les grégarines. *Tablettes Zoologiques*, 3:1–182.

Léger, L. (1893). Sur une grégarine nouvelle des Acridiens d'Algérie. *Comptes Rendus de l'Académie des Sciences de Paris*, (117):811–813.

Letunic, I., Khedkar, S., and Bork, P. (2021). SMART: recent updates, new developments and status in 2020. *Nucleic Acids Research*, 49(D1):D458–D460.

Levine, N. D. (1988). Progress in Taxonomy of the Apicomplexan Protozoa. *The Journal of Protozoology*, 35(4):518–520.

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(16):2078–2079.

Li, L., Stoeckert, C., and Roos, D. (2003). OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Research*, 13(9):2178–2189.

Limenitakis, J. and Soldati-Favre, D. (2011). Functional genetics in Apicomplexa: Potentials and limits. *FEBS Letters*, 585(11):1579–1588.

Lipa, J.-C. and Santiago-Alvarez, C. (1996). Gregarines (Eugregarinorida: Apicomplexa) in natural populations of *Dociostaurus maroccanus, Calliptamus italicus* and other Orthoptera. *Acta Protozoologica*, (35):49–59.

Liu, S., Wang, L., Zheng, H., Xu, Z., Roellig, D. M., Li, N., Frace, M. A., Tang, K., Arrowood, M. J., Moss, D. M., Zhang, L., Feng, Y., and Xiao, L. (2016). Comparative genomics reveals *Cyclospora cayetanensis* possesses coccidia-like metabolism and invasion components but unique surface antigens. *BMC Genomics*, 17(1):316.

Lourido, S. and Moreno, S. N. (2015). The calcium signaling toolkit of the Apicomplexan parasites *Toxoplasma gondii* and *Plasmodium* spp. *Cell Calcium*, 57(3):186–193.

Lovett, J. (2000). Molecular characterization of a thrombospondin-related anonymous protein homologue in *Neospora caninum*. *Molecular and Biochemical Parasitology*, 107(1):33–43.

Mahé, F., de Vargas, C., Bass, D., Czech, L., Stamatakis, A., Lara, E., Singer, D., Mayor, J., Bunge, J., Sernaker, S., Siemensmeyer, T., Trautmann, I., Romac, S., Berney, C., Kozlov, A., Mitchell, E. A. D., Seppey, C. V. W., Egge, E., Lentendu, G., Wirth, R., Trueba, G., and Dunthorn, M. (2017). Parasites dominate hyperdiverse soil protist communities in Neotropical rainforests. *Nature Ecology & Evolution*, 1(4):0091.

Mathur, V., Kolísko, M., Hehenberger, E., Irwin, N. A., Leander, B. S., Kristmundsson, Á., Freeman, M. A., and Keeling, P. J. (2019). Multiple Independent Origins of Apicomplexan-Like Parasites. *Current Biology*, 29(17):2936–2941.e5.

Mathur, V., Kwong, W. K., Husnik, F., Irwin, N. A. T., Kristmundsson, Á., Gestal, C., Freeman, M., and Keeling, P. J. (2021a). Phylogenomics Identifies a New Major Subgroup of Apicomplexans, Marosporida *class nov.*, with Extreme Apicoplast Genome Reduction. *Genome Biology and Evolution*, 13(2):evaa244.

Mathur, V., Wakeman, K. C., and Keeling, P. J. (2021b). Parallel functional reduction in the mitochondria of apicomplexan parasites. *Current Biology*, page S0960982221005418.

Mayhew, P. J. (2007). Why are there so many insect species? Perspectives from fossils and phylogenies. *Biological Reviews*, 82(3):425–454.

Medina-Durán, J. H., Mayén-Estrada, R., Mariño-Pérez, R., and Song, H. (2019). Morphology and Phylogenetic Position of Two New Gregarine Species (Apicomplexa: Eugregarinorida) Parasitizing the Lubber Grasshopper *Taeniopoda centurio* (Drury, 1770) (Insecta: Orthoptera: Romaleidae) in Mexico. *Journal of Eukaryotic Microbiology*.

Mehta, S. and Sibley, L. D. (2011). Actin depolymerizing factor controls actin turnover and gliding motility in *Toxoplasma gondii*. *Molecular Biology of the Cell*, 22(8):1290–1299.

Meissner, M., Schluter, D., and Soldati, D. (2002). Role of *Toxoplasma gondii* Myosin A in Powering Parasite Gliding and Host Cell Invasion. *Science*, 298(5594):837–840.

Milotic, M., Lymbery, A., Thompson, A., Doherty, J.-F., and Godfrey, S. (2020). Parasites are endangered by the conservation of their hosts: Meta-analyses of the effect of host captivity on the odds of parasite infection. *Biological Conservation*, 248:108702.

Mohamed, A. R., Cumbo, V. R., Harii, S., Shinzato, C., Chan, C. X., Ragan, M. A., Satoh, N., Ball, E. E., and Miller, D. J. (2018). Deciphering the nature of the coral–Chromera association. *The ISME Journal*, 12(3):776–790.

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(1):602.

Morahan, B. J., Wang, L., and Coppel, R. L. (2009). No TRAP, no invasion. *Trends in Parasitology*, 25(2):77–84.

Morrissette, N. S. and Sibley, L. D. (2002). Cytoskeleton of Apicomplexan Parasites. *Microbiology and Molecular Biology Reviews*, 66(1):21–38.

Mueller, C., Graindorge, A., and Soldati-Favre, D. (2017). Functions of myosin motors tailored for parasitism. *Current Opinion in Microbiology*, 40:113–122.

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. *Acta Carsologica*, 48(1).

Nawrocki, E. P. and Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, 29(22):2933–2935.

Ninham, J. A. (1995). *Phylogenetic analysis of five protoctist parasites of insects*. PhD Thesis.

Nocciolini, C., Cucini, C., Leo, C., Francardi, V., Dreassi, E., and Carapelli, A. (2018). Assessing the Efficiency of Molecular Markers for the Species Identification of Gregarines Isolated from the Mealworm and Super Worm Midgut. *Microorganisms*, 6(4):119.

O'Hara, S. P. and Chen, X.-M. (2011). The cell biology of *Cryptosporidium* infection. *Microbes and Infection*, 13(8-9):721–730.

Omoto, C. K., Toso, M., Tang, K., and Sibley, L. D. (2004). Expressed sequence tag (EST) analysis of Gregarine gametocyst development. *International Journal for Parasitology*, 34(11):1265–1271.

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(3):597–604.

Otto, T. D., Böhme, U., Jackson, A. P., Hunt, M., Franke-Fayard, B., Hoeijmakers, W. A. M., Religa, A. A., Robertson, L., Sanders, M., Ogun, S. A., Cunningham, D., Erhart, A., Billker, O., Khan, S. M., Stunnenberg, H. G., Langhorne, J., Holder, A. A., Waters, A. P., Newbold, C. I., Pain, A., Berriman, M., and Janse, C. J. (2014). A comprehensive evaluation of rodent malaria parasite genomes and gene expression. *BMC Biology*, 12(1).

Paing, M. M. and Tolia, N. H. (2014). Multimeric Assembly of Host-Pathogen Adhesion Complexes Involved in Apicomplexan Invasion. *PLoS Pathogens*, 10(6):e1004120.

Palmieri, N., Shrestha, A., Ruttkowski, B., Beck, T., Vogl, C., Tomley, F., Blake, D. P., and Joachim, A. (2017). The genome of the protozoan parasite *Cystoisospora suis* and a reverse vaccinology approach to identify vaccine candidates. *International Journal for Parasitology*, 47(4):189–202.

Pawlowski, J., Audic, S., Adl, S., Bass, D., Belbahri, L., Berney, C., Bowser, S. S., Cepicka, I., Decelle, J., Dunthorn, M., Fiore-Donno, A. M., Gile, G. H., Holzmann, M., Jahn, R., Jirků, M., Keeling, P. J., Kostka, M., Kudryavtsev, A., Lara, E., Lukeš, J., Mann, D. G., Mitchell, E. A. D., Nitsche, F., Romeralo, M., Saunders, G. W., Simpson, A. G. B., Smirnov, A. V., Spouge, J. L., Stern, R. F., Stoeck, T., Zimmermann, J.,

Schindel, D., and de Vargas, C. (2012). CBOL Protist Working Group: Barcoding Eukaryotic Richness beyond the Animal, Plant, and Fungal Kingdoms. *PLoS Biology*, 10(11):e1001419.

Pawlowski, J., Lejzerowicz, F., Apotheloz-Perret-Gentil, L., Visco, J., and Esling, P. (2016). Protist metabarcoding and environmental biomonitoring: Time for change. *European Journal of Protistology*, 55:12–25.

Piganeau, G., Eyre-Walker, A., Grimsley, N., and Moreau, H. (2011). How and Why DNA Barcodes Underestimate the Diversity of Microbial Eukaryotes. *PLoS ONE*, 6(2):e16342.

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(6):824–834.

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(2):77–87.

Portes, J., Barrias, E., Travassos, R., Attias, M., and de Souza, W. (2020). *Toxoplasma gondii* Mechanisms of Entry Into Host Cells. *Frontiers in Cellular and Infection Microbiology*, 10:294.

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(2):58–64.

Prensier, G., Dubremetz, J.-F., and Schrevel, J. (2008). The Unique Adaptation of the Life Cycle of the Coelomic Gregarine *Diplauxis hatti* to its Host *Perinereis cultrifera* (Annelida, Polychaeta): an Experimental and Ultrastructural Study. *Journal of Eukaryotic Microbiology*, 55(6):541–553.

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(1):98–101.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.

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

Reid, A. J., Blake, D. P., Ansari, H. R., Billington, K., Browne, H. P., Bryant, J., Dunn, M., Hung, S. S., Kawahara, F., Miranda-Saavedra, D., Malas, T. B., Mourier, T., Naghra, H., Nair, M., Otto, T. D., Rawlings, N. D., Rivailler, P., Sanchez-Flores, A., Sanders, M., Subramaniam, C., Tay, Y.-L., Woo, Y., Wu, X., Barrell, B., Dear, P. H., Doerig, C., Gruber, A., Ivens, A. C., Parkinson, J., Rajandream, M.-A., Shirley, M. W., Wan, K.-L., Berriman, M., Tomley, F. M., and Pain, A. (2014). Genomic analysis of the causative agents of coccidiosis in domestic chickens. *Genome Research*, 24(10):1676–1685.

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., Sanders, M., Shanmugam, D., Sohal, A., Wasmuth, J. D., Brunk, B., Grigg, M. E., Howard, J. C., Parkinson, J., Roos, D. S., Trees, A. J., Berriman, M., Pain, A., and Wastling, J. M. (2012). Comparative Genomics of the Apicomplexan Parasites *Toxoplasma gondii* and *Neospora caninum*: Coccidia Differing in Host Range and Transmission Strategy. *PLoS Pathogens*, 8(3):e1002567.

Ricklefs, R. E. and Outlaw, D. C. (2010). A Molecular Clock for Malaria Parasites. *Science*, 329(5988):226–229.

Robert-Gangneux, F. and Dardé, M.-L. (2012). Epidemiology of and Diagnostic Strategies for Toxoplasmosis. *Clinical Microbiology Reviews*, 25(2):264–296.

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(3). doi/10.1128/mSphere.00823-20.

Ronquist, F. and Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572–1574.

Roux, C., Fraïsse, C., Romiguier, J., Anciaux, Y., Galtier, N., and Bierne, N. (2016). Shedding Light on the Grey Zone of Speciation along a Continuum of Genomic Divergence. *PLOS Biology*, 14(12):e2000234.

Rueckert, S., Betts, E. L., and Tsaousis, A. D. (2019a). The Symbiotic Spectrum: Where Do the Gregarines Fit? *Trends in Parasitology*, 35(9):687–694.

Rueckert, S. and Horák, A. (2017). Archigregarines of the English Channel revisited: New molecular data on *Selenidium* species including early described and new species and the uncertainties of phylogenetic relationships. *PLOS ONE*, 12(11):e0187430.

Rueckert, S. and Leander, B. S. (2009). Molecular Phylogeny and Surface Morphology of Marine Archigregarines (Apicomplexa), *Selenidium* spp., *Filipodium phascolosomae* n. sp., and *Platyproteum* n. g. and comb. from North-Eastern Pacific Peanut Worms (Sipuncula). *Journal of Eukaryotic Microbiology*, 56(5):428–439.

Rueckert, S. and Leander, B. S. (2010). Description of *Trichotokara nothriae* n. gen. et sp. (Apicomplexa, Lecudinidae) – An intestinal gregarine of *Nothria conchylega* (Polychaeta, Onuphidae). *Journal of Invertebrate Pathology*, 104(3):172–179.

Rueckert, S., Pipaliya, S. V., and Dacks, J. B. (2019b). Evolution: Parallel Paths to Parasitism in the Apicomplexa. *Current Biology*, 29(17):R836–R839.

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(3):e18163.

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(2):244–262.

Russell, D. G. (1983). Host cell invasion by Apicomplexa: an expression of the parasite's contractile system? *Parasitology*, 87(2):199–209.

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 Biology*, 19(1):77.

Scalzitti, N., Jeannin-Girardon, A., Collet, P., Poch, O., and Thompson, J. D. (2020). A benchmark study of *ab initio* gene prediction methods in diverse eukaryotic organisms. *BMC Genomics*, 21(1):293.

Schrével, J. and Desportes, I. (2015). Gregarines. In Mehlhorn, H., editor, *Encyclopedia of Parasitology*, pages 1–47. Springer Berlin Heidelberg, Berlin, Heidelberg.

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(4):339–368.

Semans, F. (1941). Protozoan parasites reported from the Orthoptera, with special reference to those of Ohio. III. Protozoan parasites in relation to the host and the host ecology. *Ohio Journal of Science*, (41):457–464.

Seppey, M., Manni, M., and Zdobnov, E. M. (2019). BUSCO: Assessing Genome Assembly and Annotation Completeness. In *Gene Prediction*, volume 1962, pages 227–245. Springer New York, New York, NY. Series Title: Methods in Molecular Biology.

Shaw, M. K. (2003). Cell invasion by *Theileria* sporozoites. *Trends in Parasitology*, 19(1):2–6.

Shen, B., Buguliskis, J. S., Lee, T. D., and Sibley, L. D. (2014). Functional Analysis of Rhomboid Proteases during *Toxoplasma* Invasion. *mBio*, 5(5):e01795–14.

Sibbald, S. J. and Archibald, J. M. (2017). More protist genomes needed. *Nature Ecology & Evolution*, 1(5).

Sibley, L., Håkansson, S., and Carruthers, V. B. (1998). Gliding motility: An efficient mechanism for cell penetration. *Current Biology*, 8(1):R12–R14.

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(12):1996–2006.

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(3):241–262.

Simdyanov, T. G., Guillou, L., Diakin, A. Y., Mikhailov, K. V., Schrével, J., and Aleoshin, V. V. (2017). A new view on the morphology and phylogeny of eugregarines suggested by the evidence from the gregarine *Ancora sagittata* (Leuckart, 1860) Labbé, 1899 (Apicomplexa: Eugregarinida). *PeerJ*, 5:e3354.

Simdyanov, T. G. and Kuvardina, O. N. (2007). Fine structure and putative feeding mechanism of the archigregarine *Selenidium orientale* (Apicomplexa: Gregarinomorpha). *European Journal of Protistology*, 43(1):17–25.

Simdyanov, T. G., Paskerova, G. G., Valigurová, A., Diakin, A., Kováčiková, M., Schrével, J., Guillou, L., Dobrovolskij, A. A., and Aleoshin, V. V. (2018). First Ultrastructural and Molecular Phylogenetic Evidence from the Blastogregarines, an Early Branching Lineage of Plesiomorphic Apicomplexa. *Protist*, 169(5):697–726.

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 Journal of Microbiology & Biotechnology*, 31(12):1923–1934.

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(12):2486–2495.

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 Pathogens*, 7(10):e1002280.

Slater, G. and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, 6(1):31.

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

Song, H., Mariño-Pérez, R., Woller, D. A., and Cigliano, M. M. (2018). Evolution, Diversification, and Biogeography of Grasshoppers (Orthoptera: Acrididae). *Insect Systematics and Diversity*, 2(4):3.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313.

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(Web Server):W435–W439.

Sultan, A. A., Thathy, V., Frevert, U., Robson, K. J., Crisanti, A., Nussenzweig, V., Nussenzweig, R. S., and Ménard, R. (1997). TRAP Is Necessary for Gliding Motility and Infectivity of *Plasmodium* Sporozoites. *Cell*, 90(3):511–522.

Tardieux, I. and Baum, J. (2016). Reassessing the mechanics of parasite motility and host-cell invasion. *Journal of Cell Biology*, 214(5):507–515.

Tarigo, J. L., Scholl, E. H., Bird, D. M., Brown, C. C., Cohn, L. A., Dean, G. A., Levy, M. G., Doolan, D. L., Trieu, A., Nordone, S. K., Felgner, P. L., Vigil, A., and Birkenheuer, A. J. (2013). A Novel Candidate Vaccine for Cytauxzoonosis Inferred from Comparative Apicomplexan Genomics. *PLoS ONE*, 8(8):e71233.

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(2):235–248.

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):1–17.

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*, page 12;8:e42669.

Uvarov, B. (1977). *Grasshoppers and Locusts.*, volume 2. London, centre for overseas pest research edition.

Valigurová, A. and Florent, I. (2021). Nutrient Acquisition and Attachment Strategies in Basal Lineages: A Tough Nut to Crack in the Evolutionary Puzzle of Apicomplexa. *Microorganisms*, 9(7):1430.

Valigurová, A., Hofmannová, L., Koudela, B., and Vávra, J. (2007). An Ultrastructural Comparison of the Attachment Sites Between Gregarina steini and Cryptosporidium muris. *The Journal of Eukaryotic Microbiology*, 0(0):071116223551001–???

Valigurová, A. and Koudela, B. (2008). Morphological analysis of the cellular interactions between the eugregarine *Gregarina garnhami* (Apicomplexa) and the epithelium of its host, the desert locust *Schistocerca gregaria*. *European Journal of Protistology*, 44(3):197–207.

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(6):e0179709.

Valigurová, A., Vaškovicová, N., Musilová, N., and Schrével, J. (2013). The enigma of eugregarine epicytic folds: where gliding motility originates? *Frontiers in Zoology*, 10(1):57.

Van Beneden (1869). Sur une nouvelle espèce de Grégarine désignée sous le nom de *Gregarina gigantea*. *Bulletins de l'Académie Royale de Belgique*, 28(7):444–456.

Wakeman, K. C. and Leander, B. S. (2012). Molecular Phylogeny of Pacific Archigregarines (Apicomplexa), Including Descriptions of *Veloxidium leptosynaptae* n. gen., n. sp., from the Sea Cucumber *Leptosynapta clarki* (Echinodermata), and Two New Species of *Selenidium*. *Journal of Eukaryotic Microbiology*, 59(3):232–245.

Weiss, L. M. and Kim, K. (2014). *Toxoplasma gondii, The Model Apicomplexan - Perspectives and Methods*. Elsevier.

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., Ward, G. E., and Meissner, M. (2017). Surface attachment, promoted by the actomyosin system of *Toxoplasma gondii* is important for efficient gliding motility and invasion. *BMC Biology*, 15(1).

Wickham, H. (2009). *Ggplot2: elegant graphics for data analysis*. Use R! Springer, New York. OCLC: ocn382399721.

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(7):691–702.

Woo, Y. H., Ansari, H., Otto, T. D., Klinger, C. M., Kolisko, M., Michálek, J., Saxena, A., Shanmugam, D., Tayyrov, A., Veluchamy, A., Ali, S., Bernal, A., del Campo, J., Cihlář, J., Flegontov, P., Gornik, S. G., Hajdušková, E., Horák, A., Janouškovec, J., Katris, N. J., Mast, F. D., Miranda-Saavedra, D., Mourier, T., Naeem, R., Nair, M., Panigrahi, A. K., Rawlings, N. D., Padron-Regalado, E., Ramaprasad, A., Samad, N., Tomčala, A., Wilkes, J., Neafsey, D. E., Doerig, C., Bowler, C., Keeling, P. J., Roos, D. S., Dacks, J. B., Templeton, T. J., Waller, R. F., Lukeš, J., Oborník, M., and Pain, A. (2015). Chromerid genomes reveal the evolutionary path from photosynthetic algae to obligate intracellular parasites. *eLife*, 4.

Woods, K., Perry, C., Brühlmann, F., and Olias, P. (2021). *Theileria*'s Strategies and Effector Mechanisms for Host Cell Transformation: From Invasion to Immortalization. *Frontiers in Cell and Developmental Biology*, 9:662805.

Wu, T. D. and Watanabe, C. K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21(9):1859–1875.

Yamagishi, J., Asada, M., Hakimi, H., Tanaka, T. Q., Sugimoto, C., and Kawazu, S.-i. (2017). Whole-genome assembly of *Babesia ovata* and comparative genomics between closely related pathogens. *BMC Genomics*, 18(1).

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(9):e12745.

Yu, L., Liu, Q., Zhan, X., Huang, Y., Sun, Y., Nie, Z., Zhao, Y., An, X., Li, M., Wang, S., Ao, Y., Huang, C., He, L., and Zhao, J. (2018). Identification and molecular characterization of a novel *Babesia orientalis* thrombospondin-related anonymous protein (BoTRAP1). *Parasites & Vectors*, 11(1):667.

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(2):190–198.

Biol. Cell (2020) **112,** 173–185 DOI: 10.1111/boc.202000006

Review

Why the –omic future of Apicomplexa should include gregarines

Julie Boisard*† and Isabelle Florent*[1]

*Molécules de Communication et Adaptation des Microorganismes (MCAM, UMR 7245), Département Adaptations du Vivant (AVIV), Muséum National d'Histoire Naturelle, CNRS, Paris Cedex 05, France and †Structure et instabilité des génomes (STRING UMR 7196 CNRS / INSERM U1154), Département Adaptations du Vivant (AVIV), Muséum National d'Histoire Naturelle, Paris Cedex 05, France

Gregarines, a polyphyletic group of apicomplexan parasites infecting mostly non-vertebrates hosts, remains poorly known at taxonomic, phylogenetic and genomic levels. However, it represents an essential group for understanding evolutionary history and adaptive capacities of apicomplexan parasites to the remarkable diversity of their hosts. Because they have a mostly extracellular lifestyle, gregarines have developed other cellular developmental forms and host–parasite interactions, compared with their much better studied apicomplexan cousins, intracellular parasites of vertebrates (Hemosporidia, Coccidia, Cryptosporidia). This review highlights the promises offered by the molecular exploration of gregarines, that have been until now left on the side of the road of the comparative –omic exploration of apicomplexan parasites. Elucidating molecular bases for both their ultrastructural, functional and behavioural similarities and differences, compared with those of the typical apicomplexan models, is expected to provide entirely novel clues on the adaptive capacities developed by Apicomplexa over evolution. A challenge remains to identify which gregarines should be explored in priority, as recent metadata from open and host-associated environments have confirmed how underestimated is our current view on true gregarine biodiversity. It is now time to turn to gregarines to widen the currently highly skewed view we have of adaptive mechanisms developed by Apicomplexa.

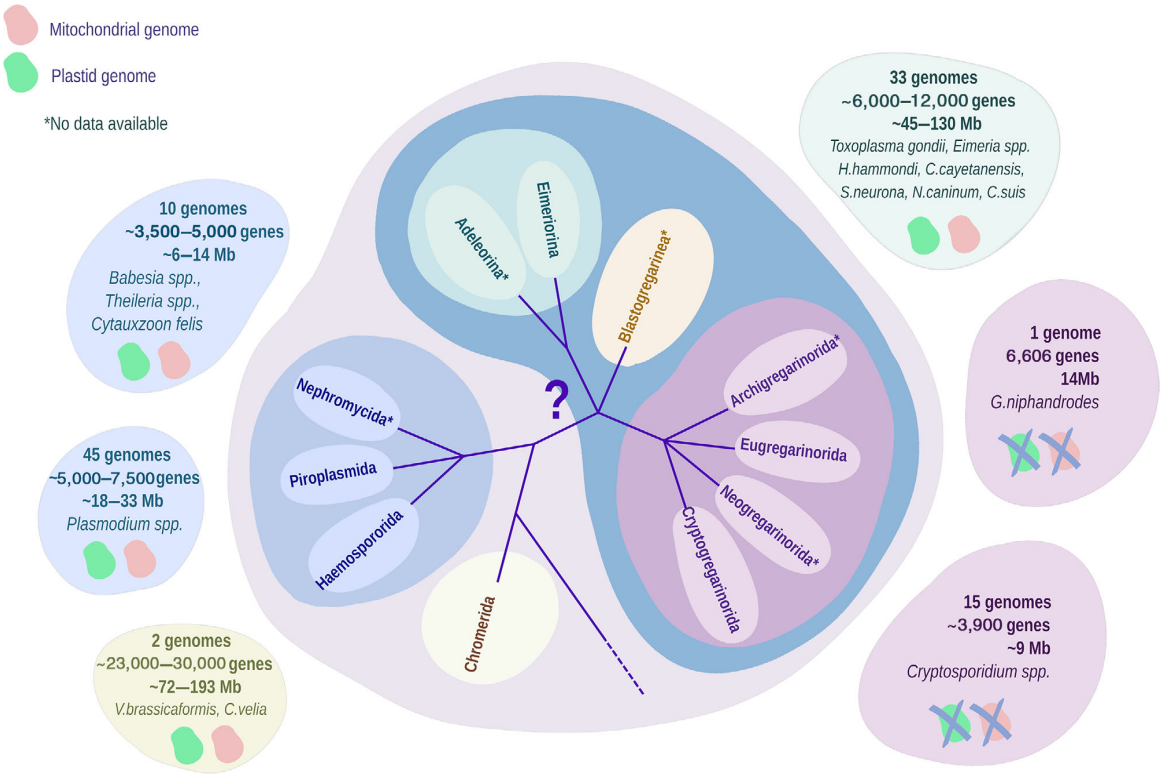
Introduction

Apicomplexa are unicellular eukaryotes (protists) collectively corresponding to ~350 genera and ~6000 named species, the wide majority of which have adopted a strict parasitic lifestyle in a very wide diversity of metazoan hosts (Portman and Slapeta, 2014; Adl et al., 2018). Apicomplexa, together with the two sister groups Dinoflagellata and Ciliata, form the supergroup Alveolata, at the base of which are Rhizaria and Stramenopiles/Phaeophyta (Adl et al., 2018). Apicomplexa are mostly known for comprising infamous intracellular parasites of vertebrates responsible for important human diseases such as malaria due to *Plasmodium* spp., cryptosporidiosis due to *Cryptosporidium* spp. and toxoplasmosis due to *Toxoplasma gondii*. Apicomplexa also comprise diverse other intracellular parasites of vertebrates, with economical or veterinary importance such as *Eimeria* spp., *Babesia* spp. and *Theileria* spp. These apicomplexan parasites have simple to very complex life cycles. Some are restricted to single hosts (monoxenous parasites, e.g. *Cryptosporidium, Eimeria*), other alternate between two successive hosts (dixenous parasites such as *Plasmodium*, *Babesia* and *Theileria*, completing sexual reproduction in various insects or arthropods and asexual phases in various tissues of vertebrates) and few have the capacity to infect multiple hosts (polyxenous parasites such as *Toxoplasma*, completing its sexual reproduction in cats and several asexual phases in various tissues of diverse warm blooded vertebrates). Due to their medical, veterinary or economical importance, and because it has been

[1]To whom correspondence should be addressed (email: isabelle.florent@mnhn.fr)
Key words: Apicomplexa, Evolutionary history, Genomics, Parasitology, Protozoa.
Abbreviations: SEM, scannning electron microscopy; TEM, transmission electron microscopy.

Figure 1 | The genomic panorama of Apicomplexa

On this schematic representation inspired by Portman and Slapeta (2014) and using the most recent taxonomy by Adl et al. (2018), we have compiled the genomic information currently available on Apicomplexa (104 genomes) and proto-Apicomplexa (Chromerida, two genomes), mostly available from EupathDB (Aurrecoechea et al., 2017). For each group of data (six in total) we have indicated: the number of available genomes, specifying the concerned species, the number of protein-coding genes, the nuclear genome size in Mb and the presence (or absence in *Cryptosporidium* spp; the question being still open in the case of the (unpublished) *G. niphandrodes* genomic data), of mitochondrial or plastid genomes. The question mark symbolises the currently unresolved branching order of the various apicomplexan groups.



possible to cultivate most of them in laboratory conditions, their genomes have been deciphered (~100, deposited into the EupathDB database (Aurrecoechea et al., 2017), Figure 1) and are major references for medical investigations, comparative genomics studies and exploration of evolutionary history of apicomplexan parasites (Janouskovec et al., 2015; Woo et al., 2015; Janouskovec et al., 2019; Kwong et al., 2019).

But Apicomplexa also comprise another group of organisms collectively known as 'gregarines', that are principally monoxenous parasites of a wide diversity of non-vertebrate metazoan hosts, ranging from Polychaeta annelids to tunicates, arthropods and mollusks, in which they develop mostly extracellularly – contrary to the above-mentioned parasites – in the intestinal and coelomic cavities of their hosts (Desportes and Schrével, 2013). These endoparasites are mostly considered as being non-pathogenic, with a few reported cases of recognised pathogenicity, while it is clear that experimental studies focusing on gregarine pathogenicity are mostly lacking (Rueckert et al., 2019a).

Gregarines: well described Apicomplexa though forgotten at molecular level

There are several consequences to the mostly extracellular lifestyle displayed by gregarines, a feature that indeed distinguishes them from their intracellular parasites relatives (Desportes and Schrével, 2013; Schrével and Desportes, 2015). First, gregarines can reach very large sizes even for unicellular eukaryotes, from less than a μm to more than a mm for respectively the zoite and trophozoite forms of the marine eugregarine *Porospora gigantea,* intestinal parasite of the lobster *Homarus gammarus* (Desportes and Schrével, 2013; Schrével and Desportes, 2015). Insect eugregarines, such as *Gregarina garnhami*, intestinal parasite of the desert locust *Schistocerca gregaria,* display trophozoite forms reaching a dozen to several hundred μm (Desportes and Schrével, 2013; Schrével and Desportes, 2015). Archigregarine trophozoites, such as *Selenidium pendula*, intestinal parasite of the Polychaeta *Scolelepis (Nerine) squamata* also reach several dozen μm (Schrével et al., 2016). The large sizes of these developmental stages and their very common occurrence in a large diversity of non-vertebrate metazoan hosts have facilitated the discovery and biological studies on gregarines, resulting in an abundant literature on their morphologies, ultrastructures and life cycles, that have been examined by photonic and electron microscopy (SEM, TEM) imaging or by using cryoelectron microscopy or immunofluorescence (see (Desportes and Schrével, 2013) for exhaustive bibliography). Dynamic recordings are also available notably concerning their movements or behaviours (Desportes and Schrével, 2013). This rich literature offers a wide panorama of the adaptive capacities of these organisms to their hosts and environments, awaiting now exploration by –omic approaches to decipher the molecular bases of their functioning and variations, as it has been developed so far for their intracellular apicomplexan cousins (Rueckert et al., 2019b). Indeed, the genomic knowledge on apicomplexan is currently highly biased in favour of intracellular parasites of vertebrate hosts, belonging to Haematozoa, Coccidia and to a lesser extent, Cryptosporidia (Figure 1). Gregarines have been so far mostly excluded from this –omic exploration, to the exception of (unpublished) genome of the terrestrial insect eugregarine *Gregarina niphandrodes* (accessible at EupathDB (Aurrecoechea et al., 2017), section CryptoDB), intestinal parasite of *Tenebrio molitor*, very partial genomic data on insect eugregarine *Ascogregarina taiwanensis* (Templeton et al., 2010) intestinal parasite of *Aedes albopictus* and partial and recently emerging transcriptomic data for a dozen of terrestrial and marine gregarine species (Omoto et al., 2004; Janouskovec et al., 2019; Mathur et al., 2019) (see Figure 1 for illustration on Apicomplexa genomic data knowledge).

There are several reasons why the acquisition of –omic knowledge on gregarines is lagging behind that of their intracellular vertebrate parasite cousins. (1) As there are infecting 'only' non-vertebrates hosts and are mostly considered non-pathogenic (Rueckert et al., 2019a), they have been neglected. (2) The current lack of *in vitro* culture methods for these parasites complicates the isolation of biological material in adequate amounts and quality for accurate usage in molecular investigations. Indeed, gregarine biological studies must rely on field collections, mostly from infected hosts (alternatively their feces), which exposes the collected material to contaminations by host cells and environmental microorganisms. The ability to maintain the hosts of a diversity of gregarines in laboratory conditions, offers a good compromise as it allows regular access to different developmental stages amenable to a variety of cellular (microscopy, test of inhibitors) and molecular (–omics) studies (see (Desportes and Schrével, 2013) for exhaustive descriptions). For example, *Neanthes* (*Nereis*) *divesicolor*, the host of the marine eugregarine *Lecudina tuzetae*, can be maintained for several months in natural or artificial sea water in the laboratory (Kuriyama et al., 2005; Desportes and Schrével, 2013). Also, several insect raising facilities may be used to get access to their infecting gregarines (Clopton, 2009; Desportes and Schrével, 2013). Also, it can be expected that the concomitant progression of –omics and microscopic methodologies, allowing using increasingly reduced amounts of biological material, will also facilitate in a near future the bridging of this molecular knowledge gap, between the currently very poorly documented non cultivable gregarines and the increasingly well documented cultivable intracellular parasites of vertebrates (Gawad et al., 2016).

Why should we study gregarines?

But why should we focus on gregarines at molecular level? What could they tell us that we do not already know about apicomplexan parasites?

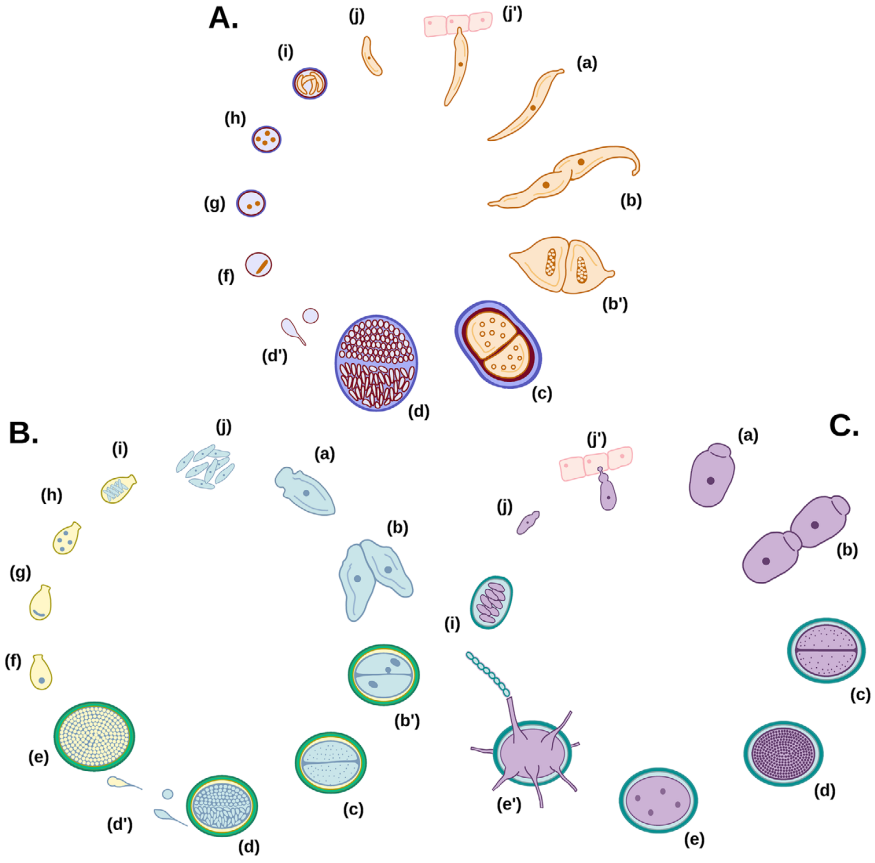
Because gregarines are Apicomplexa, although particular ones, with unique differences notably their mostly extracellular life mode and its biological consequences (Desportes and Schrével, 2013; Schrével and Desportes, 2015; Adl et al., 2018; Rueckert et al., 2019a). Like all Apicomplexa, gregarines present at least once during their life cycle a developmental form called zoite, a polarised cell comprising the so-called 'apical complex' composed of scaffolding cytoskeletal elements enclosing specialised apical organelles (rhoptries, micronemes, dense granules) that gave the name to this phylum replacing the former Sporozoa (Morrissette and Sibley, 2002; Tardieux and Baum, 2016; Adl et al., 2018). Gregarines (notably, Archigregarines) do have a conoid, composed of spirally arranged array of microtubules as found in Coccidia and Cryptospodia (forming Conoidasida), but secondarily lost in Haemosporidia (forming Aconoidasida) ((Portman and Slapeta, 2014; Adl et al., 2018), Figure 1). In Apicomplexa having intracellular developmental phase(s) this 'apical complex' has been clearly involved in the recognition and invasion of host cells, allowing parasites establishment and development in this novel ecological niche (Tardieux and Baum, 2016; Hakimi et al., 2017). In Apicomplexa displaying extracellular lifestyle as most intestinal gregarines do, this 'apical complex' appears to have a different role. Yet involved in parasite attachment to host cells at sporozoite stage, it subsequently appears mostly used for parasite feeding, sustaining spectacular growth phases, and not for achieving tissue penetration or parasite internalisation within host cells (Valigurova and Koudela, 2008; Simdyanov and Kuvardina, 2007; Schrével et al., 2016). This to the notable exception of coelomic (eu)gregarines capable of intestinal barrier crossing and neogregarines, capable of reaching intracellular niches in some of their hosts tissues (Desportes and Schrével, 2013; Schrével and Desportes, 2015).

There are several consequences to this extracellular growth of these trophozoite forms in gregarines: (1) an extremely wide diversity of shapes and sizes, as mentioned above and largely used for taxonomical purposes (Desportes and Schrével, 2013); (2) a sexual phase (gamogony followed by sporogony) that also occurs extracellularly, producing developmental forms that are particular to gregarines, starting with the syzygy (Desportes and Schrével, 2013; Schrével and Desportes, 2015).

Syzygy is the named given to the developmental stage that precludes the gregarine sexual reproduction. It corresponds to the bi-association of two trophozoites after they have detached from host cell and which are future gamonts (Figure 2, stages (a) and (b/b′)). Although morphologically similar both in size and shape, the two partners of the syzygy are committed to evolve into respectively male and female gamonts (Figure 2, stages (c)). Depending on the species, the bi-association may be caudo-caudal (Figure 2A, stage (c)), lateral (Figure 2B, stage (c)), fronto-frontal (not shown, see (Desportes and Schrével, 2013)) or caudo-frontal (Figure 2C, stage (c)). In this latter case, found in *Gregarina garnhami*, the anterior partner of the syzygy is called primite the other satellite (Figure 2C, stage (c)). Occasionally, syzygy associations may involve more than two partners but the evolution of such 'ménage à trois' has not been yet examined at molecular level (Desportes and Schrével, 2013). The evolution of the syzygy bi-association is a globular structure called gametocyst, initially composed of two hemispheres of similar shape and volume around which a thick wall is elaborated (Figure 2, stages (c)). Series of nuclear divisions with final cytokinesis (gamogony) then occur within each hemisphere producing male gametes within one hemisphere and female gametes within the other one (Figure 2, stages (d)). A clear anisogamy is commonly observed between male gametes – more pyriform and usually flagellated – and female gametes – more globular and non-flagellated (Figures 2A and 2B, stages (d′)). It is therefore only after complete gametes production that the 'sex' of gamonts may be deduced. Numerous imaging recordings have been performed to study the cellular events occurring during this first phase of the gregarine sexual reproduction called gamogony, with a remarkable confocal imaging analysis performed in the case of the marine eugregarine *Lecudina tuzetae*, intestinal parasite of the Polychaeta *Hediste* (*Nereis*) *diversicolor*, in which ~5000 male and as many female gametes are produced per gametocyst (Kuriyama et al., 2005).

Figure 2 | Representative development cycles for three gregarines

The developmental cycles of: (**A**) the (marine) archigregarine *Selenidium pendula*, intestinal parasite of the Polychaeta *Scolelepis* (*Nerine*) *squamata*, adapted from (Schrével and Desportes, 2015); (**B**) the marine eugregarine *Lecudina tuzetae,* intestinal parasite of the Polychaeta *Hediste* (*Nereis*) *diversicolor*, adapted from (Schrével and Desportes, 2015); (**C**) the terrestrial eugregarine *Gregarina garnhami* (C), intestinal parasite of the desert locust *Schistocerca gregaria* (C), based on (Canning, 1956) and personal observations. The drawings use identical 'legend letters' to designate similar developmental stages across the three cycles. (**A**) *Selenidium pendula*: (a) detached trophozoite; (b) caudal syzygy; (b′) particularity in syzygy for this species (nuclear modifications before encystment); (c) gametocyst undergoing gamogony; (d) gametocyst with fully differentiated gametes; (d′) details of male (flagellated) and female (ovoid) gamete; (e) zygote ready to undergo sporogony yielding stages with two nuclei (g), then four nuclei (h); (i) spore containing four sporozoites; (j) released sporozoite (in host) starting vegetative phase. (**B**) *Lecudina tuzetae*. (a) detached trophozoite; (b) lateral syzygy; (c) gametocyst undergoing gamogony; (d) gametocyst with fully differentiated gametes; (d′) details of male (flagellated) and female (ovoid) gamete; (e) sporokyst enclosing ~5000 zygotes ready to evolve into spores (f) eventually undergoing sporogony yielding stages with two nuclei (g), four nuclei (h); (i) spore containing eight sporozoites; (j) released sporozoite (in host) starting vegetative phase including attachment to host epithelial cell (not shown). (**C**) *Gregarina garnhami*. (a) detached trophozoite; (b) caudo-frontal syzygy (primite ahead, satellite following); (c) gametocyst undergoing gamogony; (d) gametocyst with fully differentiated gametes; (e) sporokyst enclosing zygotes ready to evolve into spores undergoing sporogony (details not shown); these spores are released in the environment as spore chains (ch) through sporoducts (sp) emerging from the sporokyst; (i) spore containing eight sporozoites; (j) released sporozoite (in host) starting vegetative phase including attachment to host epithelial cell (j′). Cyst or spore walls surround developmental stages from (c) to (i).



It is interesting to notice that, depending on the species, the length of the syzygy phase might be particularly long; in the atypical case of *Diplauxis hatti*, coelomic parasite of the Polychaeta *Perinereis cultrifera*, this bi-association remains stable for more than two years, awaiting host's sexual maturation to engage into gamogony (Prensier et al., 2008).

Once the male and female gametes have been produced within their respective compartments, their mixing occurs within the gametocyst. Thousands of fertilisations then take place simultaneously during a process called 'gamete dance' in *L. tuzetae,* which lasts ~4 h and produces ~5000 zygotes per gametocyst (Kuriyama et al., 2005) (Figure 2B stage (e)). Sporogony then begins and the gametocyst takes the name of sporokyst (Figures 2B and 2C, stages (e/e′)). Each zygote secretes a cyst wall (stage called immature sporocyst) and undergoes meiosis and additional mitosis (in eugregarines and neogregarines) leading to sporozoites formation (Figure 2, stages (f) to (i)). Each spore, also called sporocyst, therefore possesses a thick wall and is the form of dissemination of the gregarine in the environments (Desportes and Schrével, 2013; Schrével and Desportes, 2015). It contains 4 sporozoites in the case of archigregarines and 8 sporozoites in eugregarines and neogregarines ((Desportes and Schrével, 2013; Schrével and Desportes, 2015), see also Figure 2 stages (i)). In Coccidia, sporozoites are also formed within a spore that however is called oocyst, whereas in gregarines it is called sporocyst (Desportes and Schrével, 2013; Schrével and Desportes, 2015). In marine eugregarines, spores are eventually released in the environments with the breaking down of the sporokyst (Figure 2B stage (e) in the case of *L. tuzetae*). In terrestrial gregarines, spores are released in the environments *via* sporoducts that are formed at the surface of the sporokyst (see for example *Gregarina garnhami* Figure 2B stage (e′)). The progeny in terms of number of spores, resulting from the evolution of a single syzygy, is considerable and can reach few thousands to several millions depending on the gregarine species (Desportes and Schrével, 2013; Schrével and Desportes, 2015). Sporokysts are therefore obviously a material of choice to isolate genomic DNA. Sporocysts on the other hand, are the developmental forms that are likely collected within soils and sediments from environments (see below).

Once ingested by hosts, the spores undergo dehiscence after passage through the host's digestive system and sporozoites are released (four in the case of archigregarines, eight in the case of eugregarines and neogregarines, Figure 2 stages (j)). These will be able to attach to their host's intestinal cells (Figure 2C, (j′)), using their apical end displaying typical 'apical complex' features, and these attached sporozoites will start to grow dramatically, evolving into trophozoites. In the case of *G. garnhami* for example, trophozoites grow from less than 10 μm to over 400 μm in length within a single host. The size range of trophozoites may therefore cover two to three orders of magnitude depending on the species (Desportes and Schrével, 2013; Schrével and Desportes, 2015). This gigantism, achieved for a large number of gregarine species, offers a remarkable material for cell biology explorations and immunofluorescence imaging for example (Kuriyama et al., 2005; Valigurova et al., 2013).

This radical difference in lifestyle including gamogonic and sporogonic phases, contrasting from those observed in Hematozoa and Coccidia, results also in very different interactions between gregarines and their hosts, which raises many questions about the adaptive pathways they have developed, as apicomplexan parasites, to survive over the course of evolution. Which molecular solutions have they then developed to survive within the host's intestinal tracts or other cavities, maintain survival, acquire nutriments, complete (a)sexual reproduction, with a remarkable success if one considers their wide occurrence in such a high diversity of endoparasitic non-vertebrate hosts contexts? (Desportes and Schrével, 2013; Schrével and Desportes, 2015).

In order to widen our view on host-adaptive modes in Apicomplexa, we propose thereafter some possible exploration topics focusing on gregarines, which are obviously not exhaustive as the number of biological questions that may be asked is far more important.

Gregarines, as apicomplexan parasites, do possess a fully developed apical complex, at least in some developmental stages (sporozoites, trophozoites) (Desportes and Schrével, 2013)

Biological and morphological studies have established that in gregarines, this apical complex is used for host cell attachment to allow the parasite to feed from its host cell by a process known as myzocytosis (Schrével et al., 2016; Simdyanov and Kuvardina, 2007). The host cell penetration by the parasite is

not complete as the gregarine remains extracellular with only its apical end intimately engaged in a host–parasite interplay that has been studied at microscopic level but whose molecular actors are poorly defined (Valigurova et al., 2007; Schrével et al., 2016; Simdyanov et al., 2017). See also (Desportes and Schrével, 2013) for exhaustive descriptions of several additional examples. Therefore, the biological function of 'gregarine apical complex' only partly overlaps the biological function currently attributed to 'apicomplexan parasite apical complex', that is recognition and invasion of their host cell (Tardieux and Baum, 2016).

Several questions therefore emerge: to which extent does the molecular architecture of 'gregarine apical complexes' compares to that of their best known cousins, that is *Toxoplasma* and *Plasmodium* (Boucher and Bosch, 2015; Tardieux and Baum, 2016)? Are the 'scaffolding', 'recognition', 'invasive' and 'nutrition' functions fulfilled by the same molecular components or by other ones? What are the composition and functional roles of micronemes, rhoptries and dense granules in gregarines? In intracellular apicomplexan parasites of vertebrates (*Plasmodium* and *Toxoplasma*), these secretory organelles are documented to intervene sequentially in an orchestrated manner, with first micronemes secreting parasite proteins involved in host cell recognition as well as AMA-1, which, when combined to a defined cortege of RON proteins (secreted by the neck of rhoptries) will assemble into the so-called mobile junction essential to the invasion process of host cells by *Plasmodium* merozoites or *Toxoplasma* tachyzoites (reviewed in (Tardieux and Baum, 2016)). Subsequently, ROP proteins (secreted by the bulb of rhoptries) and GRA proteins (secreted by the dense granules) will allow establishing the intracellular niche for these parasites, either at the parasitophorous vacuole level (facilitating metabolite exchanges) or beyond this border to manipulate the host cell program to the benefit of the parasite (Hakimi et al., 2017). Although it may be expected that gregarine micronemes, rhoptries and dense granules will have their own protein repertoires (there are currently very limited overlaps of these repertoires between currently described apicomplexan genus (Counihan et al., 2013; Boucher and Bosch, 2015; Hakimi et al., 2017)) it will be interesting to decipher their specific roles and how they contribute (or not) to establishing also the host-parasite interface. It is important here to indicate that there are alternative invasive modes in apicomplexan parasites such as *Theileria* and *Cryptosporidium* that differ from the better described *Toxoplasma*/*Plasmodium* mode (Gubbels and Duraisingh, 2012). What are the molecular similarities between the *T. gondii* and *P. falciparum* parasitophorous vacuole make up and the food vacuole of gregarines, that forms at the gregarine–host cell interface (Valigurova et al., 2007; Schrével et al., 2016; Simdyanov et al., 2017)? Or is the similarity stronger to the feeder organelle of epicellular *Cryptosporidium?* (Barta and Thompson, 2006; Bartosova-Sojkova et al., 2015).

Their mostly extracellular development

A first morphological and biological consequence of this particular behaviour selected over evolution is the fact mentioned above that gregarine trophozoites can reach very large sizes contrary to intracellular parasites of vertebrates (1–10 μm at most). A second is that their sexual phase is also extracellular, starting with the syzygy that evolves into gametocysts then sporokysts producing sporocysts (Figure 2). These developmental forms are strikingly distinct from the developmental forms encountered in *Toxoplasma*, *Plasmodium* and even *Cryptosporidium* (Aly et al., 2009; Robert-Gangneux and Darde, 2012; Bouzid et al., 2013). Interestingly, the oocysts forms of *Toxoplasma* and *Cryptosporidium*, which are also extracellular and disseminated with their hosts' feces as resistant forms in the environments, are however elaborated intracellularly, within their hosts' intestinal cells (Robert-Gangneux and Darde, 2012; Bouzid et al., 2013). Indeed, in *Toxoplasma* and *Cryptosporidium* the gamogony remains intracellular, whereas it is extracellular in gregarines. An important consequence is that gregarines thus display totally different types of host-parasite interactions, having other constraints to face such as surviving in host-gut environment. Several studies have explored the permeability of the trophozoite membrane in link with the question of their nutrition mode (Desportes and Schrével, 2013; Schrével and Desportes, 2015). This questions the molecular nature and the biological role of their inner membrane complex (imc), which, interestingly, may be acting as a continuous 'shield' all around the trophozoite as it is only interrupted at the conoidal opening through which nutrition occurs (see (Schrével et al., 2016) for the case of *S. pendula* or (Kuriyama et al., 2005) for the case of *L. tuzetae*).

Interestingly, as the syzygy evolves towards the gametocyst form, this imc appears to be disassembled concomitantly with the secretion, by the gametocyst, of the protective cyst wall (see (Kuriyama et al., 2005) for details on *L. tuzetae*). This suggests that in gregarines, one form of shielding (imc) in trophozoites and syzygies gives place to another form of shielding (cyst wall) during gamogony then sporogony, both of which being intended to isolate the gregarine from its hostile (gut) environment. Obviously, the molecular exploration of such a parasite–environment (host) interplay will reveal novel adaptive features developed by Apicomplexa over evolution. To which extent the host-gut environment is less hostile in an invertebrate host rather than in a vertebrate one, notably regarding immune system response and microbiota regulation, should certainly be taken into account. First, invertebrate hosts rely mainly on innate immunity to fight intruders while vertebrates also have adaptive defence mechanisms (Buchmann, 2014). In addition, the coexisting microbiome is notoriously less complex and diverse in invertebrates than vertebrates (Bahrndorff et al., 2016). These differences could explain the capacity of gregarines to self-maintain in such host-gut environment for extended life cycle times while parasites of vertebrates have been constrained to invade host cells to achieve their maintenance in hosts. Further studying not only gregarines but also their host's immune and microbiota responses will certainly clarify the contribution of these host-specific features to the diversity of gregarines behaviours and life traits over evolution.

Gregarines have developed a wider diversity of motility and mobility modes than what is mostly described (and deeply studied at molecular level) for intracellular parasites of vertebrates: the gliding motility (Frenal et al., 2017)

This movement, governed by an acto-myosin motor, involves dozens of proteins that appear (so far) well conserved between apicomplexan species (Boucher and Bosch, 2015; Frenal et al., 2017; Mueller et al., 2017). Whether the gliding components are conserved also in gregarines that move by gliding (i.e., most eugregarines) remains to be established. However, gregarines do display other modes of motility such as rolling, bending (notably archigregarines (Desportes and Schrével, 2013)), the molecular bases of which are currently totally unknown. Do these alternative modes involve molecular components shared with those of the glideosome? Do they involve other components, inherited from the putative ancestor and possibly lost secondarily in intracellular parasites of vertebrates (Janouskovec et al., 2015; Woo et al., 2015; Füssy and Oborník, 2017)? Or do they involve novel components, repositioned from the ancestor heritage or acquired by horizontal gene transfer? All this remains to be established for the diversity of known and to-be-discovered gregarines. Answers to these questions will be precious to understand how such a diversity of motility/mobility modes may have emerged for apicomplexan from a common ancestral genetic heritage, as apicomplexan derive from ancestral biflagellated organisms with repositioning of some of the former flagellar components into the apical complex structure and functioning (Janouskovec et al., 2015; Woo et al., 2015; Füssy and Oborník, 2017). This knowledge will also be precious to widen our current knowledge of the adaptive capacities to hosts developed by these remarkable apicomplexan parasites. Indeed, evolutionary molecular studies on this point have established that gliding components partially existed in the common ancestor of Apicomplexa (Janouskovec et al., 2015; Woo et al., 2015; Füssy and Oborník, 2017) but have been repositioned to be functionally operational in intracellular parasites of vertebrates. What paths of specialisation did they then follow to generate such a diversity of movements in gregarines? And, corollary to these observations, can we formulate the hypothesis that the lack of host-cell fully invasive capacities of gregarines may be due either to absence of gliding capacities despite a developed/expected to be functional apical complex (in the case of archigregarines) or, conversely, to an under-developed apical complex despite operational gliding capacities (in the case of eugregarines)?

It is therefore now time to turn to gregarines to explore these adaptive traits at molecular level, but the next question is; which ones to select and from which extent of diversity? Indeed, a recent convergence of novel data clearly indicates that the current inventory of the true gregarine diversity is dramatically underestimated – and therefore, corollary, all the relevant biological models may have not yet been discovered.

The true gregarine biodiversity

Our current understanding of gregarine biodiversity comes from three sources of information that

only partly overlap: (1) the number of formally inventoried species (Portman and Slapeta, 2014); (2) the number of species theoretically computable based on the inventory and diversity of their hosts (Desportes and Schrével, 2013); (3) environmental or host-associated metagenomic or metabarcoding approaches, that have revealed novel molecular signatures, sufficiently related to gregarines to allow taxonomic affiliation to this group but sufficiently divergent to strongly suggest novel taxonomic species (de Vargas et al., 2015; Mahé et al., 2017; Del Campo et al., 2019). The regular cross-referencing of these three sources of data leads to a permanent readjustment of both the taxonomy and phylogeny of these species, so that it is safe to say that the current biodiversity of gregarines is a field of investigation under construction, whose physiognomy is likely to evolve considerably in the coming years.

Formally described species

Regarding the first point, there are currently ~1770 formally described gregarine species, unequally distributed between archigregarines (~20), eugregarines (~1700) and neogregarines (~50) (Portman and Slapeta, 2014). In parallel, taxonomic and phylogenetic revisions concerning gregarines are a currently very active – but quite unstable – field with a diversity of successive proposals regarding their phylogenetic inter-relations as well as with other apicomplexan parasites (Cavalier-Smith, 2014; Schrével et al., 2016; Rueckert and Horak, 2017; Simdyanov et al., 2017; Simdyanov et al., 2018). Molecular phylogenies are nowadays mostly based on usage of SSU rDNA marker, more rarely complete ribosomal loci (Diakin et al., 2016; Diakin et al., 2017; Simdyanov et al., 2018). Studies based on the SSU rDNA marker alone are fairly effective in defining monophyletic groups at the genus or family levels, but fail to robustly resolve the respective branches' relationships at higher taxonomic level. Attempts to improve phylogenies using the full ribosomal marker (18S SSU + 28S LSU rDNA) have provided some progress, but have the important disadvantage to be currently available for only very few gregarine species (~20) (Diakin et al., 2016; Diakin et al., 2017; Simdyanov et al., 2018). Phylogenies relying on multiple genes (or more accurately proteins) sequences derived from genome/transcriptome investigations are now emerging but remain restrained to a dozen of gregarine species (Janouskovec et al., 2019; Mathur et al., 2019). Although they are indeed expected to be more resolving, the number of concerned species is even smaller and ambiguities remain in the interrelationships between groups, since the position of the genus *Cryptosporidium* is for example unstable between the two studies (Janouskovec et al., 2019; Mathur et al., 2019). These genome/transcriptome studies have however also shown another interest: some species historically described as gregarines (*Platyproteum* sp. and *Digyalum oweni*) do not actually appear anymore to be part of Apicomplexa (Janouskovec et al., 2019; Mathur et al., 2019).

Extrapolation based on host diversity and host-parasite behaviours

Concerning the second point, it is clearly documented that gregarines parasite virtually all non-vertebrate metazoan groups, from Polychaeta annelids to tunicates, arthropods and mollusks (Desportes and Schrével, 2013; Schrével and Desportes, 2015). For a long time, experts in the gregarine field have argued that, given the currently documented diversity of gregarines, their lifestyle principally monoxenous and their apparently narrow host-range specificity, the real biodiversity of gregarines is probably several orders of magnitude underestimated in particular for those infecting insects (~half of metazoan diversity) (Desportes and Schrével, 2013; Schrével and Desportes, 2015). In addition, most known hosts are infected by several gregarine species, as for example the mealworm *Tenebrio molitor* that is infected by at least three gregarine species (*Gregarina cuneat*a, *Gregarina polymorpha* and *Gregarina steini*) (Clopton et al., 1992) and also *G. niphandrodes*. In consequences, some experts have even proposed that the real gregarine diversity should exceed that of their hosts, making it one of the most widespread groups of organisms in the environments (Desportes and Schrével, 2013; Schrével and Desportes, 2015).

Environmental metadata uncovering a wide and mostly undescribed diversity

Thirdly, several environmental and host-associated metagenomics/metabarcoding surveys have recently not only confirmed the gregarine experts' predictions: they have started to document this tremendous diversity at molecular level (de Vargas et al., 2015; Mahé et al., 2017; Del Campo et al., 2019). A very wide

diversity of apicomplexan-related sequences (~80% of which appears more closely related to gregarines) is present in terrestrial soils and marine sediments (Clopton et al., 1992; Mahé et al., 2017; Del Campo et al., 2019). This diversity may be paralleled to the biological cycles of gregarines for which cyst forms, enclosing each thousand to millions of sporocyst progenies, are frequently released with the feces of their infected hosts, easily contaminating therefore such soils. In addition, the high resistance of these sporocyst forms to environmental conditions certainly explains the high abundance and maintenance (in environments) of their enclosed DNA/RNA (Desportes and Schrével, 2013; Schrével and Desportes, 2015). However, and even more remarkably, the diversity of apicomplexan-related sequences with gregarine-like affinities is also high in marine environments (both pelagic and benthic), suggesting the presence of either developing or resistant (sporocyst) forms of these gregarines in association with planktonic elements – biological or even mineral – that remain to be identified or 'freely' floating in these marine environments (de Vargas et al., 2015). However, the flip side of these discoveries is that there is a remarkably high number of gregarine-like molecular data that cannot be related to formally described species (Del Campo et al., 2019). This is in part due to the very low molecular knowledge we currently have of taxonomically and morphologically described species. Only talking about the most commonly used molecular marker, SSU rDNA, it is available in databases for only about one hundred of the ~1770 formally described gregarine species, which corresponds to just ~5% of them, not taking into account the many gregarine species to be still discovered. So not only do we need to generate a much higher number of molecular markers for these known species, but we must also deploy appropriate strategies to morphologically and biologically describe the increasing number of 'molecular-species' the existence of which is clearly emerging from metagenomics studies, pointing to entirely novel phylogenetic groups within gregarines (de Vargas et al., 2015; Mahé et al., 2017; Del Campo et al., 2019).

Missing data blur apicomplexan phylogeny

The missing information regarding gregarines is not restricted to this lack of connection between taxonomic data derived from morphological studies and molecular data emerging from metagenomic approaches. The confrontation of these data indicates that our understanding of the apicomplexan biodiversity remains very limited, biased because we have too long neglected the large spectrum of their pathogenicity focusing preferentially on the deadliest ones and ignoring the vast majority of poorly pathogenic ones (Rueckert et al., 2019a). This biased position has probably blurred not only our comprehension of the extent of their extraordinary host-adaptive capacities, but altogether a full section of their evolutionary history. So that we still do not know where the emergence of these species is taking root; is it within a group of non-pathogenic symbionts that has become even more diversified than we imagine, advocating for multiple emergence within a radiation that is still incompletely understood (Janouskovec et al., 2019; Kwong et al., 2019; Mathur et al., 2019; Rueckert et al., 2019b)?

A close examination of the recent taxonomy of Apicomplexa clearly mentions that most groups are currently polyphyletic or paraphyletic, notably: Aconoidasida, Coccidia, Gregarinasina, Archigregarinorida and Eugregarinorida (Adl et al., 2018). This is mostly due to lack of sufficiently informative and resolving data concerning these organisms, still mostly based on SSU rRNA phylogenies (see (Cavalier-Smith, 2014; Simdyanov et al., 2017; Simdyanov et al., 2018) to give few examples) and recent, emerging, phylogenomic analyses (Janouskovec et al., 2019; Kwong et al., 2019; Mathur et al., 2019). Certainly, insufficient sampling also prevents a solid and integrated vision of phylogenetic interrelationships for all of these species (Del Campo et al., 2019; Janouskovec et al., 2019; Kwong et al., 2019; Mathur et al., 2019).

Conclusions

With just over 100 genomes deciphered for ~350 genera and ~6000 described species, Apicomplexa is a group for which there is still a lot to discover even if, it is not the poorest documented branch of the tree of life (Sibbald and Archibald, 2017). It is far from having revealed all the secrets of the diversity of its molecular innovations, developed during its evolutionary history as it had to adapt to such a wide diversity of hosts and environments (Desportes and Schrével, 2013; Schrével and Desportes, 2015).

As to date, molecular exploration of apicomplexan parasites have mainly concerned a very small number of species that have in common (1) to infect humans causing threatening and global diseases such as malaria or less threatening but worldwide spread diseases such as toxoplasmosis and cryptosporidiosis; (2) to be cultivable in the laboratory, at least for some developmental stages; (3) to have been the subject of extremely sophisticated methodological developments such as genetic manipulation, –omics in all their variations and static and dynamic microscopy.

In this panorama, gregarines, full members of the Apicomplexa phylum, have been so far left on the side of the road for exactly corollary reasons: (1) they do not infect humans, (2) they are not easy to cultivate and (3) while there is a very abundant literature of their life cycles, morphologies and ultrastructure, they are almost unknown at the genomic/transcriptomic levels and have been the subject of very few biochemical studies. But their future is now open for exploration.

The terrain is marked out for the emergence of genomic data. A first move should be in favour of known species, selected either for their biological characteristics (intestinal, coelomic, motile, non-motile) or their particular phylogenetic position (archigregarines vs eugregarines as a broad distinction). They have the potential to teach us novel adaptive aspects of the group (relative to intracellular parasites of vertebrates), particularly in relation to their extracellular mode of living and the associated specific constraints to survival capacities. Such exploration should also reveal a part of the ancestral heritage, now possibly lost in the other branches of Apicomplexa. Studying gregarines can also allow understanding finer adaptive mechanisms, for example regarding the functionalisation of the apical complex, which obviously, compared to intracellular parasites of vertebrates displays common properties (recognition) with additional functions (nutrition) and missing ones (full invasion), the molecular bases of which are to be discovered. But it cannot be overlooked that the currently documented diversity of gregarines is lagging, by probably several orders of magnitude, far beyond the true diversity of these organisms in open and host-associated environments (Desportes and Schrével, 2013; Schrével and Desportes, 2015; Mahé et al., 2017; Del Campo et al., 2019; Mathur et al., 2019). This diversity is also likely, once it is explored and named (in terms of species and lifestyle) to reveal even more original and unknown adaptive mechanisms for this fascinating group, the Apicomplexa.

Author contribution

I.F. conducted the study. J.B. elaborated all illustrations. I.F. and J.B. wrote the manuscript.

Funding

This work was supported by a grant from Agence Nationale de la Recherche (LabEx ANR-10-LABX-0003-BCDiv), in the program 'Investissements d'avenir' (ANR-11-IDEX-0004-02), as well as several interdisciplinary Programs of the MNHN (ATM-Microorganismes, ATM-Génomique et Collections, ATM-Emergence, AVIV department), and by CNRS (PhD fellowship to Julie Boisard, 2018–2021).

Acknowledgements

We deeply thank Professor J. Schrével, Emeritus Professor, UMR7245 CNRS MNHN for careful reading and editing of text and figures as well as valuable suggestions.

Conflict of interest statement

The authors have declared no conflict of interest.

References

Adl, S.M., Bass, D., Lane, C.E., Lukes, J., Schoch, C.L., Smirnov, A., Agatha, S., Berney, C., Brown, M.W., Burki, F., Cardenas, P., Cepicka, I., Chistyakova, L., Del Campo, J., Dunthorn, M., Edvardsen, B., Eglit, Y., Guillou, L., Hampl, V., Heiss, A.A., Hoppenrath, M., James, T.Y., Karpov, S., Kim, E., Kolisko, M., Kudryavtsev, A., Lahr, D.J.G., Lara, E., Le Gall, L., Lynn, D.H., Mann, D.G., Massana, I.M.R., Mitchell, E.A.D., Morrow, C., Park, J.S., Pawlowski, J.W., Powell, M.J., Richter, D.J., Rueckert, S., Shadwick, L., Shimano, S., Spiegel, F.W., Torruella, I.C.G., Youssef, N., Zlatogursky, V. and Zhang, Q. (2018) Revisions to the classification, nomenclature, and diversity of eukaryotes. J. Eukaryot. Microbiol. **66**, 4–119

Aly, A.S., Vaughan, A.M. and Kappe, S.H. (2009) Malaria parasite development in the mosquito and infection of the mammalian host. Annu. Rev. Microbiol. **63**, 195–221

Aurrecoechea, C., Barreto, A., Basenko, E.Y., Brestelli, J., Brunk, B.P., Cade, S., Crouch, K., Doherty, R., Falke, D., Fischer, S., Gajria, B., Harb, O.S., Heiges, M., Hertz-Fowler, C., Hu, S., Iodice, J., Kissinger, J.C., Lawrence, C., Li, W., Pinney, D.F., Pulman, J.A., Roos, D.S., Shanmugasundram, A., Silva-Franco, F., Steinbiss, S., Stoeckert C.J., Jr., Spruill, D., Wang, H., Warrenfeltz, S. and Zheng, J. (2017) EuPathDB: the eukaryotic pathogen genomics database resource. Nucleic Acids Res. **45**, D581–D591

Bahrndorff, S., Alemu, T., Alemneh, T. and Lund Nielsen, J. (2016) The microbiome of animals: implications for conservation biology. Int. J. Genomics **2016**, 5304028

Barta, J.R. and Thompson, R.C. (2006) What is Cryptosporidium? Reappraising its biology and phylogenetic affinities. Trends Parasitol. **22**, 463–468

Bartosova-Sojkova, P., Oppenheim, R.D., Soldati-Favre, D. and Lukes, J. (2015) Epicellular apicomplexans: parasites "on the way in". PLoS Pathog. **11**, e1005080

Boucher, L.E. and Bosch, J. (2015) The apicomplexan glideosome and adhesins - structures and function. J. Struct. Biol. **190**, 93–114

Bouzid, M., Hunter, P.R., Chalmers, R.M. and Tyler, K.M. (2013) Cryptosporidium pathogenicity and virulence. Clin. Microbiol. Rev. **26**, 115–134

Buchmann, K. (2014) Evolution of innate immunity: clues from invertebrates via fish to mammals. Front. Immunol. **5**, 459

Canning, E.U. (1956) A new eugregarine of locusts, Gregarina garnhami n.sp., parasitic in Schistocerca gregaria Forsk. J. Protozool. **3**, 50–62

Cavalier-Smith, T. (2014) Gregarine site-heterogeneous 18S rDNA trees, revision of gregarine higher classification, and the evolutionary diversification of Sporozoa. Eur. J. Protistol. **50**, 472–495

Clopton, R.E. (2009) Phylogenetic relationships, evolution, and systematic revision of the septate gregarines (Apicomplexa:Eugregarinorida:Septatorina). Comp. Parasitol. **76**, 167–190

Clopton, R.E., Janovy J., Jr. and Percival, T.J. (1992) Host stadium specificity in the gregarine assemblage parasitizing Tenebrio molitor. J. Parasitol. **78**, 334–337

Counihan, N.A., Kalanon, M., Coppel, R.L. and de Koning-Ward, T.F. (2013) Plasmodium rhoptry proteins: why order is important. Trends Parasitol. **29**, 228–236

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., Lara, E., Berney, C., Le Bescot, N., Probert, I., Carmichael, M., Poulain, J., Romac, S., Colin, S., Aury, J.M., Bittner, L., Chaffron, S., Dunthorn, M., Engelen, S., Flegontova, O., Guidi, L., Horak, A., Jaillon, O., Lima-Mendez, G., Lukes, J., Malviya, S., Morard, R., Mulot, M., Scalco, E., Siano, R., Vincent, F., Zingone, A., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Acinas, S.G., Bork, P., Bowler, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Not, F., Ogata, H., Pesant, S., Raes, J., Sieracki, M.E., Speich, S., Stemmann, L., Sunagawa, S., Weissenbach, J., Wincker, P. and Karsenti, E. (2015) Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. Science **348**, 1261605

Del Campo, J., Heger, T.J., Rodriguez-Martinez, R., Worden, A.Z., Richards, T.A., Massana, R. and Keeling, P.J. (2019) Assessing the diversity and distribution of apicomplexans in host and free-living environments using high-throughput amplicon data and a phylogenetically informed reference framework. Front. Microbiol. **10**, 2373

Desportes, I. and Schrével, J. (2013) Treatise on zoology—Anatomy, taxonomy, biology. The Gregarines: The early branching Apicomplexa (Volumes 1 and 2). Leiden, The Netherlands; Boston, MA: Brill. 791 pp

Diakin, A., Paskerova, G.G., Simdyanov, T.G., Aleoshin, V.V. and Valigurova, A. (2016) Morphology and molecular phylogeny of coelomic gregarines (Apicomplexa) with different types of motility: urospora ovalis and U. travisiae from the Polychaete Travisia forbesii. Protist **167**, 279–301

Diakin, A., Wakeman, K.C. and Valigurova, A. (2017) Description of Ganymedes yurii sp. n. (Ganymedidae), a new gregarine species from the antarctic amphipod gondogeneia sp. (Crustacea). J. Eukaryot. Microbiol. **64**, 56–66

Frenal, K., Dubremetz, J.F., Lebrun, M. and Soldati-Favre, D. (2017) Gliding motility powers invasion and egress in Apicomplexa. Nat. Rev. Microbiol. **15**, 645–660

Füssy, Z. and Oborník, M. (2017) Reductive evolution of apicomplexan parasites from phototrophic ancestors. In *Evolutionary Biology: Self/Nonself Evolution, Species and Complex Traits Evolution, Methods and Concepts*. P. Pontarotti, editor. Cham: Springer International Publishing. 217–236

Gawad, C., Koh, W. and Quake, S.R. (2016) Single-cell genome sequencing: current state of the science. Nat. Rev. Genet. **17**, 175–188

Gentekaki, E., Kolisko, M., Boscaro, V., Bright, K.J., Dini, F., Di Giuseppe, G., Gong, Y., Miceli, C., Modeo, L., Molestina, R.E., Petroni, G., Pucciarelli, S., Roger, A.J., Strom, S.L. and Lynn, D.H. (2014) Large-scale phylogenomic analysis reveals the phylogenetic position of the problematic taxon Protocruzia and unravels the deep phylogenetic affinities of the ciliate lineages. Mol. Phylogenet Evol. **78**, 36–42

Gubbels, M.J. and Duraisingh, M.T. (2012) Evolution of apicomplexan secretory organelles. Int. J. Parasitol. **42**, 1071–1081

Hakimi, M.A., Olias, P. and Sibley, L.D. (2017) Toxoplasma effectors targeting host signaling and transcription. Clin. Microbiol. Rev. **30**, 615–645

Janouskovec, 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

Janouskovec, J., Tikhonenkov, D.V., Burki, F., Howe, A.T., Kolisko, M., Mylnikov, A.P. and Keeling, P.J. (2015) Factors mediating plastid dependency and the origins of parasitism in apicomplexans and their close relatives. Proc. Natl. Acad. Sci. U.S.A. **112**, 10200–10207

Kuriyama, R., Besse, C., Geze, M., Omoto, C.K. and Schrével, J. (2005) Dynamic organization of microtubules and microtubule-organizing centers during the sexual phase of a parasitic protozoan, Lecudina tuzetae (Gregarine, Apicomplexa). Cell Motil. Cytoskeleton **62**, 195–209

Kwong, W.K., Del Campo, J., Mathur, V., Vermeij, M.J.A. and Keeling, P.J. (2019) A widespread coral-infecting apicomplexan with chlorophyll biosynthesis genes. Nature **568**, 103–107

Mahé, F., de Vargas, C., Bass, D., Czech, L., Stamatakis, A., Lara, E., Singer, D., Mayor, J., Bunge, J., Sernaker, S., Siemensmeyer, T., Trautmann, I., Romac, S., Berney, C., Kozlov, A., Mitchell, E.A.D., Seppey, C.V.W., Egge, E., Lentendu, G., Wirth, R., Trueba, G. and Dunthorn, M. (2017) Parasites dominate hyperdiverse soil protist communities in Neotropical rainforests. Nat. Ecol. Evol. **1**, 91

Mathur, V., Kolisko, M., Hehenberger, E., Irwin, N.A.T., Leander, B.S., Kristmundsson, A., Freeman, M.A. and Keeling, P.J. (2019) Multiple independent origins of apicomplexan-like parasites. Curr. Biol. **29**, 2936–2941 e2935

Morrissette, N.S. and Sibley, L.D. (2002) Cytoskeleton of apicomplexan parasites. Microbiol. Mol. Biol. Rev. **66**, 21–38; table of contents

Mueller, C., Graindorge, A. and Soldati-Favre, D. (2017) Functions of myosin motors tailored for parasitism. Curr. Opin. Microbiol. **40**, 113–122

Omoto, C.K., Toso, M., Tang, K. and Sibley, L.D. (2004) Expressed sequence tag (EST) analysis of Gregarine gametocyst development. Int. J. Parasitol. **34**, 1265–1271

Portman, N. and Slapeta, J. (2014) The flagellar contribution to the apical complex: a new tool for the eukaryotic Swiss Army knife? Trends Parasitol. **30**, 58–64

Prensier, G., Dubremetz, J.F. and Schrével, J. (2008) The unique adaptation of the life cycle of the coelomic gregarine Diplauxis hatti to its host Perinereis cultrifera (Annelida, Polychaeta): an experimental and ultrastructural study. J. Eukaryot. Microbiol. **55**, 541–553

Robert-Gangneux, F. and Darde, M.L. (2012) Epidemiology of and diagnostic strategies for toxoplasmosis. Clin. Microbiol. Rev. **25**, 264–296

Rueckert, S., Betts, E.L. and Tsaousis, A.D. (2019a) The symbiotic spectrum: where do the gregarines fit? Trends Parasitol. **35**, 687–694
Rueckert, S. and Horak, A. (2017) Archigregarines of the English Channel revisited: new molecular data on Selenidium species including early described and new species and the uncertainties of phylogenetic relationships. PLoS One **12**, e0187430
Rueckert, S., Pipaliya, S.V. and Dacks, J.B. (2019b) Evolution: parallel paths to parasitism in the Apicomplexa. Curr. Biol. **29**, R836–R839
Schrével, J. and Desportes, I. (2015) Gregarines. In *Encyclopedia of Parasitology*. H. Mehlhorn, editor. Berlin Heidelberg: Springer-Verlag
Schrével, J., Valigurova, 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
Sibbald, S.J. and Archibald, J.M. (2017) More protist genomes needed. Nat. Ecol. Evol. **1**, 145
Simdyanov, T.G., Guillou, L., Diakin, A.Y., Mikhailov, K.V., Schrével, J. and Aleoshin, V.V. (2017) A new view on the morphology and phylogeny of eugregarines suggested by the evidence from the gregarine Ancora sagittata (Leuckart, 1860) Labbe, 1899 (Apicomplexa: Eugregarinida). PeerJ. **5**, e3354
Simdyanov, T.G. and Kuvardina, O.N. (2007) Fine structure and putative feeding mechanism of the archigregarine Selenidium orientale (Apicomplexa: Gregarinomorpha). Eur. J. Protistol. **43**, 17–25
Simdyanov, T.G., Paskerova, G.G., Valigurova, A., Diakin, A., Kovacikova, M., Schrével, J., Guillou, L., Dobrovolskij, A.A. and Aleoshin, V.V. (2018) First ultrastructural and molecular phylogenetic evidence from the blastogregarines, an early branching lineage of plesiomorphic Apicomplexa. Protist **169**, 697–726
Tardieux, I. and Baum, J. (2016) Reassessing the mechanics of parasite motility and host-cell invasion. J. Cell Biol. **214**, 507–515
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. Mol. Biol. Evol. **27**, 235–248
Valigurova, A., Hofmannova, L., Koudela, B. and Vavra, J. (2007) An ultrastructural comparison of the attachment sites between Gregarina steini and Cryptosporidium muris. J. Eukaryot. Microbiol. **54**, 495–510
Valigurova, A. and Koudela, B. (2008) Morphological analysis of the cellular interactions between the eugregarine Gregarina garnhami (Apicomplexa) and the epithelium of its host, the desert locust Schistocerca gregaria. Eur. J. Protistol. **44**, 197–207
Valigurova, A., Vaskovicova, N., Musilova, N. and Schrével, J. (2013) The enigma of eugregarine epicytic folds: where gliding motility originates? Front. Zool. **10**, 57
Woo, Y.H., Ansari, H., Otto, T.D., Klinger, C.M., Kolisko, M., Michalek, J., Saxena, A., Shanmugam, D., Tayyrov, A., Veluchamy, A., Ali, S., Bernal, A., del Campo, J., Cihlar, J., Flegontov, P., Gornik, S.G., Hajduskova, E., Horak, A., Janouskovec, J., Katris, N.J., Mast, F.D., Miranda-Saavedra, D., Mourier, T., Naeem, R., Nair, M., Panigrahi, A.K., Rawlings, N.D., Padron-Regalado, E., Ramaprasad, A., Samad, N., Tomcala, A., Wilkes, J., Neafsey, D.E., Doerig, C., Bowler, C., Keeling, P.J., Roos, D.S., Dacks, J.B., Templeton, T.J., Waller, R.F., Lukes, J., Obornik, M. and Pain, A. (2015) Chromerid genomes reveal the evolutionary path from photosynthetic algae to obligate intracellular parasites. eLife **4**, e06974

Received: 10 January 2020; Revised: 3 March 2020; Accepted: 10 March 2020; Accepted article online: 16 March 2020

Parasite **28**, 12 (2021)

https://doi.org/10.1051/parasite/2021009

PARASITE
Available online at:
www.parasite-journal.org

RESEARCH ARTICLE OPEN ACCESS

Integrative taxonomy confirms that *Gregarina garnhami* and *G. acridiorum* (Apicomplexa, Gregarinidae), parasites of *Schistocerca gregaria* and *Locusta migratoria* (Insecta, Orthoptera), are distinct species

Isabelle Florent[1,*], Marie Pierre Chapuis[2,3], Amandine Labat[1], Julie Boisard[1,4], Nicolas Leménager[2,3], Bruno Michel[2,3], and Isabelle Desportes-Livage[1]

[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] CBGP, Univ Montpellier, CIRAD, INRAE, Institut Agro, IRD, 34060 Montpellier, France
[3] CIRAD, UMR CBGP, 34398 Montpellier, France
[4] 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

Received 28 July 2020, Accepted 2 February 2021, Published online 23 February 2021

Abstract – Orthoptera are infected by about 60 species of gregarines assigned to the genus *Gregarina* Dufour, 1828. Among these species, *Gregarina garnhami* Canning, 1956 from *Schistocerca gregaria* (Forsskål, 1775) was considered by Lipa et al. in 1996 to be synonymous with *Gregarina acridiorum* (Léger 1893), a parasite of several orthopteran species including *Locusta migratoria* (Linné, 1758). Here, a morphological study and molecular analyses of the SSU rDNA marker demonstrate that specimens of *S. gregaria* and specimens of *L. migratoria* are infected by two distinct *Gregarina* species, *G. garnhami* and *G. acridiorum,* respectively. Validation of the species confirms that molecular analyses provide useful taxonomical information. Phenotypic plasticity was clearly observed in the case of *G. garnhami*: the morphology of its trophozoites, gamonts and syzygies varied according to the geographical location of *S. gregaria* and the subspecies infected.

Key words: Gregarines, Orthoptera, Species delimitation, SSU rDNA phylogeny, Phenotypic plasticity, Biodiversity.

Résumé – La taxonomie intégrative confirme que *Gregarina garnhami* et *G. acridiorum* (Apicomplexa, Gregarinidae), parasites de *Schistocerca gregaria* et *Locusta migratoria* (Insecta, Orthoptera), sont des espèces distinctes. Les orthoptères sont parasités par environ soixante espèces de grégarines affiliées au genre *Gregarina* Dufour, 1828. Parmi ces espèces *Gregarina garnhami* Canning, 1956 décrite chez *Schistocerca gregaria* (Forskål, 1775), a été mise en synonymie par Lipa et al. en 1996 avec *Gregarina acridiorum* (Léger 1893), parasite de plusieurs espèces d'orthoptères dont *Locusta migratoria* (Linné, 1758). Ici, une étude morphologique et des analyses moléculaires du marqueur SSU rDNA démontrent que les spécimens de *S. gregaria* et ceux de *L. migratoria* sont infectés par 2 espèces distinctes de grégarines, *Gregarina garnhami* et *Gregarina acridiorum*, respectivement. La validation de ces espèces confirme l'importance des informations fournies par les analyses moléculaires dans les études taxonomiques. Une plasticité phénotypique a été clairement observée dans le cas de *G. garnhami :* la morphologie de ses trophozoïtes, gamontes et syzygies varie selon la localisation géographique et la sous-espèce de *S. gregaria* infectée.

Introduction

Gregarines are a heterogeneous group of apicomplexan parasites that infect a very wide range of non-vertebrate hosts, in which they mostly occupy intestinal tracts and coelomic spaces [17]. The biodiversity of gregarines currently corresponds to 1600-1700 formally described species [32], but according to experts in the field, this number may be vastly underestimated [1, 17]. Recent metagenomic surveys of terrestrial soils and marine environments further confirmed the high occurrence and abundance of gregarine-like sequences in these environments that remain to be ascribed to formally described species [15, 16, 28]. In the past, ascribing gregarine species assignations was based on combinations of morphological and behavioral

*Corresponding author: isabelle.florent@mnhn.fr

features including parasitic life traits (host and host range specificities), the different locations occupied by the parasite in hosts (i.e. intestine or coelom), descriptions of life-cycle development stages (morphological measurements, duration of the stages, scanning and transmission electron microscopy), gamont pairing (frontal, lateral, caudo-frontal), and modes of gametocyst dehiscence [11, 17, 26]. The increasing use of molecular data in recent decades has led to the confirmation, but also sometimes to the revision of the taxonomic and phylogenetic view we have of gregarines, and has revealed that some species that were once considered distinct are in fact the same [19] or, the reverse, novel cryptic species, i.e. morphologically indistinguishable but clearly distinct at the molecular level [30].

Orthoptera (Ensifera (crickets and katydids) and Caelifera (grasshoppers, ground-hoppers and pygmy mole crickets)) are reported to be parasitized by about 60 species assigned to the genus *Gregarina* Dufour, 1828 (see [17] for a recent, extensive review of the literature). Based on morphological descriptions, some gregarine species have been found to be restricted to one host family or superfamily, while others seem to have the capacity to infect a wide range of hosts distributed worldwide [14, 17, 36, 37]. Problems of identification based on morphological characters likely arose from phenotypic plasticity in response to wide-range host species and/or other contrasted environmental conditions. As a result, species delimitation within the genus *Gregarina* has been the subject of debate, with confusion, descriptions and synonymies, in particular for gregarines that infect the Caelifera suborder, as illustrated below. Species delimitation is, however, a global and recurrent issue in protistology [6].

In 1893, Léger described *Clepsidrina acridiorum* [24], which, a few years later, was termed *Gregarina* by Labbé (1899) [21]. This parasite was found in Acridoidea collected in Algeria [24]. As the infected specimens belonged to different genera of Caelifera (*Truxalis*, *Pamphagus*, *Sphingonotus*), Léger concluded that "other acridians from Africa should be investigated for potential *G. acridiorum* infections" [24]. Interestingly, he noticed that *G. acridiorum* was not found in the desert locust, *Schistocerca gregaria* [24]. Later in 1956, Canning described a gregarine she named *Gregarina garnhami*, sampled from this *S. gregaria* host [7]. Interestingly, *G. garnhami* was also reported by the same author in both the migratory locust, *L. migratoria* and in the Egyptian locust, *Anacridium aegyptium* [7]. According to data in the literature, *G. acridiorum* and *G. garnhami* share common morphological and behavioral characteristics, such as their development in the midgut of their hosts, a small globular epimerite, stout bodied gamonts, and barrel-shaped (or dolioform) oocysts [7, 23, 27]. In 1996, Lipa et al. concluded that the species described in 1956 by Canning in *S. gregaria* was in fact *G. acridiorum* [27]. This interpretation was supported by the fact that in 1956, Canning had not been aware of the existence of *G. acridiorum* [27]. *Gregarina acridiorum* has been reported in a range of Orthoptera hosts (Ensifera and Caelifera: Acrididae, Tetrigidae) including *L. migratoria* and *A. aegyptium* [12, 27], two species also described as hosts of *G. garnhami* [7]. Consequently, the two acridian species could be infected by the two gregarines species.

Gregarina acridiorum and *G. garnhami* also closely resemble *Gregarina rigida* (Hall, 1907) Ellis, 1913, described in a broad range of widely distributed orthopteran hosts [17] and also similar to *Gregarina ronderosi*, a parasite of the argentine grasshopper *Dichroplus elongatus* [22]. The developmental and morphological similarities of these four gregarines question their species definition as well as their host specificities and indeed, based on these similarities, in 1968, Corbel even proposed that *G. rigida* and *G. acridiorum* were the same [13]. Table 1 lists the main biological and morphological characters of these four very similar gregarines of acridians, plus data concerning a fifth species, *Gregarina caledia* (nomen nudum), a parasite of the Australian grasshopper *Caledia captiva*, described in the PhD Thesis of Jennifer Ann Ninham (1995) and reported to be very similar to *G. garnhami* [30]. Table 1 illustrates how tenuous some of these differences can be when these five gregarines of acridians are compared. The limited availability of DNA sequences corresponding to these species is an obstacle to the resolution of these controversies (only partial SSU rDNA sequences (1210 bp) are available for *G. caledia* (L31799) and *Gregarina chortiocetes* (L31841)). The latter species, an intestinal parasite of *Chortiocetes terminifera,* is however poorly described at the morphological level [30].

In 2002, Lange & Wittenstein indicated that: "given the great similarity of *Gregarina* spp. associated with Acrididae, it would probably be very informative to study, at the molecular level, as many species as possible" [22]. To achieve this objective, we combined morphological and molecular data to better explore the species boundaries of gregarines that infect two orthopteran Acrididae hosts, *S. gregaria* (Forsskål, 1775) and *L. migratoria* (Linné, 1758). These two hosts are locusts, i.e. grasshoppers that can form dense migrating swarms, that are often destructive to agriculture, through an extreme form of density-dependent phenotypic plasticity, known as phase polyphenism [3, 41]. Here we sought to determine whether they are infected by the same or distinct gregarine species, as the information in the current literature is not congruent [7, 12–14, 21, 24, 27].

Morphological observations of the developmental stages of gregarines from *L. migratoria* and two subspecies of *S. gregaria* were performed and completed with the sequencing of their SSU rDNA loci. The results revealed clear molecular differences in this genetic marker, despite extremely similar morphological features, strongly supporting the hypothesis that these two acridian hosts are not infected by the same gregarine species. Some subtle morphological differences have also been identified between the two gregarine species.

Materials and methods

Collection of hosts and isolation of parasites

Specimens of *L. migratoria* (Linné, 1758) were obtained from the vivarium belonging to the French National Museum of Natural History (French acronym MNHN) (Source uncertain; time of establishment > 15 years, regularly replenished from Insect Raising SARL (2, Chemin Champthiaud,

25410 Dannemarie-sur-Crète, France). Two sub-species of the desert locust, *S. gregaria,* were studied: *S. gregaria gregaria* (Forsskål, 1775) and *S. gregaria flaviventris* (Burmeister, 1838), isolated in distinct regions along a north–south axis in Africa [10, 41]. The *S. g. gregaria* insects came from either a long-standing laboratory strain belonging to the team involved in molecular developmental physiology and signal transduction of the Department of Biology of Leuven University, Belgium (https://bio.kuleuven.be/df/jv; geographical source: Mauritania; time since establishment: several decades) or a long-standing laboratory strain acquired from the National Anti-Locust Centre in Agadir, Morocco, regularly replenished with individuals sampled in the field (geographical source: between Draa wadi and the Dakhla region; time of establishment: from the 1990s to 2014). *Schistocerca gregaria gregaria* egg pods from the two strains were received at the SEPA platform in CBGP on May 30 and June 12, 2014, respectively, and hatchlings were crowd-reared before treatment (July 17 and 23, 2014) in a breeding chamber at 32 °C, with 50% humidity, with a 12 h:12 h photoperiod, and fed with seedling wheat, supplemented by wheat bran for adults. The *S. g. flaviventris* insects came from a natural population in Tankwa Karoo National Park, South Africa (20.03° E; −32.23°), in which 70 nymphs were collected on February 23, 2014 and taken to the SEPA platform in CBGP for two generations of maintenance before treatment on July 11 and July 18, 2014, in the same breeding chamber and under the same conditions.

The hosts used in this study and the dates of sampling for biological analyses are listed in Table 2. All acridian specimens were anesthetized with chloroform. Their digestive tract was dissected in 0.22 µm-filtered sterile PBS 1X and gamonts (solitary or in syzygies) and gametocysts were isolated from ceca and midguts (*S. gregaria*) or only midguts (*L. migratoria*) using tweezers and sterile elongated Pasteur pipettes, under a stereomicroscope. Gametocysts were also occasionally isolated from insect feces and kept at room temperature to observe dehiscence. All isolated gregarines were washed at least three times in 0.22 µm-filtered sterile PBS 1X to eliminate host tissue and environmental bacteria prior to being used for microscopic observations, fixed for scanning electron microscopy, or stored as cell pellets at −20 °C prior to genomic DNA extraction.

Morphological studies

Isolated parasites were first observed on slides using light microscopy. Images were acquired using a Nikon DXM 1200C camera and a micrometric slide to set the scales, and the images were processed using ImageJ software (https://imagej.nih.gov/ij/). In parallel, pools of isolated and washed gamonts and gametocysts and relevant sections of infected acridian ceca and midguts were prepared for scanning electron microscopy (SEM). After appropriate washing in 0.22 µm-filtered sterile PBS 1X, the samples were fixed in 5% (v/v) glutaraldehyde in 0.2M cacodylate buffer (pH 7.2) at 4 °C for 6–12 h then washed twice in 0.2M cacodylate buffer (pH 7.2) before undergoing successive series of dehydration in 50, 70, 90 and 100% ethanol. Samples were critical point-dried in liquid CO_2 (Emitech K850, Quorum Technologies, Lewes, United Kingdom) then coated with 20 nm gold (JFC-1200 Fine coater, JEOL, Tokyo, Japan). Samples were then examined with a Hitachi Scanning Electron Microscope SU3500 Premium (Hitachi, Tokyo, Japan), as previously described [2]. Quantitative measurements were length and width at the different life stages, including length of protomerites and deutomerites for trophozoites and gamonts.

Statistical tests

In order to compare the averages of the measurements carried out for the gregarines infecting either *S. gregaria* or *L. migratoria* hosts, statistical tests were performed as follows. For the group of measurements with $n = 18$, we used a Shapiro–Wilk Test to assess the normality of the data, which established normality. For this sample and all the other groups of measurements tested with $n > 30$, we used parametric tests. First, a Fisher test was conducted to test the homoscedasticity of the variances within the groups. When homoscedasticity was retrieved, we conducted a Student's *t*-test to compare the means of each group. When homoscedasticity was not retrieved, we then used a Welch's *t*-test. Analyses were performed using R software.

DNA extraction and sequencing

Total genomic DNA was extracted from pools of parasites (gamonts or gametocysts), isolated from individual host specimens as indicated in Table 3, using standard phenol-chloroform extractions [34] or MasterPure™ Complete DNA and RNA Purification kits (Epicentre Biotechnologies, Madison, WI, USA), as previously described [35]. Isolated nucleic acids were subsequently used as templates in standard PCR reactions designed to amplify most of the SSU rDNA loci (V1–V8) [18], using forward WL1 – 5′–GCGCTACCTGGTTGATCCTGCC–3′ and reverse EukP3 5′–GACGGGCGGTGTGTAC–3′ primers, as previously described [35]. After confirmation of the appropriate amplicon size by agarose-gel electrophoresis, PCR products were purified using an Illustra™ GFX™ PCR DNA and Gel Band Purification Kit (GE Healthcare, France), and cloned into a pGEM®-T Easy Vector (Promega, Madison WI, USA), as previously described [35, 39]. DNA sequences were obtained by Sanger technology (Beckman Coulter Genomics, Takeley, United Kingdom) from positive clones selected by PCR using the T7 and Sp6 universal primers that flank the pGEM®-T Easy Vector cloning site, as previously described [39]. In addition to using T7 and Sp6 as sequencing primers, several internal primers were used (LWA3 5′–AAACTTAAAGGAATTGACGG–3′; PIF4F 5′–CCGTTACTTTGAGCAAATTGG–3′; PIF4R 5′–CTTAGAATTTCACCTCTCTCC–3′). SSU rDNA loci were then aligned and assembled from raw data using MEGA X [20]. The 43 novel sequences were deposited in the European Nucleotide Archive (ENA) database under accession numbers: LR814064–LR814106 (http://www.ebi.ac.uk/ena/data/view/LR814064-LR814106).

Phylogenetic analyses

Using maximum likelihood (ML) and Bayesian methods, phylogenetic trees were built from 69 sequences from

Table 1. Morphological differences between five very similar gregarines of acridians reported in the literature. This table is based on individual descriptions provided by the authors of [7, 21, 22, 24, 30]; see also [17]. *D*, diameter; *L*, length; *W*, width; TL, total length.

Gregarine	*Gregarina acridiorum* (Léger, 1893) Labbé, 1899 [24] [21]	*Gregarina garnhami* Canning, 1956 [7]	*Gregarina rigida* (Hall, 1907) Ellis, 1913 [17]	*Gregarina ronderosi,* Lange & Wittenstein, 2002 [22]	*Gregarina caledia,* Ninham, 1995 [30]
Hosts	Caelifera: Acrididae, Tetrigidae; Ensifera: Tettigoniidae	Caelifera: Acrididae	Caelifera: Acrididae; Ensifera: Tettigoniidae	Caelifera: Acrididae	Caelifera: Acrididae
Infected sites in hosts	Midgut	Early stages in gastric ceca and occasionally in the midgut; gamonts in the midgut	Early stages in gastric ceca, near the anterior end of the midgut	Trophozoites, solitary or associated gamonts in gastric ceca and gut; gametocysts in the hindgut	Trophozoites, solitary or associated gamonts in gastric ceca and midgut; gametocysts in the hindgut
Trophozoites – gamonts					
Gamonts	Gamonts: cylindrical, ovoid in older forms, endocyte yellow orange. *L*: ~ 400 µm, *W*: 160 µm	Gamonts: rather stout bodied in older forms, endocytes are pale yellow. *L*: 250–554 µm	Gamonts: rather stout bodied, endocytes are brownish orange. *L*: 250–750 µm *W*: 130–210 µm	Trophozoite (epimerite): *L*: 10.4–275 µm, more slender than gamonts; Gamonts: rather stout bodied, endocytes are pale yellow. *L*: 80–348 µm	Gamonts: pale-yellow, ovoid then cylindrical *L*: 180–264 µm *W*: 60–70 µm Mean: 222 µm × 65 µm
Association Length	TL: up to 1000 µm	TL: 500–1110 µm	TL: up to 1425 µm (average: 550 µm). Protomerite smaller in the satellite than in the primite	TL: 160–700 µm (average: 425 µm). Primites and satellites are similar in size and shape	TL: 515 µm. Primites and satellites are similar in size. Also seen: primite with 2 small satellites
Epimerite	Small, spherical with a short stalk.	Small, globular with a short stalk.	Small spherical hyaline knob.	Conical when attached, globular and smaller in free trophozoites	Globular
Protomerite	Sub-globular in primites, depressed at the anterior end in satellites.	Conical in young stages, subspherical in older stages, widest at the septum, tapering towards the anterior end, flattened in satellite.	Somewhat flattened, 3 times wider than long, generally less constriction at septum more or less indistinct.	Sub-globular in primites, depressed at anterior end in satellites, less flattened than in *G. garnhami*	Rounded anterior end
Deutomerite	Cylindrical, rounded posterior end	Cylindrical or rounded, with sharply pointed posterior end	Cylindrical or barrel-shaped, little wider than protomerite, broadly rounded end or flattened "cornered" extremity	Cylindrical, rounded posterior end, wider than protomerite, barrel-shaped in older forms	Cylindrical, in small gamonts, wider than in protomorites and rounded in older forms
Gametocysts – oocysts					
Gametocysts	*D*: 500 µm, thick ectocyst	*D*: 114–470 µm (exclusive of the ectocyst)	*D*: 300 µm in average. Yellow orange color. Thick ectocyst	*D*: 96–376 µm. Thick ectocyst	*D*: 228–312 µm (mean 270 µm). Yellow orange color. Ectocyst (24–100 µm thick)

(Continued on next page)

Table 1. (Continued)

Gregarine	*Gregarina acridiorum* (Léger, 1893) Labbé, 1899 [24] [21]	*Gregarina garnhami* Canning, 1956 [7]	*Gregarina rigida* (Hall, 1907) Ellis, 1913 [17]	*Gregarina ronderosi,* Lange & Wittenstein, 2002 [22]	*Gregarina caledia,* Ninham, 1995 [30]
Basal discs	Yellow orange	Yellow orange	Not mentioned	Orange	Orange
Sporoducts	12–15, with a swollen basal part, *L* > 1/2 cyst diameter	8, *L*: 1/3 cyst diameter (without ectocyst)	10 or more, short	12–15, *L*: up to 60 µm	5 to more than 10
Oocysts (sporocysts)	Dolioform*, double wall 7.6 µm × 3.3 µm	Dolioform*, thick wall 6.5–7 µm × 4 µm	Barrel-shaped* 8 µm × 5 µm	Dolioform* or Barrel-shaped* 5 µm × 3.2 µm	Barrel-shaped* 12 µm × 6 µm

* Depending on the authors, the terms "dolioform" and/or "barrel-shaped" were used to describe the shape of oocysts. Note also that oocysts were called sporocysts in all these historical descriptions.

Table 2. Acrididae hosts used in this study, sampling dates, host status and sampled gregarines. "Sick" hosts died rapidly (within days) in laboratory conditions in contrast to "healthy" hosts that were maintained for weeks.

Acrididae host/designation in study	Source	Host status	Gregarines sampled
Schistocerca gregaria gregaria (2014)/**SG-M**	Long-standing laboratory strain from CNLA Agadir, Morocco	Sick	Young trophozoites in ceca, gamonts, syzygies and gametocysts in the midgut, occasionally gametocysts in feces; high infection level
Schistocerca gregaria gregaria (2014)/**SG-B**	Long-standing laboratory strain from KU Leuven, Belgium	Healthy	Young trophozoites in ceca, gamonts, syzygies and gametocysts in the midgut, occasionally gametocysts in feces; high infection level
Schistocerca gregaria flaviventris (2014)/**SG-SA**	Natural population from Tankwa Karoo National Park, South Africa	Sick	Young trophozoites in ceca, gamonts, syzygies and gametocysts in the midgut, occasionally gametocysts in feces; high infection level
Locusta migratoria (2012, 2014, 2015)/**LM-M**	Long-standing laboratory strain from MNHN Vivarium, France	Healthy	Gamonts, syzygies and gametocysts in the midgut, occasionally gametocysts in feces; mild infection level

gregarines infecting either *S. gregaria* (20 sequences), *L. migratoria* (23 sequences), a range of different insect hosts (22 sequences) or marine crustaceans, chosen as the gregarine outgroup specimen (4 sequences) [11, 30, 35]. Using a previously published alignment [35], the new gregarine sequences were added manually to yield a confident alignment of 1433 positions, after selection of conserved blocks defined using Gblocks 0.91b [8] (parameters used: Minimum Number Of Sequences For A Conserved Position: 35; Minimum Number Of Sequences For A Flanking Position: 58; Maximum Number Of Contiguous Nonconserved Positions: 8; Minimum Length Of A Block: 3; Allowed Gap Positions: With Half Use Similarity Matrices: Yes). A GTR substitution model with gamma-distributed rate variation across sites and a proportion of invariant sites was suggested as the best-fit model by MEGA X [20]. A Bayesian phylogenetic tree was constructed with MrBayes v3.2.3 [33] using lset *n* = 6 rates = invgamma parameters; Monte Carlo Markov Chain parameters were mcmc ngen = 100 00 000 relburnin = yes burninfrac = 0.25 samplefreq = 1000 printfreq = 10 000 nchains = 4 nruns = 2. A consensus tree was constructed from the post burn-in trees and posterior probabilities were calculated in MrBayes. Posterior probabilities > 0.95 were considered strong support. Maximum likelihood analyses were performed using RAxML version 8.2.12 [40] using the GTR + G + I model; bootstraps were estimated from 1000 replicates. Bootstrap percentages > 75% were considered good support. Trees were visualized and edited with FigTree and Inkscape.

Estimates of genetic divergence between and within groups

The numbers of base differences per site from averaging over all sequence pairs between and within each group were calculated using MEGA X [20]. This analysis involved 44 nucleotide sequences: 20 from gregarines that infect *S. gregaria,* 23 from gregarines that infect *L. migratoria*, and the sequence of *G. caledia* that infects *C. captiva* (L31799). For each sequence pair, all ambiguous positions were removed (pairwise deletion option) leaving a total of 1784 positions in the final dataset. From this dataset, we also constructed a minimum spanning network to analyze the relationships among the cloned SSU rDNA sequences using POPART [25].

Table 3. Gregarine specimens isolated for molecular investigation.

Host	Geographical origin and collection date	Number of isolated parasite stages	gDNA preparation (name, method)	Parasite clones (clone designations)
Locusta migratoria	MNHN 2012	Gamonts (50)	LW, Phenol chloroform	LM1.01.M.2012-1
Locusta migratoria	MNHN 2014	Gamonts (50)	JF, MasterPure	LM2.01.M.2014-2
Locusta migratoria	MNHN 2015	Gametocysts (20)	JS310, MasterPure	LM3.01.M.2015-3
				LM3.02.M.2015-4
				LM3.03.M.2015-5
				LM3.04.M.2015-6
				LM3.05.M.2015-7
Locusta migratoria	MNHN 2015	Gametocysts (17)	JS311, MasterPure	LM4.01.M.2015-8
				LM4.02.M.2015-9
				LM4.03.M.2015-10
				LM4.04.M.2015-11
Locusta migratoria	MNHN 2015	Gametocysts (13)	JS312, MasterPure	LM5.01.M.2015-12
				LM5.02.M.2015-13
				LM5.03.M.2015-14
				LM5.04.M.2015-15
Locusta migratoria	MNHN 2015	Gametocysts (13)	JS313, MasterPure	LM6.01.M.2015-16
				LM6.02. M.2015-17
				LM6.03. M.2015-18
				LM6.04. M.2015-19
				LM6.05. M.2015-20
Locusta migratoria	MNHN 2015	Gametocysts (17)	JS314, MasterPure	LM7.01. M.2015-21
				LM7.02. M.2015-22
				LM7.03. M.2015-23
Schistocerca gregaria flaviventris	South Africa 2014	Gamonts (10) and Gametocysts (10)	JS260, MasterPure	SG1.01.SA.2014-24
				SG1.02.SA.2014-25
				SG1.03.SA.2014-26
				SG1.04.SA.2014-27
Schistocerca gregaria flaviventris	South Africa 2014	Gametocysts (9)	JS261, MasterPure	SG2.01.SA.2014-28
				SG2.02.SA.2014-29
				SG2.03.SA.2014-30
				SG2.04.SA.2014-31
				SG2.05.SA.2014-32
Schistocerca gregaria flaviventris	South Africa 2014	Gamonts (~250)	JS269, MasterPure	SG3.01.SA.2014-33
				SG3.02.SA.2014-34
				SG3.03.SA.2014-35
Schistocerca gregaria gregaria	Belgium 2014	Gamonts (~200)	JS267, MasterPure	SG4.01.B.2014-36
				SG4.02.B.2014-37
				SG4.03.B.2014-38
				SG4.04.B.2014-39
				SG4.05.B.2014-40
				SG4.06.B.2014-41
				SG4.07.B.2014-42
Schistocerca gregaria gregaria	Morocco 2014	Young trophozoites in ceca (~400)	JS272, MasterPure	SG5.01.Ma.2014-43

Results

Gregarines isolated from the intestinal tracts of various acridian *S. gregaria* and *L. migratoria* host specimens (Table 2) were mostly located between the host intestine epithelial cells and digested food material. In addition, in all *S. gregaria* specimens, young trophozoite stages were invariably observed in the host's ceca, whereas this was never observed in *L. migratoria*. Occasionally, gametocysts were also isolated from insect feces and kept at room temperature to observe dehiscence. The observed stages were trophozoites, solitary gamonts, gamonts associated in caudo-frontal syzygies, and gametocysts enclosing oocysts or emitting them as chains through sporoducts (Fig. 1).

Morphological description of gregarines of *Schistocerca gregaria*

Young trophozoite stages (also referred to as cephalonts in historical publications [7, 17, 30]) (Fig. 1A) were observed in the two subspecies, regardless of the geographical location/raising facilities (Table 2). The globular epimerite with a short neck was visible in their anterior part (Fig. 1A). The density of infections could be very high, as shown by the number of trophozoites attached to the gut epithelium of an *S. g. gregaria* host from Morocco (Fig. 1B). The epimerite of attached trophozoites was enclosed in the host epithelial cell (Fig. 1C). High densities of trophozoites were also found in the ceca (data not shown) and midgut (solitary gamonts and syzygies

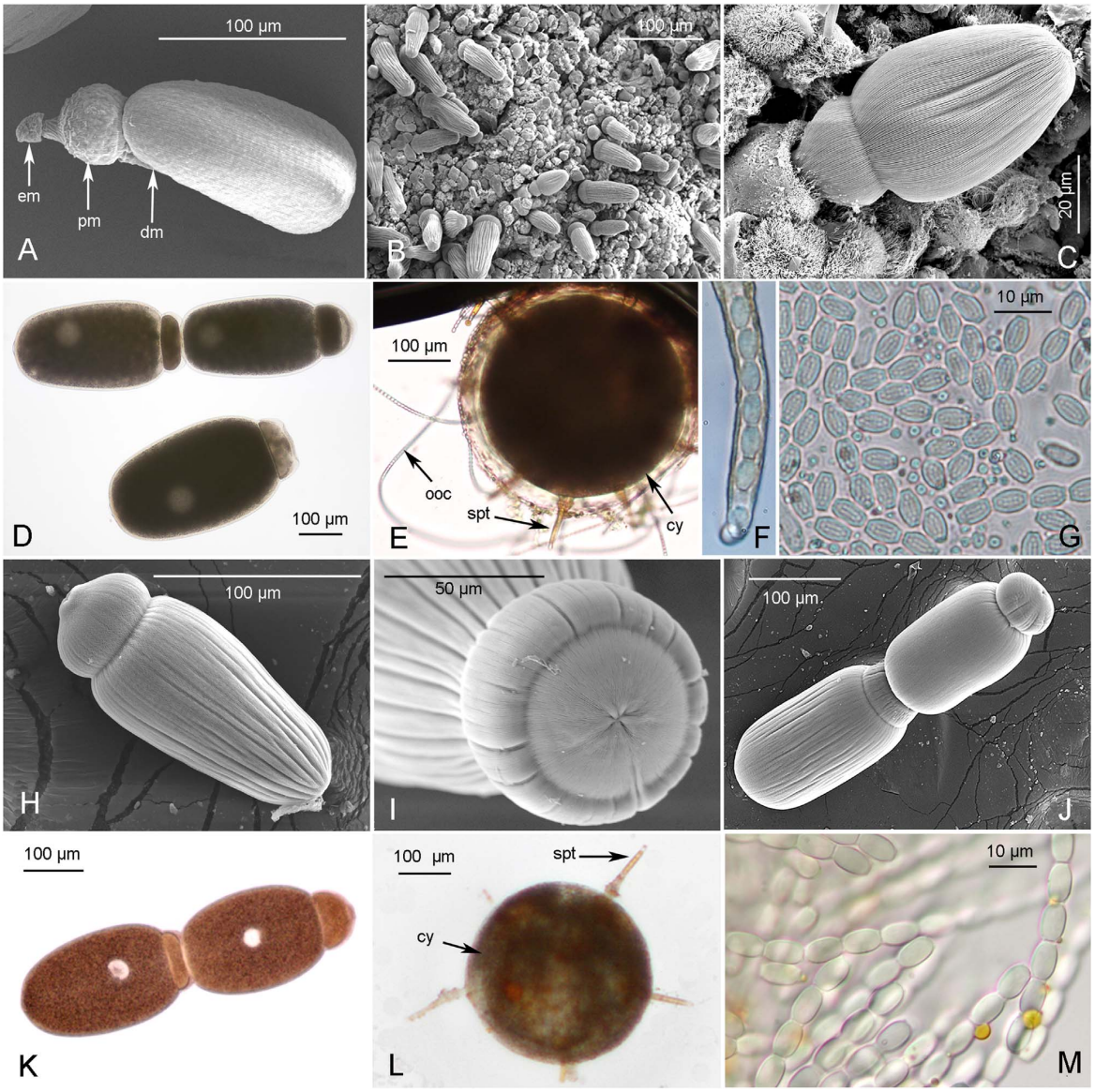


Figure 1. Scanning Electron Microscopy (A–C, H–J) and photonic imaging (D–G, K–M) of gregarines infecting *S. gregaria* (A–G) and *L. migratoria* (H–M). *S. gregaria* gregarines: A, young trophozoite (epimerite (em) protomerite (pm) and deutomerite (dm)), (South Africa); B, intestinal tract infected by numerous gregarines (Morocco); C, gregarine encased in an intestinal host cell, enlargement of B (Morocco); D. Solitary gamont and syzygy (Belgium); E. Gametocyst form (cy) with developed sporoducts (spt) releasing oocyst chains (ooc); F, zoom on sporoduct extremity showing enclosed oocysts; G. released oocysts. *L. migratoria* gregarines: H, solitary gamont detached from intestinal host cell; I. zoom on gamont protomerite; J–K, gamonts associated in syzygies; L, Gametocyst form (cy) with developed sporoducts (spt); M. released oocysts. Scales are given for each figure.

(Fig. 1D)). The protomerite of trophozoites and gamonts was oval or slightly conical (Figs. 1A–1D); in syzygies, it appeared to be flattened at the top of the satellite with a ridge formed during pairing with the primite (Fig. 1D). Scanning electron microscopy revealed a similar ridge at the top of the satellite in *G. garnhami* syzygies [42]. The deutomerite was cylindrical or ovoid, and quite stocky in older trophozoites and syzygies (Fig. 1D). A constriction of the septum was visible between the posterior part of the protomerite and the anterior part of the deutomerite (Fig. 1D). The nucleus was seen in the opaque endocyte of the deutomerite. Longitudinal epicytic folds were visible at the surface of these trophozoite/gamont stages (Figs. 1A–1C). Thickening of the ectocyte was visible above the endocyte at the apex of the primite protomerite (Fig. 1D).

The gamonts in *S. g. flaviventris* from South Africa (L (length) = 402 ± 79 µm, W (width) = 172 ± 42 µm, n = 27) were very similar in size to gamonts in *S. g. gregaria* from Belgium (L = 410 ± 53 µm, W = 200 ± 39 µm, n = 34),

but slightly smaller in *S. g. gregaria* from Morocco ($L = 332 \pm 43$ µm, $W = 96 \pm 16$ µm, $n = 4$). Moreover, smaller and much thinner trophozoites were observed in the latter ($L = 192 \pm 15$ µm, $W = 34 \pm 4$ µm, $n = 12$) (Fig. 1A). Also interestingly, gamonts in *S. g. gregaria* from Belgium were much stockier ($L/W = 2.1 \pm 0.2$) than gamonts in *S. g. flaviventris* from South Africa ($L/W = 2.4 \pm 0.3$) and gamonts ($L/W = 3.5 \pm 0.2$) and trophozoites ($L/W = 5.8 \pm 1.0$) in *S. g. gregaria* from Morocco. The ratios of protomerite (P) to deutomerite (D) lengths were, however, similar for gamonts, regardless of the infected hosts (P/D = 0.25 ± 0.04 (South Africa, $n = 27$); P/D = 0.23 ± 0.06 (Belgium, $n = 34$); P/D = 0.23 ± 0.07 (Morocco, $n = 4$), and also for the thinner trophozoites found in Moroccan *S. g. gregaria* specimens (P/D = 0.26 ± 0.04, $n = 12$). Overall, for trophozoites and gamonts infecting these hosts, regardless of the subspecies and their geographical location, the values were: $L = 370 \pm 98$ µm; $W = 159 \pm 69$; $L/W = 2.83 \pm 1.38$ ($n = 77$).

Gametocysts in dehiscence were observed, producing ~8 (but sometimes more) pale orange basal discs, circular cellular structures with a central opening that eventually developed across the mucilaginous layer (ectocyst) into sporoducts with swollen bases (Fig. 1E). Their length was ~1/3 that of the diameter of the gametocyst (Fig. 1E). Gametocysts diameters were 350 ± 56 µm, $n = 36$ (from 210 to 420 µm). Oocysts extruding as chains through these sporoducts (Fig. 1F) were barrel-shaped with a thick wall enclosing eight sporozoites (Fig. 1G). Their size was quite uniform ($L = 6.54 \pm 0.32$ µm, $W = 4.32 \pm 0.23$ µm, $n = 89$) (Fig. 1G).

Morphological description of gregarines of *Locusta migratoria*

Trophozoite stages attached to the gut epithelium of hosts were not seen, but a scar remained visible where the epimerite had been present at the top of the protomerite of detached gamonts (Figs. 1H–1I). These gamonts were rather cylindrical with a sub-globular protomerite, flattened and slightly constricted at the proto-deutomerite septum (Figs. 1H, 1J, 1K). The deutomerite was much longer and more slender towards the posterior end (Fig. 1H). The size of the gamonts varied but the mean size ($L = 219 \pm 48$ µm, $W = 93 \pm 30$ µm, $n = 37$) was smaller than the mean size observed in *S. gregaria* specimens (see above). Gamonts were also quite stocky ($L/W = 2.5 \pm 0.6$, $n = 37$). In caudo-frontal syzygies, the protomerite was sub-globular in the primite, but shorter and flattened with a circular anterior edge in the satellite (Figs. 1J–1K). The deutomerite was cylindrical, slightly wider in the anterior part (Fig. 1J), ovoid in syzygies (Fig. 1K), with a rounded posterior end. The spherical nucleus could be seen in the opaque cytoplasm (endocyte) of the deutomerite (Fig. 1K). Longitudinal epicytic folds were seen at the surface of these stages (Figs. 1H–1J). The length of these syzygies was ($L = 456 \pm 73$ µm, $W = 93 \pm 30$ µm, $n = 16$) in our studies. The ratio of protomerite (P) deuteromerite (D) lengths was ~1/4 (P/D = 0.25 ± 0.05, $n = 21$). Gametocysts were spherical with a mucilaginous layer (ectocyst). Under this layer, and as observed in gregarines that infect *S. gregaria*, basal discs of the future sporoducts differentiated at the surface of encysted gametocysts. These basal discs were also orange with a central white aperture, but were fewer in number (< 8, $n = 15$). Like in the case of gregarines that infect *S. gregaria*, chains of oocysts were extruded through sporoducts (Figs. 1L–1M) whose length in gregarines of *L. migratoria* is longer and represents ~1/2 the diameter of the gametocyst (Fig. 1L). Gametocysts diameters were 227 ± 35 µm, $n = 18$ (from 190 to 296 µm). Oocysts, that were also emitted as chains from sporoducts, were also barrel-shaped with a double wall but were slightly longer and slimmer ($L = 6.83 \pm 0.27$ µm, $W = 3.99 \pm 0.19$ µm, $n = 40$, Fig. 1M) than the oocysts emitted by gregarines that infect *S. gregaria* (Figs. 1G, 1F).

Statistical comparison of morphological measurements

For the gamonts, the means of the lengths (p-value = 2.2e-16; df (degree of freedom) = 111.97) and of the widths (p-value = 8.574e-11; $df = 111.13$) were significantly different between the gregarines infecting *S. gregaria* and *L. migratoria*. However, there were no significant differences between the length/width ratios between these two groups. Concerning the gametocysts diameters, the mean was significantly different (p-value = 1.986e-13; $df = 49.386$). Finally, for the oocysts, both mean length (p-value = 6.664e-07; $df = 89.407$) and mean width (p-value = 5.722e-13; $df = 88.967$) were significantly different.

SSU rDNA sequences

To further characterize these gregarines, a molecular study was designed to sequence most of the SSU rDNA locus from gamonts and gametocysts, isolated from several host specimens belonging to *L. migratoria* and two subspecies of *S. gregaria* (Table 2). A total of 23 sequences were generated from gregarines found in 7 specimens of *L. migratoria* on three collection dates, and 20 sequences were generated from gregarines found in five specimens of *S. gregaria* from a total of three geographical origins and/or raising facilities (Table 3). Regardless of the subspecies and the geographical location of hosts and their maintenance facilities, all the gregarines isolated from *S. gregaria* specimens shared the same "type 1" sequence (1638-bp long), presumably corresponding to *G. garnhami*, whereas all the gregarines isolated from *L. migratoria* specimens presented a clearly distinct "type 2" sequence (1637-bp long), presumably corresponding to *G. acridiorum*. Multiple sequence alignment and distance analyses were performed to qualify intra-species and inter-species variations, and clearly revealed two distinct clusters (Fig. 2A). Within the sequence group of gregarines from the host *S. gregaria*, the mean level of divergence was 0.2%, whereas within the sequence group of gregarines from the host *L. migratoria*, the mean level of divergence was 0.3%. The mean level of genetic distance between gregarine sequences from *S. gregaria* and those from *L. migratoria* was 1.5%, whereas the genetic divergence from *G. caledia,* parasite of *C. captiva*, was 1.1% with the gregarine group from *L. migratoria*, but 2.2% with the gregarine group from *S. gregaria*. In all, 22 conserved polymorphic positions, rather evenly distributed along the SSU rDNA locus, were

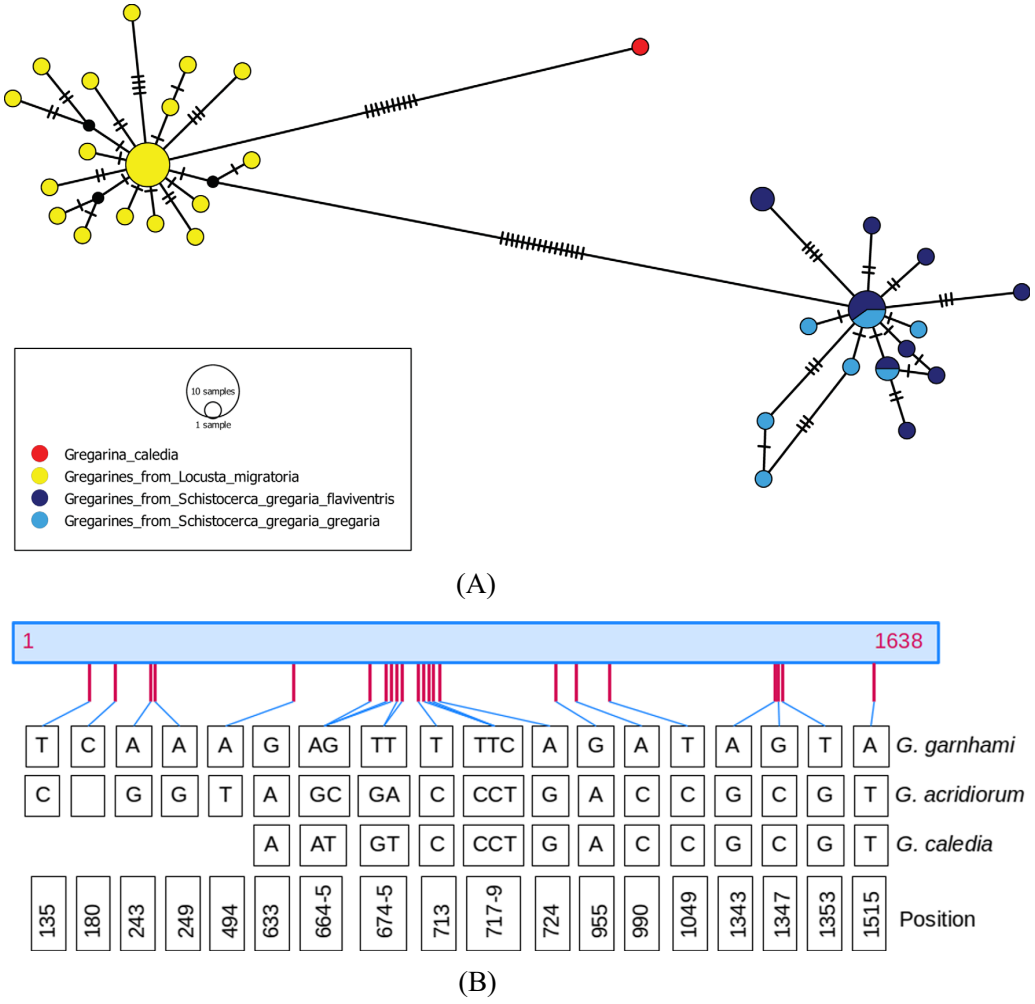


Figure 2. A: Minimum spanning network for the 43 cloned sequences of the SSU rDNA region studied, and the published sequence of *G. caledia* (L31799). Each link between haplotypes indicates one mutation, including indel events. The colors indicate the species or subspecies of the host. This network was inferred using POPART [25]. B: Distribution of the 22 polymorphic positions in SSU rDNA locus regions V1-V8 (1638-bp), between type 1 (presumably *G. garnhami*) (n = 20) and type 2 (presumably *G. acridiorum*) (n = 23) sequences, amplified from gregarines parasitizing respectively *S. gregaria* and *L. migratoria*. The corresponding positions are also given for *G. caledia* (L31799, 1210 bp) parasitizing *Caledia captiva*. Eleven additional positions, otherwise conserved between *G. garnhami* and *G. acridiorum* sequences, are modified in *G. caledia* sequence: site 1059, G deletion; sites 1161-1164: GAGC substituted by AG-G; site 1181: G substituted for C; site 1187: G substituted for A; sites 1231 and 1240: T substituted for C; site 1493: T insertion; site 1584: G substituted for A.

identified between "type 1" and "type 2" sequences (assumed to be *G. garnhami* and *G. acridiorum*, respectively), as schematized in Figure 2B.

Phylogenetic analysis

A phylogenetic approach, using partial SSU rDNA sequences and both maximum likelihood and Bayesian inference reconstructions, indicated that gregarine sequences from the two different host species studied clustered with sequences from other Gregarinoidea species (as described in [9, 11, 35]) with a high ML bootstrap value and Bayesian posterior probability (Fig. 3). These novel gregarine sequences form two clearly distinct clades according to their host species, and it thus appears that all *S. gregaria* hosts, regardless of their subspecies and the geographical location at which they were maintained, were infected by the same species (based on their SSU rDNA sequence) that was clearly distinct from the parasitic species infecting *L. migratoria*. The SSU rDNA sequence from *G. caledia* showed closer affinity to gregarine sequences from the host *L. migratoria* than from the host *S. gregaria* (see also Fig. 2). Furthermore, we observed that hosts of the "type 2" (presumably *G. acridiorum*) and *G. caledia* sequences, i.e. *L. migratoria* and *C. captiva*, belong to the same clade B of the acridian phylogeny as defined by Song et al. 2018 [38], while *S. gregaria*, infested with *G. garnhami* ("type 1" sequences), belongs to a distinct clade D, as indicated in Figure 3. Thus, gregarine distribution appears to parallel the

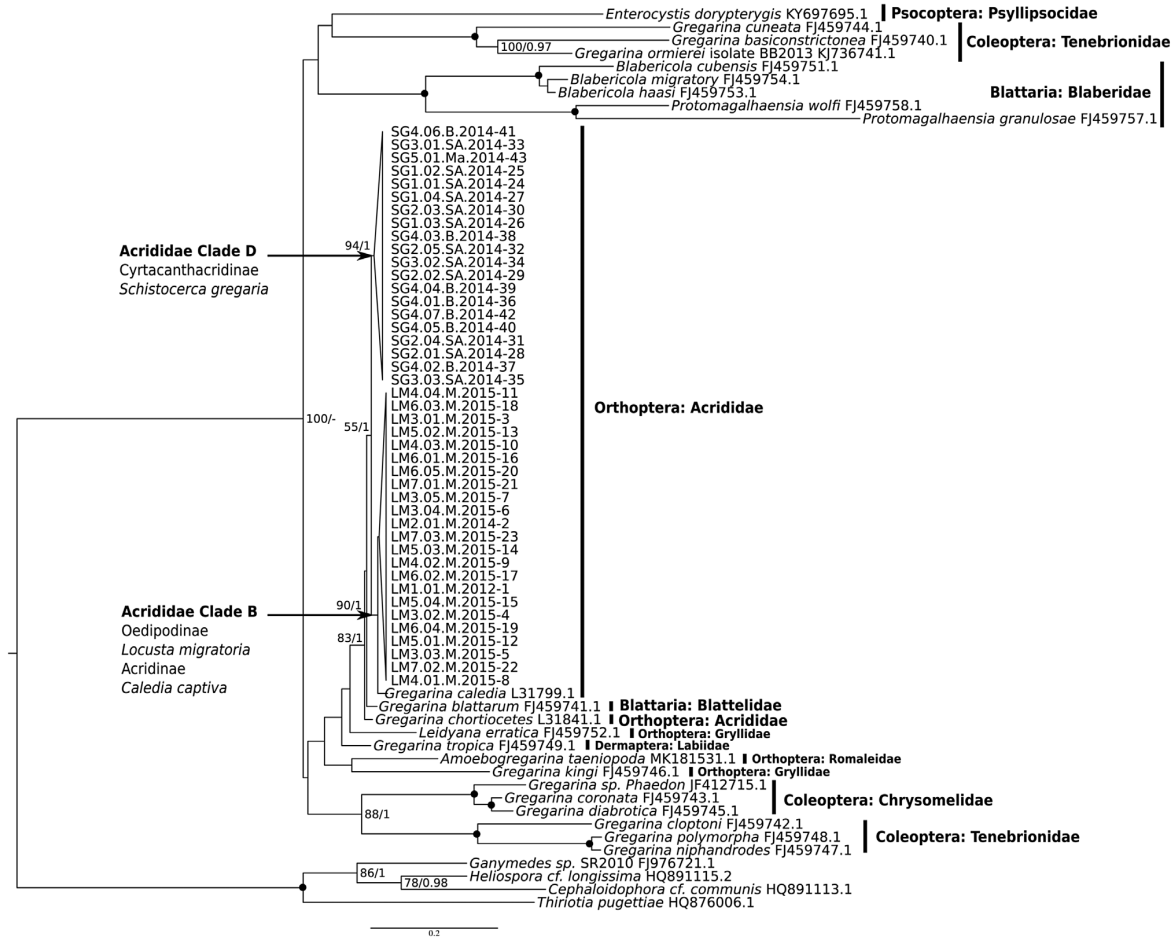


Figure 3. Phylogenetic tree built using 69 SSU rDNA sequences for 1,433 sites in order to zoom in on the clade Gregarinoidea including gregarines parasites of Orthoptera [11]. Outgroup consists of 4 sequences from Cephaloidophoroidea species that infect crustaceans, currently considered as the sister group of Gregarinoidea [29]. Evolutionary history is inferred by maximum likelihood and Bayesian inference using a GTR substitution model with gamma-distributed rate variation across sites plus invariant sites. Maximum likelihood topology is shown, with supports from both methods. Bootstrap < 75% and posterior probabilities < 0.95 are not shown. Black spots indicate 100/1 supports. The gregarines infecting *L. migratoria* clustered with *G. caledia*, isolated from the grasshopper *Caledia captiva* [30], the gregarines infecting *S. gregaria* forming a distinct independent clade. *G. chortiocetes,* infecting the locust *Chortiocetes terminifera* [30], and *Gregarina blattarum*, infecting the cockroach *Blatella germanica* [11] form sister branches to this group. The taxonomy of locust hosts is indicated, as established by Song et al, [38].

taxonomy of these three hosts. This observation will however need to be confirmed, as the ML bootstrap support remains low (55), despite high Bayesian posterior probability (Fig. 3).

Discussion

To determine whether the acridian orthopterans *S. gregaria* and *L. migratoria* are infected by the same gregarine species, their parasites were isolated and morphological and molecular analyses were performed using a series of host specimens of both species collected from a range of different locations and insect raising facilities (Table 2). While morphological investigations confirmed highly similar parasites with only tenuous morphological and behavioral differences, molecular investigations yielded unambiguous results strongly supporting different gregarine species in these *S. gregaria* and *L. migratoria* hosts.

Molecular data support distinct species

Molecular characterization, based on the partial SSU rDNA marker (V1–V8 region [18]) of all gregarines studied, unambiguously demonstrated that all *S. gregaria* hosts – regardless of their subspecies and raising facilities – are infected by the same gregarine species (presumably *G. garnhami*), whereas all *L. migratoria* hosts are infected by a distinct species (presumably *G. acridiorum*). Both gregarine sequences clustered in the previously reported Gregarinoidea clade [11, 29, 31].

Overall, 22 different bases were identified in this 1638 bp region that could be used to delimit the species. The 1.5% genetic distance between the two sequences is in agreement with previously described inter-specific levels of genetic divergence that, for example, distinguish *Gregarina niphandrodes* from *Gregarina polymorpha* (1.44%) [31]. However, it should be noted that, according to the same authors, such "low" genetic divergence could also correspond to "intra-specific" variability [31]. Certainly, additional studies will be needed to clarify this issue, but we recently demonstrated that two marine gregarines with an almost identical SSU rDNA sequence (1 bp difference for 1702 positions, i.e. ~0.05% divergence) displayed ~10% overall nucleic acid divergence at the genomic level, preventing genetic crossing, i.e. arguing for different species (I. Florent and J. Boisard, unpublished data).

Based on these molecular results and on data in the literature, notably the identification of their hosts, we propose that the "type 1" sequence found in gregarines infecting *S. gregaria* hosts may correspond to the species named *G. garnhami*, reported by several authors and collected from *S. gregaria* [7, 42]. The gregarine species found in *L. migratoria* likely corresponds to *G. acridiorum*, in agreement with Léger [24], but not with the proposal of Lipa et al. [27].

Some morphological and behavioral features discriminate the two species

To further confirm that two distinct gregarine species infect *S. gregaria* vs. *L. migratoria*, we focused on their possibly discriminating morphological and behavioral differences. Several morphological characters have been proposed in the literature to discriminate acridian gregarines, including: (1) the number and length of sporoducts, (2) the size of oocysts, and (3) the presence of a sharply pointed posterior extremity in *G. garnhami* versus a rounded extremity in *G. acridiorum* gamonts (see Table 1), even though, as indicated by Lange and Wittenstein, 2002, "such morphological features are probably not sufficient to delimit species, as very similar values in ranges and ratios were found between them" [22].

The sporoducts were indeed shorter in gregarines that infect *S. gregaria* (Fig. 1E) than in gregarines that infect *L. migratoria*, (Fig. 1L), supporting the hypothesis that *S. gregaria* can be infected by *G. garnhami* (~1/3 of the diameter of the gametocysts, Table 1) and *L. migratoria* by *G. acridiorum* (~1/2 of the diameter of the gametocysts, Table 1). The comparative study of sizes of barrel-shaped oocysts led to a less definitive indication. In gregarines that infect *S. gregaria*, the measurements (6.54 ± 0.32 µm × 4.32 ± 0.23 µm, n = 89) closely matched those reported in the literature for *G. garnhami* (6.5–7 µm × 4 µm, Table 1), compared to the remaining four species (Table 1). In gregarines that infect *L. migratoria*, these measurements (6.83 ± 0.27 µm × 3.99 ± 0.19 µm, n = 40) somewhat resemble those proposed in the literature for *G. acridiorum* (7.6 × 3.3 µm, Table 1), but are also very similar to the values reported for *G. garnhami* (6.5–7 µm × 4 µm, Table 1). However, these measurements are clearly more distantly related to the measurements reported for oocysts of the three other morphologically similar species: *G. rigida* (8 µm × 5 µm), *G. ronderosi* (5 µm × 3.2 µm), and *G. caledia* (12 µm × 6 µm) (Table 1).

However, the sharp (*G. garnhami*) versus round (*G. acridiorum*) posterior extremity of gamonts, proposed as a distinguishing feature between these two species, was not always reliably observed in our study and was therefore not retained as a distinguishing feature. Also, the number of sporoducts per gametocyst, currently reported in the literature to be larger in *G. acridiorum* (12–15) than in *G. garnhami* (8) (see Table 1), does not support our hypothesis that *G. acridiorum* is present in *L. migratoria* and *G. garnhami* is present in *S. gregaria*, as we observed the contrary: the number of sporoducts was less than eight for gregarines infecting *L. migratoria* (Fig. 1L) and more than eight for gregarines infecting *S. gregaria* (Fig. 1E). However, as previously mentioned by Clopton et al., 2009, the number of sporoducts is probably an unreliable taxonomical character [11]. Gametocysts diameters were also larger in *G. garnhami* (350 ± 56 µm, n = 36) vs. *G. acridiorum* (227 ± 35 µm, n = 18), but with overlapping values (210–420 µm for *G. garnhami*; 190–296 µm for *G. acridiorum*).

In the course of this study, we identified a third distinctive feature that is rarely mentioned in the literature: the fact that gregarines were systematically observed in the ceca of *S. gregaria* but never in the ceca of *L. migratoria*. The presence of *G. garnhami* but also *G. rigida*, *G. ronderosi* and *G. caledia* in the ceca of their hosts has also been systematically reported (Table 1) but interestingly, only the midgut was reported to be infected in the host specimens examined by Léger 1893, which included *L. migratoria* [24]. Whether this behavioral difference results from differences between gregarine species, in terms of ecological niche or host-parasite relationship, or from anatomical specificities in the two infected hosts, as already suggested [4], needs to be investigated experimentally. This third difference further supports the hypothesis that the two gregarines that infect either *S. gregaria* or *L. migratoria* should be considered distinct species. Remarkably, the gregarines recorded by Lipa et al. [27] in different acridian species, developed in the midgut but also in the gastric intestinal ceca of their hosts, a habitat that could indicate that they were infected by *G. garnhami* rather than by *G. acridiorum*. Alternatively, these acridian species may have hosted entirely novel (cryptic) gregarine species that remain to be characterized.

In addition to the morphological and developmental differences described above, these two gregarines share many peculiarities such as the ectocyst and the orange basal discs involved in gametocyst encystment then dehiscence [17]. The ectocyst, which designates the thick outer gelatinous layer or translucent hyaline coat of the gametocyst, is found in a wide range of gregarines of Orthoptera and is probably an adaptation to the host environment that makes it possible to keep the developing gregarine in a moist atmosphere [17]. Basal discs, involved in the extrusion of the sporoducts of all gregarines belonging to the superfamily Gregarinoidea, are widely observed in Hexapoda hosts [17]. The basal discs are orange in all the gregarines of Orthoptera and the gamonts are often pale yellow, as we observed here for gregarines infecting both *S. gregaria* and *L. migratoria* hosts. Importantly, it is possible that these morphological features are the product of

plasticity, so their taxonomical significance remains to be explored.

Taxonomic consequences

Based on these differences and on the available literature, we thus endorse the hypothesis that the species that infect *S. gregaria* should bear the species name *G. garnhami,* in agreement with the morphological characters established for this species (Table 1) and in agreement with a previous proposal by Valigurova and Koudela [42]. Indeed, these authors already disputed the interpretation of Lipa et al. (1996) [27] arguing that in their studies, Lipa et al. did not observe the developmental stages that are able to differentiate these species, i.e. the number and length of the sporoducts involved in the dehiscence process and the size and shape of their oocysts [42]. Concerning the species that infect *L. migratoria,* we maintain our proposal to name them *G. acridiorum*, even though only in partial agreement with the morphological characters established for this species (Table 1). This proposal is logical given the taxonomic history of this species, as the first *Gregarina* species found to infect *L. migratoria* was called *Gregarina (Clepsidrina) acridiorum* [21, 24], and the absence of colonization of the hosts's ceca. Although the size and shape of the oocysts we observed in the gregarine infecting *L. migratoria* do not perfectly match the measurements reported for *G. acridiorum* in the literature (Table 1), the size and shape of the oocysts we observed in gregarines that infect *S. gregaria* perfectly match the measurements reported for *G. garnhami* in the literature. However, the oocyst in gregarines found in *L. migratoria* were clearly thinner and longer than the oocysts in gregarines found in *S. gregaria*, observed in similar conditions. The observed length of sporoducts also agrees with data reported for both species in the literature, unlike the observed number of basal discs/sporoducts developing at the surface of gametocysts in dehiscence (Table 1). As mentioned above, this point should be interpreted with caution as it has been reported that the number of basal discs and the development of sporoducts may vary according to environmental conditions (temperature, hygrometry) as well as possibly the size of the gametocysts [7, 11].

Morphological plasticity and host conditions

The morphological data showed that the developmental stages of the gregarines infecting *S. gregaria* (Figs. 1A–1G) were generally very similar, though slightly longer and larger than the developmental stages of the gregarines infecting *L. migratoria* (Figs. 1H–1M). However, depending on the raising facility and/or geographic origin, gregarines – notably trophozoites and gamonts – appeared to be slimmer in *S. g. gregaria* hosts from Morocco and *S. g. flaviventris* hosts from South Africa (not shown) than in gregarines infecting *S. g. gregaria* hosts from Belgium (Fig. 1D). The latter, which were much stockier, were more similar to the gamont stages of the gregarines that infect *L. migratoria* (Figs. 1H–1K). Since *S. g. flaviventris* hosts and *S. g. gregaria* hosts from the South African and Moroccan facilities, respectively, were also observed to be unhealthy (mature adults behaved sluggishly and seemed soft and light from food), while *S. g. gregaria* hosts from the Belgium facility and the *L. migratoria* hosts maintained in France did not seem to be particularly affected by the presence of their infecting gregarines (see also Table 2), we favor the hypothesis that environmental differences or co-occurring microorganisms may explain the difference in "fitness" between "African" and "European" hosts, as this was not due to infections by distinct gregarine species.

How many distinct species are there for these gregarines?

The gregarine developmental stages described in *S. gregaria* and *L. migratoria* hosts are very similar morphologically, and share many characteristics including the thick mucilaginous ectocyst of the gametocyst, orange basal discs associated with great variability of size parameters. As these morphological features have also been observed in other species, particularly in *G. rigida*, *G. ronderosi* and *G. caledia* collected from different (and sometimes from identical) orthopteran hosts (Table 1), these species need to be further characterized at the molecular level to solve their phylogenetic relationships. The only molecular sequence available (*G. caledia*, L31799) although rather small (1210 bp) strongly suggests a third distinct species, closely related phylogenetically to the proposed *G. acridiorum* but still with some observed genetic distance (1.1%). *Gregarina caledia* is also potentially morphologically distinguishable by the larger size of its oocysts and its ability to infect host ceca (Table 1, [30]). Importantly though, in the first morphological reports, this species was said to be closely related to *G. garnhami* with which it also shares the ability to infect host ceca [30].

Gregarina rigida (Hall, 1907) Ellis, 1913, has also been reported in a range of orthopterans. When describing this species, the authors did not cite any literature on *G. acridiorum*, so, in 1968, Corbel concluded that *G. rigida* was a junior synonym of *G. acridiorum* [14]. To be confirmed, the status of this species (e.g. synonym of *G. acridiorum*?) therefore requires molecular data, even though available measurements of oocysts and the fact that it has also been found in host ceca (Table 1) favor a distinct species. Importantly, in 2002, *G. ronderosi*, which was found in the argentine grasshopper, *Dichroplus elongatus*, was named a novel species by Lange and Wittenstein due to the lack of infection in specimens of *L. migratoria* experimentally infected with this gregarine [22]. It thus also possibly corresponds to a fifth distinct species, also awaiting molecular characterization. Lange and Wittenstein, 2002, even suggested that *G. ronderosi* could be synonymous with *G. garnhami*, but that molecular data were required as morphometric differences did not enable conclusive delimitation of the species [22].

Conclusion

It is well documented that assigning protist species can no longer rely on morphological information alone, but should include molecular data in an integrated taxonomic approach [5, 6]. The data presented here confirm that most morphological

and morphometric differences cannot conclusively delimit closely related species, while molecular data can reveal clearly measurable differences. By strongly suggesting that *S. gregaria* is infected by *G. garnhami*, whereas *L. migratoria* is infected with *G. acridiorum*, our data suggest two important discriminating features: the respective size of the oocysts of *G. garnhami* and *G. acridiorum*, but also their location in their respective host's gut. The first consequence is that *G. garnhami* can no longer be considered a junior synonym of *G. acridiorum,* contrary to the proposal by Lipa et al. [27] and is therefore reinstated here as a valid taxon, in agreement with the proposal of Valigurova and Koudela [42].

The exact distribution of *G. garnhami* and *G. acridiorum* in Orthoptera remains to be further investigated at this stage as clearly, when synonymized, they were assumed to infect the same series of host species [17]. Additional studies, specifically molecular studies, are crucial to determine the diversity of gregarine species that infect acridians, beyond the establishment of morphological specificities (see Table 1). This could help determine whether *G. rigida* and *G. ronderosi* are in fact distinct species or should be synonymized with other species. Interestingly, *G. caledia*, a parasite of the Australian locust *C. captiva* reported to be very similar to *G. garnhami* and for which molecular data are available [30], should be considered a species distinct from both *G. garnhami* and *G. acridiorum* as argued in this paper. Based on our molecular studies, *G. caledia* presents closer phylogenetic similarity to *G. acridiorum* (Fig. 3). A major challenge concerns the precise diversity of the species *G. acridiorum* that has been described in over 60 orthopteran hosts, from both the Caelifera and Ensifera orders, as is also the case for *G. rigida*. It is likely that these two species correspond to a much greater diversity of probably cryptic species that remain to be described by this type of integrative taxonomical approach, in the diversity of their currently described hosts.

Conflict of interests

The authors declare that they have no conflict of interest.

Acknowledgements. This work was supported by a grant from the French *Agence Nationale de la Recherche* [LabEx ANR-10-LABX-0003-BCDiv], in the program "*Investissements d'avenir*" [ANR-11-IDEX-0004-02], by several interdisciplinary Programs of the MNHN (*ATM-Microorganismes*, *ATM-Génomique et Collections*, ATM-Emergence, AVIV department), the CNRS (Julie Boisard's PhD fellowship, 2018–2021) and the French Agricultural Research Centre for International Development (CIRAD). We are very grateful to Laure Wasniewski and Judykaelle Fede for providing the first molecular data on the gregarines infecting *L. migratoria*, to Lisy Ravendran for technical assistance with SEM sample preparation, to Geraldine Toutirais and the MNHN Platform (Plateau technique de Microscopie Électronique, Muséum National d'Histoire Naturelle MNHN, Paris, France, http://ptme.mnhn.fr/) for SEM image acquisition, to Laure Benoit for assistance with network figure and locust sample preparation, and to Hélène Jourdan for stimulating discussions at an early stage of the study. We are grateful to J. Vanden Broeck and C. Piou, for providing desert locust egg pods from the University of Leuven and from the CNLA of Agadir, Morocco.

References

1. Adl SM, Bass D, Lane CE, Lukes J, Schoch CL, Smirnov A, Agatha S, Berney C, Brown MW, Burki F, Cardenas P, Cepicka I, Chistyakova L, Del Campo J, Dunthorn M, Edvardsen B, Eglit Y, Guillou L, Hampl V, Heiss AA, Hoppenrath M, James TY, Karnkowska A, Karpov S, Kim E, Kolisko M, Kudryavtsev A, Lahr DJG, Lara E, Le Gall L, Lynn DH, Mann DG, Massana R, Mitchell EAD, Morrow C, Park JS, Pawlowski JW, Powell MJ, Richter DJ, Rueckert S, Shadwick L, Shimano S, Spiegel FW, Torruella G, Youssef N, Zlatogursky V, Zhang Q (2019) Revisions to the classification, nomenclature, and diversity of Eukaryotes. Journal of Eukaryotic Microbiology, 66(1), 4–119.
2. Allain T, Chaouch S, Thomas M, Vallee I, Buret AG, Langella P, Grellier P, Polack B, Bermudez-Humaran LG, Florent I. 2017. Bile-salt-hydrolases from the probiotic strain *Lactobacillus johnsonii* La1 mediate anti-giardial activity in vitro and in vivo. Frontiers in Microbiology, 8, 2707.
3. Ayali A. 2019. The puzzle of locust density-dependent phase polyphenism. Current Opinion in Insect Science, 35, 41–47.
4. Bernays EA. 1981. A specialized region of the gastric ceca in the locust, *Schistocerca gregaria*. Physiological Entomology, 6(1), 1–6.
5. Berney C, Ciuprina A, Bender S, Brodie J, Edgcomb V, Kim E, Rajan J, Parfrey LW, Adl S, Audic S, Bass D, Caron DA, Cochrane G, Czech L, Dunthorn M, Geisen S, Glockner FO, Mahe F, Quast C, Kaye JZ, Simpson AGB, Stamatakis A, Del Campo J, Yilmaz P, de Vargas C. 2017. UniEuk: Time to speak a common language in Protistology! Journal of Eukaryotic Microbiology, 64(3), 407–411.
6. Boenigk J, Ereshefsky M, Hoef-Emden K, Mallet J, Bass D. 2012. Concepts in protistology: species definitions and boundaries. European Journal of Protistology, 48(2), 96–102.
7. Canning EU. 1956. A new Eugregarine of locusts, *Gregarina garnhami* n.sp., parasitic in *Schistocerca gregaria* Forsk. Journal of Protozoology, 3(2), 50–62.
8. Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular Biology and Evolution, 17(4), 540–552.
9. Cavalier-Smith T. 2014. Gregarine site-heterogeneous 18S rDNA trees, revision of gregarine higher classification, and the evolutionary diversification of Sporozoa. European Journal of Protistology, 50(5), 472–495.
10. Chapuis MP, Bazelet CS, Blondin L, Foucart A, Vitalis R, Samways MJ. 2016. Subspecific taxonomy of the desert locust, *Schistocerca gregaria* (Orthoptera: Acrididae), based on molecular and morphological characters. Systematic Entomology, 41(3), 516–530.
11. Clopton RE. 2009. Phylogenetic relationships, evolution, and systematic revision of the septate gregarines (Apicomplexa: Eugregarinorida: Septatorina). Comparative Parasitology, 76, 167–190.
12. Corbel JC. 1967. Intensité, fréquence et facteurs des infections grégariniennes chez les Insectes Orthoptères. Annales de Parasitologie Humaine et Comparée, 42, 373–385.
13. Corbel JC. 1968. New or poorly known gregarines parasitic on orthopterus insects. Annales de Parasitologie Humaine et Comparée, 43(3), 291–320.
14. Corbel JC. 1968. Parasitic specificity of gregarines of Orthoptera. Annales de Parasitologie Humaine et Comparée, 43(1), 25–32.
15. de Vargas C, Audic S, Henry N, Decelle J, Mahe F, Logares R, Lara E, Berney C, Le Bescot N, Probert I, Carmichael M, Poulain J, Romac S, Colin S, Aury JM, Bittner L, Chaffron S, Dunthorn M, Engelen S, Flegontova O, Guidi L, Horak A, Jaillon O, Lima-Mendez G, Lukes J, Malviya S, Morard R,

Mulot M, Scalco E, Siano R, Vincent F, Zingone A, Dimier C, Picheral M, Searson S, Kandels-Lewis S, Acinas SG, Bork P, Bowler C, Gorsky G, Grimsley N, Hingamp P, Iudicone D, Not F, Ogata H, Pesant S, Raes J, Sieracki ME, Speich S, Stemmann L, Sunagawa S, Weissenbach J, Wincker P, Karsenti E. 2015. Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. Science, 348(6227), 1261605.

16. Del Campo J, Heger TJ, Rodriguez-Martinez R, Worden AZ, Richards TA, Massana R, Keeling PJ. 2019. Assessing the diversity and distribution of apicomplexans in host and free-living environments using high-throughput amplicon data and a phylogenetically informed reference framework. Frontiers in Microbiology, 10, 2373.
17. Desportes I, Schrével J. 2013. Treatise on Zoology – Anatomy, Taxonomy, Biology. The Gregarines, The early branching Apicomplexa (2 vols). Brill. p. 791.
18. Hadziavdic K, Lekang K, Lanzen A, Jonassen I, Thompson EM, Troedsson C. 2014. Characterization of the 18S rRNA gene for designing universal eukaryote specific primers. PLoS One, 9(2), e87624.
19. Hussain KJ, Krishnan SM, Johny S, Whitman DW. 2013. Phenotypic plasticity in a gregarine parasite (Apicomplexa: Eugregarinordia) infecting grasshoppers. Comparative Parasitology, 80(2), 233–239.
20. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Molecular Biology and Evolution, 35(6), 1547–1549.
21. Labbé A. 1899. Sporozoa, Das Tierreich: Eine Zusammenstellung und Kennzeichnung der rezenten Tierformen. R. Friedlander und Sohn: Berlin, Germany. p. 180.
22. Lange CE, Wittenstein E. 2002. The life cycle of *Gregarina ronderosi* n. sp. (Apicomplexa: Gregarinidae) in the Argentine grasshopper *Dichroplus elongatus* (Orthoptera: Acrididae). Journal of Invertebrate Pathology, 79(1), 27–36.
23. Léger L. 1892. Recherches sur les grégarines. Tablettes Zoologiques, 3, 1–182.
24. Léger L. 1893. Sur une grégarine nouvelle des Acridiens d'Algérie. Comptes Rendus de l'Académie des Sciences de Paris, 117, 811–813.
25. Leigh JW, Bryant D. 2015. Popart: full-feature software for haplotype network construction. Methods in Ecology and Evolution, 6(9), 1110–1116.
26. Levine ND. 1988. Progress in taxonomy of the Apicomplexan protozoa. Journal of Protozoology, 35(4), 518–520.
27. Lipa JJH-CP, Santiago-Alvarez C. 1996. Gregarines (Eugregarinorida: Apicomplexa) in natural populations of *Dociostaurus maroccanus*, *Calliptamus italicus* and other Orthoptera. Acta Protozoologica, 35, 49–59.
28. Mahé F, de Vargas C, Bass D, Czech L, Stamatakis A, Lara E, Singer D, Mayor J, Bunge J, Sernaker S, Siemensmeyer T, Trautmann I, Romac S, Berney C, Kozlov A, Mitchell EAD, Seppey CVW, Egge E, Lentendu G, Wirth R, Trueba G, Dunthorn M. 2017. Parasites dominate hyperdiverse soil protist communities in Neotropical rainforests. Nature Ecology & Evolution, 1(4), 91.
29. Medina-Duran JH, Mayen-Estrada R, Marino-Perez R, Song H. 2020. Morphology and Phylogenetic position of two new gregarine species (Apicomplexa: Eugregarinorida) parasitizing the Lubber Grasshopper *Taeniopoda centurio* (Drury, 1770) (Insecta: Orthoptera: Romaleidae) in Mexico. Journal of Eukaryotic Microbiology, 67(1), 4–17.
30. Ninham JA. 1995. Phylogenetic analysis of five protoctist parasites of insects. PhD thesis. Australian National University: Canberra.
31. Nocciolini C, Cucini C, Leo C, Francardi V, Dreassi E, Carapelli A. 2018. Assessing the efficiency of molecular markers for the species identification of Gregarines Isolated from the Mealworm and super worm midgut. Microorganisms, 6(4), 119.
32. Portman N, Slapeta J. 2014. The flagellar contribution to the apical complex: a new tool for the eukaryotic Swiss Army knife? Trends in Parasitology, 30(2), 58–64.
33. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Systematic Biology, 61(3), 539–542.
34. Sambrook J, Russell DW. 2006. Purification of nucleic acids by extraction with phenol:chloroform. Cold Spring Harbor Protocols, 2006(1).
35. Schrével J, Valigurova A, Prensier G, Chambouvet A, Florent I, Guillou L. 2016. Ultrastructure of *Selenidium pendula*, the type species of Archigregarines, and phylogenetic relations to other marine Apicomplexa. Protist, 167(4), 339–368.
36. Seck AG, Toguebaye BS. 1995. Étude taxonomique de quelques grégarines (Protozoa, Apicomplexa) parasites de criquets du Sénégal. Description d'*Actinocephalus pyrgomorphae* n. sp. Bulletin de l'Institut Fondamental d'Afrique Noire Cheikh Anta Diop Sénégal, série A, 48, 37–47.
37. Semans FM. 1941. Protozoan parasites reported from the Orthoptera, with special reference to those of Ohio. III. Protozoan parasites in relation to the host and the host ecology. Ohio Journal of Science, 41, 457–464.
38. Song H, Mariño-Pérez R, Woller DA, Cigliano MM. 2018. Evolution, diversification, and biogeography of grasshoppers (Orthoptera: Acrididae). Insect Systematics and Diversity, 2(4), 3, 1–25.
39. Souidenne D, Florent I, Dellinger M, Justine JL, Romdhane MS, Furuya H, Grellier P. 2016. Diversity of apostome ciliates, *Chromidina* spp. (Oligohymenophorea, Opalinopsidae), parasites of cephalopods of the Mediterranean Sea. Parasite, 23, 33.
40. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics, 30(9), 1312–1313.
41. Uvarov BP. 1977. Grasshoppers and Locusts, vol. 2, Centre for Overseas Pest Research: London.
42. Valigurova A, Koudela B. 2008. Morphological analysis of the cellular interactions between the eugregarine *Gregarina garnhami* (Apicomplexa) and the epithelium of its host, the desert locust *Schistocerca gregaria*. European Journal of Protistology, 44(3), 197–207.

Cite this article as: Florent I, Chapuis MP, Labat A, Boisard J, Leménager N, Michel B & Desportes-Livage I. 2021. Integrative taxonomy confirms that *Gregarina garnhami* and *G. acridiorum* (Apicomplexa, Gregarinidae), parasites of *Schistocerca gregaria* and *Locusta migratoria* (Insecta, Orthoptera), are distinct species. Parasite **28**, 12.

Marine gregarine genomes illuminate current understanding of apicomplexan glideosome

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

Author Contact:

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

Isabelle Florent: isabelle.florent@mnhn.fr

Loïc Ponger: loic.ponger@mnhn.fr

Lead Contact:

Isabelle Florent

Author Affiliation:

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

SUMMARY

Apicomplexans, parasite protists of a very wide diversity of metazoan hosts, are mostly known from species infecting human. Absence or limited data for basal lineages prevents a comprehensive view of evolutionary history and adaptive capacities of Apicomplexa. Here, we characterized the genome of the marine eugregarine *Porospora gigantea*, remarkable for the gigantic size of its vegetative feeding forms (trophozoites) and their speed of gliding movement, the fastest so far recorded for an Apicomplexa. Not a single but two highly related genomes named A and B were assembled. Highly syntenic, of similar size (9 Mb) and coding capacities (~5300 genes), they display a 10.8%

divergence at nucleotide level corresponding to 16-38 My divergent time. Orthogroups analyses across 25 (proto)Apicomplexa including *Gregarina niphandrodes* showed that A and B are highly divergent from all other known apicomplexan species, revealing an unexpected diversity. These two related species branch in phylogenetic studies at the base of Cephaloidophoroidea, forming a new family in these crustacean gregarines. Gliding proteins data mining found a strong conservation of actin-related proteins, as well as of regulatory factors, within apicomplexan. In contrast, the conservation of core glideosome proteins and adhesion proteins appears to be highly variable among apicomplexan lineages, especially in gregarines. These results confirm the importance of studying gregarines to widen our biological and evolutionary view of apicomplexan parasites, to better apprehend species diversity and revise our understanding of the molecular bases of some key functions such as observed for the glideosome.

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

INTRODUCTION

Apicomplexan are unicellular eukaryotic microorganisms that have evolved towards a strict parasitic lifestyle. Some species are extremely pathogenic such as *Plasmodium* spp., *Toxoplasma gondii* and *Cryptosporidium* spp., responsible for malaria, toxoplasmosis and cryptosporidiosis. While the genomic knowledge of apicomplexan parasites currently concerns a dozen of genera, more precisely those of highly pathogenic agents[1], it remains highly skewed towards intracellular parasites of vertebrates, notably Coccidia, Hemosporidia and *Cryptosporidium* (see references in Table S1). Yet, the Apicomplexa include about 350 genera[2] for 6,000 documented species. This includes gregarines, which represent early diverging lineages with low pathogenicity that have attracted less interest from biomedical research. However, gregarines include at least 1,770 species[3], their diversity being highly understated, as gregarines were identified as the most abundant and widely reported apicomplexan in a recent environmental study[4]. As they remain non-cultivable organisms, their study in the laboratory is technically challenging, explaining why gregarines have hardly been explored at the -omic levels[5].

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

To overcome this lack of data, we focused on the marine eugregarine *Porospora gigantea* (van Beneden, 1869) Schneider, 1875, an intestinal parasite of lobster *Homarus gammarus. Porospora gigantea* was described in 1869 by E. van Beneden, who called it *Gregarina gigantea*, by reference to the gigantic size taken by its trophozoite stages, up to 16,000 µm i.e. visible by the naked eye[14]. He reported that “cyst” forms of this parasite accumulated within the chitineous folds of the host’s rectum, the “rectal ampulla”. Schneider further showed that these cysts enclosed thousands of “gymnospores” or “heliospores”, corresponding to spherical groups of very tiny zoites radiating from a central, optically void mass, and renamed the species *Porospora gigantea* (van Beneden, 1869) Schneider, 1875[15]. As biological material from non-cultivable microorganisms is particularly

difficult to gather, we took advantage of the existence of these structures later confirmed by others[16–19]. Indeed, they provide a remarkable natural source of genomic DNA, each cyst containing several thousands of "gymnospores", themselves composed of hundreds "zoites". *Porospora gigantea* trophozoites, also known to glide at rates of up to 60µm/s[20], appears as prime candidates to study the mechanism of gliding, a specific form of motility which is a characteristic of Apicomplexa[21–25]. Currently about 40 proteins have been identified mainly in *T. gondii* and *P. falciparum,* and assembled in a commonly accepted structural model, the glideosome (see Frénal et al, 2017[26] for review).
In this study, we report the first draft genome of *P. gigantea.* Remarkably, not one but two related genomes have been assembled. We present their main characteristics and their associated proteomes in the context of the other available apicomplexan genomes. We also determined their position within the phylogeny of marine gregarines. One of the main objectives of this study was to provide an overview of the conservation of proteins involved in the gliding, at the apicomplexan level including gregarines.

RESULTS

Phenotypic characterization

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

Two highly related genomes

Four biological samples were sequenced and analyzed independently, and then secondarily assembled together (Figure 2.A). The raw assembly produced 214,938 contigs (99.6Mb) among which 13,656 contigs had a length greater than 1kb (47.9Mb). The obtained scaffolds were cleaned by removing contaminants such as bacteria, fungi and host sequences (Figure 2.B), which created a raw assembly of 1719 contigs for 18Mb.
The analysis of contigs coverage by each individual library revealed a bimodal distribution suggesting a mixture of genomes with a proportion depending on the biological sample (Figure 2, Figure S3). More precisely, while only one set of scaffolds displayed a significant coverage for the lobster tank parasite sample (JS-470, peak around 250X), the three other parasite samples, from freshly captured hosts (JS-482, JS-488, JS-489) showed two distinct sets of scaffolds with different coverage values. The *in-silico* analysis of these two sets revealed an equivalent genome size of ~9Mb. This coverage difference was used to split the contigs into two sets that were given the names A (for the set present in all four samples) and B (for the set present only in three lobsters freshly captured in the wild) (Figure 2.C). The proportion of genomes A and B in each biological DNA sample has been estimated (Figure S3) as 100%-0% for JS-470, 63.2%-36.8% for JS-482, 70.5%-29.5% for JS-488 and 62.4%-37.6% for JS-489, based on medium coverage levels. The genome A contains 786 scaffolds for a total of 8.8Mb whereas the genome B contains 933 scaffolds for a total of 9.0Mb. The contigs from both genomes can be aligned over 7.7Mb, with a percentage of divergence around 10.8% at nucleotide level.
These two genomes have a similar size (~9Mb), are highly syntenic with approximatively 10% of divergence. These highly related genomes have been named A and B and are associated to the species name *P. cf. gigantea* (Figure 2).

Genome features

Two genomes with similar coding capacities. A total of 10,631 putative genes were predicted on the raw assembly, which could be splitted into two sets of similar size: 5270 genes in genome A vs. 5361 genes in genome B (Table 1, Figure 2). The completeness of both A and B genomes was addressed by using the BUSCO software[29] and the Apicomplexa geneset (n=446). Genomes A and B respectively showed a completeness score of 70% (n=312) and 67.7% (n=302) using this Apicomplexa geneset (all BUSCO's scores are shown on Figure S4).
Orthologues were searched between A and B. The proteins of *P. cf. gigantea* A and B were splitted into 5656 orthogroups including 4443 groups (88%) with at least one orthologous gene for both A and B. This percentage of common orthogroups between genomes A and B is higher than the percentage of common orthogroups observed between *P. falciparum* and *P. berghei* (70%), documented to have diverged around 33 Mya ago (TimeTree[30]) but similar to that observed between *P. falciparum* and *P. reichenowi* (86%, 3.3 – 7.7 Mya, TimeTree).
The percentages of shared orthogroups between *P. cf. gigantea* genomes and each of the reference apicomplexan species are similar despite the highly variable divergence (*C. parvum*, 18%; *G. niphandrodes*, 17%; *P. falciparum*, 14%; *T. gondii*, 14%) but it is higher than the percentages observed with chromerid species (*C. velia*, 8%; *V. brassicaformis,* 10%). We can underline from this result that the *P. cf. gigantea* genomes don't share significantly more orthogroups with *G. niphandrodes*, the only other available gregarine genome (Figure 3).

Two gene dense genomes with small introns. The proportion of coding sequences (84%) in A and B genomes is particularly high compared to other reference species (from 25% to 76%; Table 1). The genomic compaction of non-coding DNA in genomes A and B can be explained by the small size of most introns (Figure S5). We observed a specific class of introns with a length around 25-30bp (mode at 28bp) representing 71-72% of the introns. The donor and acceptor sites of these

small introns display specific consensus patterns (Figure S5) which are different from the other *Porospora* introns. Especially, these introns exhibit a strongly conserved adenine located 6 bp upstream of the 3' of the acceptor site which could represents the intron branch point as observed for the small introns in *B. microti* (introns of 20bp)[31].

Evolutive history of both *P. cf. gigantea*

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

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

A partially conserved glideosome machinery

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

references proteins have been searched for in both *Porospora cf. gigantea* genomes and a selection of representative species.

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

Glideosome apicomplexan-specific proteins. The glideosome machinery mainly comprises specialized apicomplexan-specific proteins. The single-headed short heavy chain myosin class XIV, named myosin A (MyoA), acts as a motor generating the rearward traction required for motility, invasion and egress, as evidenced by various conditional depletion protocols[54–56]. The glideosome itself takes place between the plasma membrane and the apicomplexan-specific inner membrane complex (IMC), in which MyoA is associated with a light chain (myosin light chain 1 – MLC1 in *T. gondii*[57] and MyoA tail domain-interacting protein - MTIP in *P. falciparum*[58]) as well as glideosome associated proteins (GAP): GAP40, GAP45, GAP50[59–61], GAP70 and GAP80 being only described in *T. gondii*[55]. GAP45 binds the glideosome to the plasma membrane by recruiting MyoA that acts as a bridge[61]; GAP40 and GAP50 are thought to help anchoring MyoA to the parasite cytoskeleton[62], whereas another set of glideosome-associated proteins with multiple-membrane spans (GAPM) are believed to interact with the alveolin and subpellicular microtubules network, suggesting an indirect interaction with the IMC[26,63]. Finally, the conoid-associated myosin H is necessary for initiating gliding motility in *T. gondii*[64].
Myosins ABCDE and its associated light chain were found in all species. Myosin H is also widely distributed, although it is missing in *P. falciparum* and *V. brassicaformis*. The situation is more complex for glideosome associated proteins. Only GAP40 was found in all species, including probable homologues in chromerids. Surprisingly, given the central role attributed to GAP45 in the glideosome model, no orthologue was found in gregarines, *Cryptosporidium* or chromerids. GAP50 was found in all species except the two *P. cf. gigantea* genomes. As expected, GAP70 and GAP80, only identified in *T. gondii*, were not found in other species, except an orthologue for GAP80 in the coccidia *H. hammondi*. Concerning GAPMs, we found orthologues of at least one of its variations (GAPM 1, 2 or 3) for all species. Finally, GAC was found in all species except in chromerids, confirming its occurrence in apicomplexan only.

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

even species-specific. On the other hand, we found orthologues for ROM4 in all species, except for chromerids.
The TRAP proteins described to date have the following characteristics: an extracellular region containing one or more TSP1 domains and/or one or more vWA domains[73–75]. They are also characterized by the presence of a single transmembrane domain, a signal peptide, as well as, in some cases, a juxtaposed rhomboid protease cleavage site, and a short and charged C-terminal cytoplasmic domain, together with aromatic residues. The presence of a YXXΦ tyrosine sorting signature has also been described[74] (X meaning any amino acid, Φ a hydrophobic amino acid - isoleucine, leucine, methionine, phenylalanine, or valine).
In order to evaluate the presence of TRAP-like proteins in *P. cf. gigantea* genomes, we inventoried all proteins containing at least one TSP1 domain (Table S8), and identified potential candidates displaying several structural characteristics of TRAP-like family (Figure 6). We identified a CpTSP2 orthologue within both *P. cf. gigantea* genomes, PgTSP2. Like CpTSP2, it is a large protein (~2800aa) composed of Notch, TSP1, and Sushi domains. It has an addressing signal, a transmembrane domain and a short and charged basic cytoplasmic tail. This protein also has orthologues in *G. niphandrodes,* in chromerids and coccidia.
We also demonstrated the presence of four other protein pairs present in both A and B genomes, most of which appear to be specific to *P. cf. gigantea.* PgTSP-1 has a TSP1 domain, a signal peptide, a transmembrane domain and a short acidic and charged cytoplasmic tail. PgTSP-2, very similar in structure to PgTSP-1 also has a TSP1 domain, a signal peptide, a transmembrane domain and a short, charged but basic cytoplasmic tail. PgTSP_EGF-1 has two TSP1 domains, a signal peptide, a transmembrane domain and a short acidic charged cytoplasmic tail, and several EGF or EGF-like domains in their extracellular portion, as also described in *C. parvum* (CpTSP7, CpTSP8 and CpTSP9[79]). We also identified another protein very similar in structure, PgTSP_EGF-2.

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

Regulatory factors and signaling pathways. The increase of parasite intracellular calcium, by activating calcium-dependent protein kinases (CDPK), is involved in the regulation of motility, microneme secretion, invasion and egress[93,94]. Other proteins known in such signaling pathways include phosphatidylinositol phospholipase C (PI-PLC), diacylglycerol kinase 1 (DGK1), acylated pleckstrin homology domain-containing protein (APH) which are involved in micronemes secretion regulation [95,96], the C2 domains-containing protein DOC2.1 which mediates apical microneme exocytosis[97]; finally, the apical lysine methyltransferase (AKMT), which is involved in gliding motility, invasion and egress in *T. gondii*[98]. With the exception of the APH that we were unable to identify in *Porospora cf. gigantea* or chromerids, all the regulatory factors appeared to be largely conserved.

DISCUSSION

Molecular data support the presence of two species

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

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

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

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

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

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

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

Even among gregarines a wide diversity may be pointed out as the vast majority of proteins shared

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

The gregarine glideosome(s)

An incomplete but operational machinery. Our molecular analysis of the glideosome components shows that the currently known mechanistic model described from *T. gondii* and *P. falciparum* cannot fully account for gliding in all apicomplexans, as previously anticipated[26,62,66]. Some key molecular components such as canonical adhesins or GAP45 appear to be missing, implying an only partially retrieved glideosome in gregarines as well as in *Cryptosporidium* species (Figure 5.B). Since we have observed the gliding movement of *P. cf. gigantea* trophozoites at an impressive rate, this raised the question of how they manage to perform this rapid movement in the absence of a complete dedicated machinery.

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

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

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

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

ACKNOWLEDGEMENTS

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

This work was supported by a grant from the French *Agence Nationale de la Recherche* [LabEx ANR-10-LABX-0003-BCDiv], in the program "*Investissements d'avenir*" [ANR-11-IDEX-0004-02], by several interdisciplinary Programs of the MNHN (*ATM-Microorganismes*, *ATM-Génomique et Collections*, *ATM-Emergence*, AVIV department), the CNRS (Julie Boisard's PhD fellowship, 2018–2021).

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

AUTHOR'S CONTRIBUTION

JB, EDB, LP and IF designed the study. IF, JS and LG performed the biological sampling and IF extracted the nucleic acids. IF, JS and LG performed the photonic microscopy analyses, IF

performed the SEM while SLP and GP performed the TEM. JB, EDB and LP realized the bioinformatics analyses. AL and LD sequenced and assembled the complete ribosomal loci. JB performed the phylogenetic analyses and the glideosome expert annotation.
JB, EDB, LP and IF wrote the manuscript with contributions from all authors. All authors have read and approved the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

MAIN-TEXT FIGURE/TABLE LEGENDS

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

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

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

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

Figure 5. Comparative analysis of glideosome components. A. Table of presence/absence of glideosome proteins, distributed into functional groups. Glideosome components have been described mainly in *T. gondii* and *P. falciparum*. Proteins were searched for in both *Porospora* genomes as well as in a selection of representative species. Green indicates the presence, while white indicates the absence of a protein. Light green refers to the cases where one-to-one orthologous relationships have not been conclusively identified in *C. velia* and *V. brassicaformis*,

but related protein expansions have been observed[108]. All *P. cf. gigantea* orthologous proteins are detailed in Table S8. **B.** Schematic comparison of the canonical model of the glideosome and the elements found in *P. cf. gigantea* A and B. Missing proteins are shown in dotted line.

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

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

STAR METHODS

RESOURCE AVAILABILITY

Lead contact

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

Materials availability

This study did not generate new unique reagents.

Data and code availability

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

EXPERIMENTAL MODEL AND SUBJECT DETAILS

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

METHOD DETAILS

Electronic microscopy

For the Scanning Electron Microscopy (SEM) studies, isolated trophozoites and cysts, hosts intestines and rectal ampullas opened along their longitudinal axis, were washed as indicated above before fixation in 2.5% (v/v) glutaraldehyde in 0.1M sodium cacodylate (pH 7.2), at 4◦C, for 6 to 12

hours. After two washing steps in 0.1M sodium cacodylate (pH 7.2), biological specimens were transferred to Microporous Specimen Capsules (30µm porosity, 12mm diameter, 11mm high, ref #70187-20, Electron Microscopy Science) and dehydrated in a graded series of ethanol in ddH_2O (50, 70, 90, and 100%). Biological specimens were then critical point-dried in liquid CO_2 (Emitech K850, Quorum Technologies) in the Microporous Specimen Capsules, then transferred to adhesive carbon coated holders, and coated with 20nm of gold (JEOL Fine Coater JFC-1200). The biological specimens were then examined with a Hitachi Scanning Electron Microscope SU3500 Premium.
For the Transmission Electron Microscopy (TEM) studies, the samples were fixed for 2h in 0.2M sodium cacocylate buffer with 4% glutaraldehyde, 0.25M sucrose in 0.2M sodium cacodylate buffer pH 7.4. Cells were then washed three times in sodium cacodylate buffer containing decreasing concentrations of sucrose (0.25M, 0.12M, 0M) for 15 min each time, followed by postfixation for 1h at 4 °C in 2% osmium tetroxide in 0.1M sodium cacodylate buffer. After three rinses in 0.2M sodium cacodylate buffer, samples were dehydrated by successive transfers through a graded ethanol series (25%, 50%, 70%, 90%, 3 × 100%), then embedded in Spurr's resin. Sections were cut using a diamond knife on a Leica Ultracut UCT ultramicrotome (Leica, Wetzlar, Germany) and after staining with saturated uranyl acetate for 15 min and Reynolds' lead citrate for 3min, grids were examined with a Jeol 1400 transmission electron microscope (Jeol, Tokyo, Japan).

DNA/RNA isolations

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

DNA/RNA sequencing and assembly

The gDNA extracted for the 4 biological samples (JS-470, JS-482, JS-488 and JS-489) were sequenced individually by using Illumina NextSeq technology (2*151bp; NextSeq 500 Mid Output Kit v2; Institut du Cerveau et de la Moelle Epinière - CHU Pitié-Salpêtrière - Paris). We obtained 2*50 M to 2*70 M reads which were checked using FastQC[110] (version 0.11.5). Reads were cleaned by using Trim Galore[111] (version 0.4.4) removing remnant Nextera adaptors, clipping 15 bp in 5'-end and 1 bp in 3'-end and trimming low-quality ends (phred score < 30). The assembly was carried out by using SPAdes[112] (version 3.9.1; options: careful mode, automatic k-mers) with the pooled libraries (Figure 2.A).
The RNA was extracted from both samples (JS-555 and JS-575c) and treated with RNAse-free DNase. Library preparations (Institut du Cerveau et de la Moelle – CHU Pitié Salpétrière - Paris) were realized following manufacturer's recommendations (SMART-Seq v4 Ultra Low Input RNA Kit from Takara). Then, final samples were sequenced on NextSeq 500 Illumina device with

MidOutPut cartridge to generate a total of 2*87 M reads of 75bp. The read quality was checked by using FastQC[110] and cleaned by using Trim Galore[111] to remove remnant Nextera adaptors, clipping 15 bp in 5'-end and 1 bp in 3'-end and trimming low-quality ends (phred score < 30). The sequence reads of both samples were merged into one library which was assembled using Trinity[113,114].
All genomic contigs longer than 1kb were analyzed by using a principal component analysis (PCA) based on their 5-mer composition, which allowed classifying them into 6 groups by using a hierarchical clustering method (HCA) based on the Ward criterion (Figure 2.B).
For all contigs, the putative protein coding genes were then predicted by using Augustus[115] (version 3.3) and the only gene model available for an Apicomplexa: *T. gondii*. All the predicted proteins were thus compared with the NCBI non-redondant protein database by using BLAST[116]. The analysis of the taxonomic groups associated to the corresponding best hits, enabled us to identify five clusters as putative bacterial contaminants whereas the sixth cluster which included 1745 contigs (18.0Mb), was identified as organisms closely related to Apicomplexa, referred as "apicomplexa" cluster later (Figure 2.B).

Identification of genomes A and B

The contigs of the "apicomplexa" cluster were splitted into genomes A and B by using the difference of coverage observed for each of the 4 gDNA libraries (figure 2.C). Each gDNA library (JS-470, JS-482, JS-488 and JS-489) was individually mapped on the contigs by using Bowtie2[117] and the coverages' medians were calculated for each contig and each library by using the Samtools[118] and the Bedtools[119] libraries. This coverage information was processed with a principal component analysis and a k-means algorithm which allowed classifying the contigs into 2 clusters. Then, a linear discriminant model was trained with the coverage information and the result of this first classification before to be applied to all the contigs in order to improve the classification. The linear discriminant method (training and classification) was iteratively repeated 3 times until convergence. A similar analysis was carried out with 1kb non-overlapping windows (instead of full length contigs) to identify some putative hybrid contigs. Then, contigs classified to different genomes depending on the windows were divided into sub-contigs which were re-assigned to their respective genomes.
The nucleic divergence between genome A and genome B was estimated from the alignment of contigs built with Mummer3.0[120]. All alignments of the syntenic regions were parsed to compute the divergence by using a home-made script. Assembly metrics were assessed by using respectively QUAST[121] (version 5.0).

Prediction and annotation

All *de novo* assembled transcripts were aligned against the "apicomplexa" cluster contigs with GMAP[122] within the PASA program[123]. Then, two *ab initio* gene prediction tools, SNAP[124] (version 2017-11-15) and Augustus[115] were trained using a subset of the PASA transcriptome assemblies. A specific gene model was trained with Augustus, including meta-parameter optimization and construction of hints for introns (allowing small introns length >10bp) using our "apicomplexan cluster" repeat-masked genome assembly as reference (RepeatMasker[125], version 4.0.8). Gene predictions were then performed allowing the prediction of alternative transcripts and noncanonical intron bounds. An alternative model was also trained with SNAP (default protocol) and used for gene predictions. The Augustus and SNAP outputs having sometimes predicted genes slightly differently, the predictions were then parsed by a home-made script in order to keep, for each prediction made, as many alternative genes and transcripts as possible. The completeness of the gene prediction was assessed by using BUSCO[29] (version 4.0.6).
The predicted proteins, have been automatically annotated by using i) the best hit of a BLASTP search against VEupathdb (version 2019-20-01), ii) the results of KoFamScam against the KEGG pathway database[126] (version 2019-05-11) and iii) the signature domains obtained with Interproscan[127] (version 5.39-77.0).

The ortholog groups were identified by using orthoMCL[128] (default parameters, version 2.0.9) applied to the proteome of a selection of representative organisms available on VEuPathDB (Table S1).
The divergence time of genome A and genome B was estimated from the divergence time of *P. falciparum* and *P. reichenowi* as estimated in TimeTree[30]. Then, the coding sequences of the orthologous groups/quartets including one and only one gene for the genome A, the genome B, *P. falciparum* and *P. reichenowi* were aligned by using MacSE[129]. For each alignment, the number of synonymous substitutions per site (dS) between genomes A/B and between *P. falciparum/reichenowi* were computed with the maximum likelihood method of Yang and Nielsen (2000)[130] implemented in PAML4[131].
The Infernal software[132] (version 1.3.3) and the Rfam database[133] (version 14.2) were used together to search for transfer RNAs, spliceosomal RNAs and ribosomal RNAs. The snoReport software[134] (version 2) was used to search C/D and H/ACA small nucleolar RNAs.

Removal of contaminant sequences

Host. All the "apicomplexa" cluster contigs were screened by using the short reads available for the sequencing project of the genome of the closely related *Homarus americanus* species (PRJNA486050) in order to identify host contaminants. This dataset was supposed to be free of sequences from apicomplexan species, since it has been obtained from DNA extracted from the non-intestinal tissues (the tail, the leg or the pleiopod appendices). The mapping was carried out with Bowtie2[135] and the coverages were calculated by using the Samtools[118] library. The contigs thus identified, that were covered over more than 60% of their length by *Homarus's* short reads, were considered as host contaminants and were removed.
Prokaryote/Fungi. In parallel, predicted genes on the "apicomplexa" cluster contigs were also deeply analyzed for contamination by bacteria and fungi sequences. On scaffolds of this cluster containing at least one predicted protein, a BLASTP against NCBI NR database was launched. For contigs displaying a hit with a e-value lower than 1e-30 and covering more than 30% of query length with a Prokaryote/Fungi and more than 30% of length covered by Prokaryotes/Fungi hits on these contigs, an additional BLASTN against NCBI NR/NT was performed. For the remaining scaffolds without predicted protein, a BLASTN vs nr/nt was directly performed. At the end of this procedure, the contigs with Prokaryotes/Fungi hits covering more of 70% of length were labeled as contaminants and were removed from the genome assembly.

Experimental reconstruction of 18S/28S loci

First, a partial SSU rDNA locus was amplified by using JS-470 gDNA (including only the genome A) as DNA template and WL1 and EukP3 primers (Table S7) in a classical PCR reaction; the amplified bands were systematically cloned and sequenced as previously described[39]. This partial SSU rDNA sequence was further extended experimentally in the 3' direction still using JS-470 gDNA as DNA template and novel primers designed or re-designed based on the molecular data published for *Cephaloidophora cf. communis* and *Heliospora cf. longissima*[38] (Figure S2.A). The resulting sequence (>4 kb) was then used as anchor to reconstruct a complete ribosomal locus with the program iSeGWalker[136]. By clustering reads from JS-470 on this anchor, a 7322-bp theoretical sequence that corresponded to [partial 28S – 18S – ITS1 – 5.8S – ITS2 – partial 28S] including a perfect 1352-bp overlap between the 5' and 3' [partial 28S] segments was obtained. This allowed reconstructing a complete ribosomal locus [18S – ITS1 – 5.8S – ITS2 –28S] of 5977-pb for genome A, that was then experimentally validated by PCR amplification, cloning and sequencing (Figure S2.B). A similar clustering approach, using all reads for JS-482, JS-488 and JS-489, allowed reconstructing *in silico* the complete ribosomal locus for genome B, which was of same length, but with 30 polymorphic positions compared to the complete ribosomal locus for genome A (Figure S2.C). 60% of the complete ribosomal locus for genome B was then confirmed experimentally by PCR amplification, cloning and sequencing (positions 1187 to 4220). This second clustering also

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

Phylogeny

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

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

Expert annotation for glideosome proteins

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

For each orthogroup containing at least one of the reference proteins, the list of proteins was extracted and the protein sequences were recovered, as well as their respective coding nucleic sequences for both *P. cf. gigantea* genomes. A BLASTP was performed for extracted proteins against the proteomes of *P. cf. gigantea*, as well as a BLASTP of the candidate proteins for each *P. cf. gigantea* genome against the 25 species reference proteomes. A BLASTN was performed against NCBI NR for the coding sequences of the candidate proteins of both *P. cf. gigantea* genomes. The sequences thus collected for each described protein were aligned with mafft[142]. Maximum likelihood molecular phylogeny was performed on each alignment using RAxML software[139]. Analyses were performed using the LG model; bootstraps were estimated from 1,000 replicates.

Annotations of the conserved molecular domains were searched for in the automatic annotation and structure analyzed with SMART[143]. For each protein, all the performed analyses were then manually examined to validate the candidate proteins within the proteomes of the two *P. cf. gigantea* genomes. Presence/absence table of glideosome proteins was visualized using R using the tidyverse package[144]. The identification of TRAP-like proteins was done by searching for the TSP1 molecular domain (IPR000884) within the two *P. cf. gigantea* genomes. The structure of each candidate protein was then carefully studied. If necessary, partially predicted proteins were re-edited with Genewise[145]. Schematic representation of TRAP-like proteins was done using BioRender (biorender.com).

LEGEND FOR SUPPLEMENTAL VIDEO

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

REFERENCES

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

Freeman, M.A., and Keeling, P.J. (2019). Multiple Independent Origins of Apicomplexan-Like Parasites. Current Biology *29*, 2936-2941.e5. 10.1016/j.cub.2019.07.019.
12. Mathur, V., Kwong, W.K., Husnik, F., Irwin, N.A.T., Kristmundsson, Á., Gestal, C., Freeman, M., and Keeling, P.J. (2021). Phylogenomics Identifies a New Major Subgroup of Apicomplexans, Marosporida *class nov.* , with Extreme Apicoplast Genome Reduction. Genome Biology and Evolution *13*, evaa244. 10.1093/gbe/evaa244.
13. Salomaki, E.D., Terpis, K.X., Rueckert, S., Kotyk, M., Varadínová, Z.K., Čepička, I., Lane, C.E., and Kolisko, M. (2021). Gregarine single-cell transcriptomics reveals differential mitochondrial remodeling and adaptation in apicomplexans. BMC Biol *19*, 77. 10.1186/s12915-021-01007-2.
14. Van Beneden (1869). Sur une nouvelle espèce de Grégarine désignée sous le nom de *Gregarina gigantea*. Bulletins de l'Académie Royale de Belgique *28*, 444–456.
15. Schneider, A. (1875). Contribution à l' histoire des Grégarines des Invertébrés de Paris et de Roscoff. Arch Zool Exp Gen *4*, 493–604.
16. De Bauchamp, P. (1910). Sur une grégarine nouvelle du genre *Porospora*. C R Acad Sci Paris *151*, 997–999.
17. Hatt, P. (1931). L'évolution des Porosporides chez les mollusques. Archives de zoologie expérimentale et générale *72*, 341–415.
18. Desportes, I., I., and Theodorides, J. (1965). Ultrastructure of the Gymnospore of *Porospora* (Eugregarina, Porosporidae). C R Acad Sci Paris *260*, 1761–2.
19. Schrével, J., and Desportes, I. (2015). Gregarines. In Encyclopedia of Parasitology, H. Mehlhorn, ed. (Springer Berlin Heidelberg), pp. 1–47. 10.1007/978-3-642-27769-6_1335-2.
20. King, C., and Sleep, J. (2005). Modelling the mechanism of gregarine gliding using bead translocation. J Eukaryotic Microbiology *52*, 7S-27S. 10.1111/j.1550-7408.2005.05202003_1_41.x.
21. Russell, D.G. (1983). Host cell invasion by Apicomplexa: an expression of the parasite's contractile system? Parasitology *87*, 199–209. 10.1017/S0031182000052562.
22. King, C.A. (1988). Cell motility of sporozoan protozoa. Parasitology Today *4*, 315–319. 10.1016/0169-4758(88)90113-5.
23. Sibley, L.D., Håkansson, S., and Carruthers, V.B. (1998). Gliding motility: An efficient mechanism for cell penetration. Current Biology *8*, R12–R14. 10.1016/S0960-9822(98)70008-9.
24. Opitz, C., and Soldati, D. (2002). 'The glideosome': a dynamic complex powering gliding motion and host cell invasion by *Toxoplasma gondii*: Mechanism of host cell invasion by the Apicomplexa. Molecular Microbiology *45*, 597–604. 10.1046/j.1365-2958.2002.03056.x.
25. Keeley, A., and Soldati, D. (2004). The glideosome: a molecular machine powering motility and host-cell invasion by Apicomplexa. Trends in Cell Biology *14*, 528–532. 10.1016/j.tcb.2004.08.002.
26. Frénal, K., Dubremetz, J.-F., Lebrun, M., and Soldati-Favre, D. (2017). Gliding motility powers invasion and egress in Apicomplexa. Nature Reviews Microbiology *15*, 645–660. 10.1038/nrmicro.2017.86.
27. Valigurová, A., Vaškovicová, N., Musilová, N., and Schrével, J. (2013). The enigma of eugregarine epicytic folds: where gliding motility originates? Front Zool *10*, 57. 10.1186/1742-9994-10-57.
28. Léger, L., and Duboscq, O. (1909). Etude sur la sexualité des Grégarines. Arch. Protistenk. *17*, 19–134.
29. Seppey, M., Manni, M., and Zdobnov, E.M. (2019). BUSCO: Assessing Genome Assembly and Annotation Completeness. In Gene Prediction Methods in Molecular Biology. (Springer New York), pp. 227–245. 10.1007/978-1-4939-9173-0_14.
30. Kumar, S., Stecher, G., Suleski, M., and Hedges, S.B. (2017). TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Molecular Biology and Evolution *34*, 1812–1819. 10.1093/molbev/msx116.
31. Cornillot, E., Hadj-Kaddour, K., Dassouli, A., Noel, B., Ranwez, V., Vacherie, B., Augagneur, Y., Brès, V., Duclos, A., Randazzo, S., et al. (2012). Sequencing of the smallest

Apicomplexan genome from the human pathogen *Babesia microti*. Nucleic Acids Research *40*, 9102–9114. 10.1093/nar/gks700.
32. Neafsey, D.E., Hartl, D.L., and Berriman, M. (2005). Evolution of Noncoding and Silent Coding Sites in the *Plasmodium falciparum* and *Plasmodium reichenowi* Genomes. Molecular Biology and Evolution *22*, 1621–1626. 10.1093/molbev/msi154.
33. Reid, A.J., Vermont, S.J., Cotton, J.A., Harris, D., Hill-Cawthorne, G.A., Könen-Waisman, S., Latham, S.M., Mourier, T., Norton, R., Quail, M.A., et al. (2012). Comparative Genomics of the Apicomplexan Parasites *Toxoplasma gondii* and *Neospora caninum*: Coccidia Differing in Host Range and Transmission Strategy. PLoS Pathogens *8*, e1002567. 10.1371/journal.ppat.1002567.
34. Ricklefs, R.E., and Outlaw, D.C. (2010). A Molecular Clock for Malaria Parasites. Science *329*, 226–229. 10.1126/science.1188954.
35. Hayakawa, T., Tachibana, S.-I., Hikosaka, K., Arisue, N., Matsui, A., Horii, T., and Tanabe, K. (2012). Age of the last common ancestor of extant *Plasmodium* parasite lineages. Gene *502*, 36–39. 10.1016/j.gene.2012.04.037.
36. Crandall, K.A., Pérez-Losada, M., and Porter, M.L. (2009). Crabs, shrimps, and lobsters (Decapoda). In The Timetree Of Life (Oxford University Press), p. 551.
37. Bracken-Grissom, H.D., Ahyong, S.T., Wilkinson, R.D., Feldmann, R.M., Schweitzer, C.E., Breinholt, J.W., Bendall, M., Palero, F., Chan, T.-Y., Felder, D.L., et al. (2014). The Emergence of Lobsters: Phylogenetic Relationships, Morphological Evolution and Divergence Time Comparisons of an Ancient Group (Decapoda: Achelata, Astacidea, Glypheidea, Polychelida). Systematic Biology *63*, 457–479. 10.1093/sysbio/syu008.
38. Simdyanov, T.G., Diakin, A.Y., and Aleoshin, V.V. (2015). Ultrastructure and 28S rDNA Phylogeny of Two Gregarines: *Cephaloidophora cf. communis* and *Heliospora cf. longissima* with Remarks on Gregarine Morphology and Phylogenetic Analysis. Acta Protozoologica *54*, 241–262. 10.4467/16890027AP.15.020.3217.
39. Schrével, J., Valigurová, A., Prensier, G., Chambouvet, A., Florent, I., and Guillou, L. (2016). Ultrastructure of *Selenidium pendula*, the Type Species of Archigregarines, and Phylogenetic Relations to Other Marine Apicomplexa. Protist *167*, 339–368. 10.1016/j.protis.2016.06.001.
40. Rueckert, S., Simdyanov, T.G., Aleoshin, V.V., and Leander, B.S. (2011). Identification of a Divergent Environmental DNA Sequence Clade Using the Phylogeny of Gregarine Parasites (Apicomplexa) from Crustacean Hosts. PLoS ONE *6*, e18163. 10.1371/journal.pone.0018163.
41. Mulec, J., and Summers Engel, A. (2019). Karst spring microbial mat microeukaryotic diversity differs across an oxygen-sulphide ecocline and reveals potential for novel taxa discovery. AC *48*. 10.3986/ac.v48i1.4949.
42. Skillman, K.M., Diraviyam, K., Khan, A., Tang, K., Sept, D., and Sibley, L.D. (2011). Evolutionarily Divergent, Unstable Filamentous Actin Is Essential for Gliding Motility in Apicomplexan Parasites. PLoS Pathog *7*, e1002280. 10.1371/journal.ppat.1002280.
43. Plattner, F., Yarovinsky, F., Romero, S., Didry, D., Carlier, M.-F., Sher, A., and Soldati-Favre, D. (2008). *Toxoplasma* Profilin Is Essential for Host Cell Invasion and TLR11-Dependent Induction of an Interleukin-12 Response. Cell Host & Microbe *3*, 77–87. 10.1016/j.chom.2008.01.001.
44. Pino, P., Sebastian, S., Kim, E.A., Bush, E., Brochet, M., Volkmann, K., Kozlowski, E., Llinás, M., Billker, O., and Soldati-Favre, D. (2012). A Tetracycline-Repressible Transactivator System to Study Essential Genes in Malaria Parasites. Cell Host & Microbe *12*, 824–834. 10.1016/j.chom.2012.10.016.
45. Skillman, K.M., Daher, W., Ma, C.I., Soldati-Favre, D., and Sibley, L.D. (2012). *Toxoplasma gondii* Profilin Acts Primarily To Sequester G-Actin While Formins Efficiently Nucleate Actin Filament Formation *in Vitro*. Biochemistry *51*, 2486–2495. 10.1021/bi201704y.
46. Mehta, S., and Sibley, L.D. (2011). Actin depolymerizing factor controls actin turnover and gliding motility in *Toxoplasma gondii*. MBoC *22*, 1290–1299. 10.1091/mbc.e10-12-0939.
47. Tosetti, N., Pacheco, N.D.S., Soldati-Favre, D., and Jacot, D. (2019). Three F-actin assembly centers regulate organelle inheritance, cell-cell communication and motility in. 32.
48. Daher, W., Plattner, F., Carlier, M.-F., and Soldati-Favre, D. (2010). Concerted Action of Two

Formins in Gliding Motility and Host Cell Invasion by *Toxoplasma gondii*. PLoS Pathog *6*, e1001132. 10.1371/journal.ppat.1001132.
49. Baum, J., Tonkin, C.J., Paul, A.S., Rug, M., Smith, B.J., Gould, S.B., Richard, D., Pollard, T.D., and Cowman, A.F. (2008). A Malaria Parasite Formin Regulates Actin Polymerization and Localizes to the Parasite-Erythrocyte Moving Junction during Invasion. Cell Host & Microbe *3*, 188–198. 10.1016/j.chom.2008.02.006.
50. Ganter, M., SchÃ¼ler, H., and Matuschewski, K. (2009). Vital role for the *Plasmodium* actin capping protein (CP) beta-subunit in motility of malaria sporozoites: Plasmodium actin capping protein. Molecular Microbiology *74*, 1356–1367. 10.1111/j.1365-2958.2009.06828.x.
51. Drewry, L.L., and Sibley, L.D. (2015). *Toxoplasma* Actin Is Required for Efficient Host Cell Invasion. mBio *6*, e00557-15. 10.1128/mBio.00557-15.
52. Egarter, S., Andenmatten, N., Jackson, A.J., Whitelaw, J.A., Pall, G., Black, J.A., Ferguson, D.J.P., Tardieux, I., Mogilner, A., and Meissner, M. (2014). The *Toxoplasma* Acto-MyoA Motor Complex Is Important but Not Essential for Gliding Motility and Host Cell Invasion. PLoS ONE *9*, e91819. 10.1371/journal.pone.0091819.
53. Whitelaw, J.A., Latorre-Barragan, F., Gras, S., Pall, G.S., Leung, J.M., Heaslip, A., Egarter, S., Andenmatten, N., Nelson, S.R., Warshaw, D.M., et al. (2017). Surface attachment, promoted by the actomyosin system of *Toxoplasma gondii* is important for efficient gliding motility and invasion. BMC Biology *15*. 10.1186/s12915-016-0343-5.
54. Meissner, M., Schluter, D., and Soldati, D. (2002). Role of *Toxoplasma gondii* Myosin A in Powering Parasite Gliding and Host Cell Invasion. Science *298*, 837–840. 10.1126/science.1074553.
55. Frénal, K., Marq, J.-B., Jacot, D., Polonais, V., and Soldati-Favre, D. (2014). Plasticity between MyoC- and MyoA-Glideosomes: An Example of Functional Compensation in *Toxoplasma gondii* Invasion. PLoS Pathogens *10*, e1004504. 10.1371/journal.ppat.1004504.
56. Siden-Kiamos, I., Ganter, M., Kunze, A., Hliscs, M., Steinbüchel, M., Mendoza, J., Sinden, R.E., Louis, C., and Matuschewski, K. (2011). Stage-specific depletion of myosin A supports an essential role in motility of malarial ookinetes: Promoter swap to study *Plasmodium* myosin A function. Cellular Microbiology *13*, 1996–2006. 10.1111/j.1462-5822.2011.01686.x.
57. Herm-Gotz, A. (2002). *Toxoplasma gondii* myosin A and its light chain: a fast, single-headed, plus-end-directed motor. The EMBO Journal *21*, 2149–2158. 10.1093/emboj/21.9.2149.
58. Bergman, L.W. (2003). Myosin A tail domain interacting protein (MTIP) localizes to the inner membrane complex of *Plasmodium* sporozoites. Journal of Cell Science *116*, 39–49. 10.1242/jcs.00194.
59. Gaskins, E., Gilk, S., DeVore, N., Mann, T., Ward, G., and Beckers, C. (2004). Identification of the membrane receptor of a class XIV myosin in *Toxoplasma gondii* . Journal of Cell Biology *165*, 383–393. 10.1083/jcb.200311137.
60. Baum, J., Papenfuss, A.T., Baum, B., Speed, T.P., and Cowman, A.F. (2006). Regulation of apicomplexan actin-based motility. Nature Reviews Microbiology *4*, 621–628. 10.1038/nrmicro1465.
61. Frénal, K., Polonais, V., Marq, J.-B., Stratmann, R., Limenitakis, J., and Soldati-Favre, D. (2010). Functional Dissection of the Apicomplexan Glideosome Molecular Architecture. Cell Host & Microbe *8*, 343–357. 10.1016/j.chom.2010.09.002.
62. Tardieux, I., and Baum, J. (2016). Reassessing the mechanics of parasite motility and host-cell invasion. Journal of Cell Biology *214*, 507–515. 10.1083/jcb.201605100.
63. Bullen, H.E., Tonkin, C.J., O'Donnell, R.A., Tham, W.-H., Papenfuss, A.T., Gould, S., Cowman, A.F., Crabb, B.S., and Gilson, P.R. (2009). A Novel Family of Apicomplexan Glideosome-associated Proteins with an Inner Membrane-anchoring Role. J. Biol. Chem. *284*, 25353–25363. 10.1074/jbc.M109.036772.
64. Graindorge, A., Frénal, K., Jacot, D., Salamun, J., Marq, J.B., and Soldati-Favre, D. (2016). The Conoid Associated Motor MyoH Is Indispensable for *Toxoplasma gondii* Entry and Exit from Host Cells. PLOS Pathogens *12*, e1005388. 10.1371/journal.ppat.1005388.

65. Paing, M.M., and Tolia, N.H. (2014). Multimeric Assembly of Host-Pathogen Adhesion Complexes Involved in Apicomplexan Invasion. PLoS Pathog *10*, e1004120. 10.1371/journal.ppat.1004120.
66. Boucher, L.E., and Bosch, J. (2015). The apicomplexan glideosome and adhesins – Structures and function. Journal of Structural Biology *190*, 93–114. 10.1016/j.jsb.2015.02.008.
67. Jacot, D., Waller, R.F., Soldati-Favre, D., MacPherson, D.A., and MacRae, J.I. (2016). Apicomplexan Energy Metabolism: Carbon Source Promiscuity and the Quiescence Hyperbole. Trends in Parasitology *32*, 56–70. 10.1016/j.pt.2015.09.001.
68. Sultan, A.A., Thathy, V., Frevert, U., Robson, K.J.H., Crisanti, A., Nussenzweig, V., Nussenzweig, R.S., and Ménard, R. (1997). TRAP Is Necessary for Gliding Motility and Infectivity of *Plasmodium* Sporozoites. Cell *90*, 511–522. 10.1016/S0092-8674(00)80511-5.
69. Huynh, M.-H., and Carruthers, V.B. (2006). *Toxoplasma* MIC2 Is a Major Determinant of Invasion and Virulence. PLoS Pathog *2*, e84. 10.1371/journal.ppat.0020084.
70. Buguliskis, J.S., Brossier, F., Shuman, J., and Sibley, L.D. (2010). Rhomboid 4 (ROM4) Affects the Processing of Surface Adhesins and Facilitates Host Cell Invasion by *Toxoplasma gondii*. PLoS Pathog *6*. 10.1371/journal.ppat.1000858.
71. Shen, B., Buguliskis, J.S., Lee, T.D., and Sibley, L.D. (2014). Functional Analysis of Rhomboid Proteases during *Toxoplasma* Invasion. mBio *5*, e01795-14. 10.1128/mBio.01795-14.
72. Rugarabamu, G., Marq, J.-B., Guérin, A., Lebrun, M., and Soldati-Favre, D. (2015). Distinct contribution of *Toxoplasma gondii* rhomboid proteases 4 and 5 to micronemal protein protease 1 activity during invasion: ROM4 and ROM5 contribute to MPP1 activity. Molecular Microbiology *97*, 244–262. 10.1111/mmi.13021.
73. Kappe, S., Bruderer, T., Gantt, S., Fujioka, H., Nussenzweig, V., and Ménard, R. (1999). Conservation of a Gliding Motility and Cell Invasion Machinery in Apicomplexan Parasites. Journal of Cell Biology *147*, 937–944. 10.1083/jcb.147.5.937.
74. Morahan, B.J., Wang, L., and Coppel, R.L. (2009). No TRAP, no invasion. Trends in Parasitology *25*, 77–84. 10.1016/j.pt.2008.11.004.
75. Templeton, T.J., and Pain, A. (2016). Diversity of extracellular proteins during the transition from the 'proto-apicomplexan' alveolates to the apicomplexan obligate parasites. Parasitology *143*, 1–17. 10.1017/S0031182015001213.
76. Dessens, J.T., Beetsma, A.L., Dimopoulos, G., Wengelnik, K., Crisanti, A., Kafatos, F.C., and Sinden, R.E. (1999). CTRP is essential for mosquito infection by malaria ookinetes. EMBO J *18*, 6221–6227. 10.1093/emboj/18.22.6221.
77. Bargieri, D.Y. (2016). *Plasmodium* Merozoite TRAP Family Protein Is Essential for Vacuole Membrane Disruption and Gamete Egress from Erythrocytes. Cell Host Microbe *20*, 618–630. 10.1016/j.chom.2016.10.015.
78. Lacroix, C., and Ménard, R. (2008). TRAP-like protein of *Plasmodium* sporozoites: linking gliding motility to host-cell traversal. Trends in Parasitology *24*, 431–434. 10.1016/j.pt.2008.07.003.
79. Deng, M., Templeton, T.J., London, N.R., Bauer, C., Schroeder, A.A., and Abrahamsen, M.S. (2002). *Cryptosporidium parvum* Genes Containing Thrombospondin Type 1 Domains. IAI *70*, 6987–6995. 10.1128/IAI.70.12.6987-6995.2002.
80. Putignani, L., Possenti, A., Cherchi, S., Pozio, E., Crisanti, A., and Spano, F. (2008). The thrombospondin-related protein CpMIC1 (CpTSP8) belongs to the repertoire of micronemal proteins of *Cryptosporidium parvum*. Molecular and Biochemical Parasitology *157*, 98–101. 10.1016/j.molbiopara.2007.09.004.
81. Gaffar, F.R., Yatsuda, A.P., Franssen, F.F.J., and Vries, E. de (2004). A Babesia bovis merozoite protein with a domain architecture highly similar to the thrombospondin-related anonymous protein (TRAP) present in *Plasmodium* sporozoites. Molecular and Biochemical Parasitology *136*, 25–34. 10.1016/j.molbiopara.2004.02.006.
82. Zhou, J., Fukumoto, S., Jia, H., Yokoyama, N., Zhang, G., Fujisaki, K., Lin, J., and Xuan, X. (2006). Characterization of the *Babesia gibsoni* P18 as a homologue of thrombospondin related

adhesive protein. Molecular and Biochemical Parasitology *148*, 190–198. 10.1016/j.molbiopara.2006.03.015.
83. Yu, L., Liu, Q., Zhan, X., Huang, Y., Sun, Y., Nie, Z., Zhao, Y., An, X., Li, M., Wang, S., et al. (2018). Identification and molecular characterization of a novel *Babesia orientalis* thrombospondin-related anonymous protein (BoTRAP1). Parasites Vectors *11*, 667. 10.1186/s13071-018-3245-2.
84. Montenegro, V.N., Paoletta, M.S., Jaramillo Ortiz, J.M., Suarez, C.E., and Wilkowsky, S.E. (2020). Identification and characterization of a *Babesia bigemina* thrombospondin-related superfamily member, TRAP-1: a novel antigen containing neutralizing epitopes involved in merozoite invasion. Parasites Vectors *13*, 602. 10.1186/s13071-020-04469-5.
85. Lovett, J. (2000). Molecular characterization of a thrombospondin-related anonymous protein homologue in *Neospora caninum*. Molecular and Biochemical Parasitology *107*, 33–43. 10.1016/S0166-6851(99)00228-5.
86. Clarke, L.E., Tomley, F.M., Wisher, M.H., Foulds, I.J., and Boursnell, M.E. (1990). Regions of an *Eimeria tenella* antigen contain sequences which are conserved in circumsporozoite proteins from *Plasmodium* spp. and which are related to the thrombospondin gene family. Mol Biochem Parasitol *41*, 269–279. 10.1016/0166-6851(90)90190-w.
87. Witcombe, D.M., Belli, S.I., Wallach, M.G., and Smith, N.C. (2003). Molecular characterisation of EmTFP250: a novel member of the TRAP protein family in *Eimeria maxima*. International Journal for Parasitology *33*, 691–702. 10.1016/S0020-7519(03)00086-9.
88. Bichet, M., Joly, C., Hadj Henni, A., Guilbert, T., Xémard, M., Tafani, V., Lagal, V., Charras, G., and Tardieux, I. (2014). The toxoplasma-host cell junction is anchored to the cell cortex to sustain parasite invasive force. BMC Biol *12*, 773. 10.1186/s12915-014-0108-y.
89. Portes, J., Barrias, E., Travassos, R., Attias, M., and de Souza, W. (2020). *Toxoplasma gondii* Mechanisms of Entry Into Host Cells. Front. Cell. Infect. Microbiol. *10*, 294. 10.3389/fcimb.2020.00294.
90. Yang, A.S.P., Lopaticki, S., O'Neill, M.T., Erickson, S.M., Douglas, D.N., Kneteman, N.M., and Boddey, J.A. (2017). AMA1 and MAEBL are important for *Plasmodium falciparum* sporozoite infection of the liver. Cellular Microbiology *19*, e12745. 10.1111/cmi.12745.
91. O'Hara, S.P., and Chen, X.-M. (2011). The cell biology of *Cryptosporidium* infection. Microbes Infect *13*, 721–730. 10.1016/j.micinf.2011.03.008.
92. Singh, P., Mirdha, B.R., Srinivasan, A., Rukmangadachar, L.A., Singh, S., Sharma, P., Hariprasad, G., Gururao, H., Gururao, H., and Luthra, K. (2015). Identification of invasion proteins of *Cryptosporidium parvum*. World J Microbiol Biotechnol *31*, 1923–1934. 10.1007/s11274-015-1936-9.
93. Lourido, S., and Moreno, S.N.J. (2015). The calcium signaling toolkit of the Apicomplexan parasites *Toxoplasma gondii* and *Plasmodium* spp. Cell Calcium *57*, 186–193. 10.1016/j.ceca.2014.12.010.
94. Ghartey-Kwansah, G., Yin, Q., Li, Z., Gumpper, K., Sun, Y., Yang, R., Wang, D., Jones, O., Zhou, X., Wang, L., et al. (2020). Calcium-dependent Protein Kinases in Malaria Parasite Development and Infection. Cell Transplant *29*, 096368971988488. 10.1177/0963689719884888.
95. Bullen, H.E., Jia, Y., Yamaryo-Botté, Y., Bisio, H., Zhang, O., Jemelin, N.K., Marq, J.-B., Carruthers, V., Botté, C.Y., and Soldati-Favre, D. (2016). Phosphatidic Acid-Mediated Signaling Regulates Microneme Secretion in *Toxoplasma*. Cell Host & Microbe *19*, 349–360. 10.1016/j.chom.2016.02.006.
96. Darvill, N., Dubois, D.J., Rouse, S.L., Hammoudi, P.-M., Blake, T., Benjamin, S., Liu, B., Soldati-Favre, D., and Matthews, S. (2018). Structural Basis of Phosphatidic Acid Sensing by APH in Apicomplexan Parasites. Structure *26*, 1059–1071. 10.1016/j.str.2018.05.001.
97. Farrell, A., Thirugnanam, S., Lorestani, A., Dvorin, J.D., Eidell, K.P., Ferguson, D.J.P., Anderson-White, B.R., Duraisingh, M.T., Marth, G.T., and Gubbels, M.-J. (2012). A DOC2 Protein Identified by Mutational Profiling Is Essential for Apicomplexan Parasite Exocytosis. Science *335*, 218–221. 10.1126/science.1210829.

98. Heaslip, A.T., Nishi, M., Stein, B., and Hu, K. (2011). The Motility of a Human Parasite, *Toxoplasma gondii*, Is Regulated by a Novel Lysine Methyltransferase. PLoS Pathog *7*, e1002201. 10.1371/journal.ppat.1002201.
99. Guo, Y., Tang, K., Rowe, L.A., Li, N., Roellig, D.M., Knipe, K., Frace, M., Yang, C., Feng, Y., and Xiao, L. (2015). Comparative genomic analysis reveals occurrence of genetic recombination in virulent *Cryptosporidium hominis* subtypes and telomeric gene duplications in *Cryptosporidium parvum*. BMC Genomics *16*. 10.1186/s12864-015-1517-1.
100. Gras, S., Jackson, A., Woods, S., Pall, G., Whitelaw, J., Leung, J.M., Ward, G.E., Roberts, C.W., and Meissner, M. (2017). Parasites lacking the micronemal protein MIC2 are deficient in surface attachment and host cell egress, but remain virulent in vivo. Wellcome Open Res *2*, 32. 10.12688/wellcomeopenres.11594.2.
101. Harding, C.R., Gow, M., Kang, J.H., Shortt, E., Manalis, S.R., Meissner, M., and Lourido, S. (2019). Alveolar proteins stabilize cortical microtubules in *Toxoplasma gondii*. Nature Communications *10*. 10.1038/s41467-019-08318-7.
102. Rompikuntal, P.K., Foe, I.T., Deng, B., Bogyo, M., and Ward, G.E. (2020). Blocking palmitoylation of *Toxoplasma gondii* myosin light chain 1 disrupts glideosome composition but has little impact on parasite motility (Microbiology) 10.1101/2020.08.13.250399.
103. Valigurová, A., Vaškovicová, N., Diakin, A., Paskerova, G.G., Simdyanov, T.G., and Kováčiková, M. (2017). Motility in blastogregarines (Apicomplexa): Native and drug-induced organisation of *Siedleckia nematoides* cytoskeletal elements. PLOS ONE *12*, e0179709. 10.1371/journal.pone.0179709.
104. Heintzelman, M.B. (2004). Actin and myosin in *Gregarina polymorpha*. Cell Motility and the Cytoskeleton *58*, 83–95. 10.1002/cm.10178.
105. Heintzelman, M.B., and Mateer, M.J. (2008). GpMyoF, a WD40 Repeat-Containing Myosin Associated with the Myonemes of *Gregarina polymorpha*. Journal of Parasitology *94*, 158–168. 10.1645/GE-1339.1.
106. Kováčiková, M., Simdyanov, T.G., Diakin, A., and Valigurová, A. (2017). Structures related to attachment and motility in the marine eugregarine *Cephaloidophora cf. communis* (Apicomplexa). European Journal of Protistology *59*, 1–13. 10.1016/j.ejop.2017.02.006.
107. Diakin, A., Wakeman, K.C., and Valigurová, A. (2017). Description of *Ganymedes yurii* sp. n. (Ganymedidae), a New Gregarine Species from the Antarctic Amphipod *Gondogeneia* sp. (Crustacea). Journal of Eukaryotic Microbiology *64*, 56–66. 10.1111/jeu.12336.
108. Woo, Y.H., Ansari, H., Otto, T.D., Klinger, C.M., Kolisko, M., Michálek, J., Saxena, A., Shanmugam, D., Tayyrov, A., Veluchamy, A., et al. (2015). Chromerid genomes reveal the evolutionary path from photosynthetic algae to obligate intracellular parasites. eLife *4*. 10.7554/eLife.06974.
109. Butler, M., Cockcroft, A., MacDiarmid, A., and Wahle, R. (2011). *Homarus gammarus*. The IUCN Red List of Threatened Species. e.T169955A69905303. 10.2305/IUCN.UK.2011-1.RLTS.T169955A69905303.en.
110. Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc.
111. Krueger, Felix (2015). Trim galore. A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files. https://github.com/FelixKrueger/TrimGalore.
112. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., et al. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology *19*, 455–477. 10.1089/cmb.2012.0021.
113. Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol *29*, 644–652. 10.1038/nbt.1883.
114. Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger,

M.B., Eccles, D., Li, B., Lieber, M., et al. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc *8*, 1494–1512. 10.1038/nprot.2013.084.
115. Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Research *34*, W435–W439. 10.1093/nar/gkl200.
116. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. Journal of Molecular Biology *215*, 403–410. 10.1016/S0022-2836(05)80360-2.
117. Langmead, B., Wilks, C., Antonescu, V., and Charles, R. (2019). Scaling read aligners to hundreds of threads on general-purpose processors. Bioinformatics *35*, 421–432. 10.1093/bioinformatics/bty648.
118. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. 10.1093/bioinformatics/btp352.
119. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. 10.1093/bioinformatics/btq033.
120. Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S.L. (2004). Versatile and open software for comparing large genomes. Genome Biol *5*, R12. 10.1186/gb-2004-5-2-r12.
121. Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics *29*, 1072–1075. 10.1093/bioinformatics/btt086.
122. Wu, T.D., and Watanabe, C.K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics *21*, 1859–1875. 10.1093/bioinformatics/bti310.
123. Haas, B.J. (2003). Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. Nucleic Acids Research *31*, 5654–5666. 10.1093/nar/gkg770.
124. Korf, I. (2004). Gene finding in novel genomes. BMC Bioinformatics *5*, 59. 10.1186/1471-2105-5-59.
125. Smit, AFA. and Hubley, R. (2015). RepeatMasker Open-4.0. http://www.repeatmasker.org.
126. Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S., and Ogata, H. (2020). KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. Bioinformatics *36*, 2251–2252. 10.1093/bioinformatics/btz859.
127. Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics *30*, 1236–1240. 10.1093/bioinformatics/btu031.
128. Li, L., Stoeckert, CJ., and Roos, D. (2003). OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Research *13*, 2178–2189. 10.1101/gr.1224503.
129. Ranwez, V., Harispe, S., Delsuc, F., and Douzery, E.J.P. (2011). MACSE: Multiple Alignment of Coding SEquences Accounting for Frameshifts and Stop Codons. PLoS ONE *6*, e22594. 10.1371/journal.pone.0022594.
130. Yang, Z., and Nielsen, R. (2000). Estimating Synonymous and Nonsynonymous Substitution Rates Under Realistic Evolutionary Models. Molecular Biology and Evolution *17*, 32–43. 10.1093/oxfordjournals.molbev.a026236.
131. Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. Molecular Biology and Evolution *24*, 1586–1591. 10.1093/molbev/msm088.
132. Nawrocki, E.P., and Eddy, S.R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics *29*, 2933–2935. 10.1093/bioinformatics/btt509.
133. Kalvari, I., Nawrocki, E.P., Argasinska, J., Quinones-Olvera, N., Finn, R.D., Bateman, A., and Petrov, A.I. (2018). Non-Coding RNA Analysis Using the Rfam Database. Current Protocols in Bioinformatics *62*, e51. 10.1002/cpbi.51.
134. de Araujo Oliveira, J.V., Costa, F., Backofen, R., Stadler, P.F., Machado Telles Walter, M.E.,

and Hertel, J. (2016). SnoReport 2.0: new features and a refined Support Vector Machine to improve snoRNA identification. BMC Bioinformatics *17*, 464. 10.1186/s12859-016-1345-6.
135. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357–359. 10.1038/nmeth.1923.
136. Karadjian, G., Hassanin, A., Saintpierre, B., Gembu Tungaluna, G.-C., Ariey, F., Ayala, F.J., Landau, I., and Duval, L. (2016). Highly rearranged mitochondrial genome in *Nycteria* parasites (Haemosporidia) from bats. Proceedings of the National Academy of Sciences *113*, 9834–9839. 10.1073/pnas.1610643113.
137. Castresana, J. (2000). Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. Molecular Biology and Evolution *17*, 540–552. 10.1093/oxfordjournals.molbev.a026334.
138. Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Molecular Biology and Evolution *35*, 1547–1549. 10.1093/molbev/msy096.
139. Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313. 10.1093/bioinformatics/btu033.
140. Ronquist, F., and Huelsenbeck, J.P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics *19*, 1572–1574. 10.1093/bioinformatics/btg180.
141. Rambaut (2018). FigTree. tree.bio.ed.ac.uk/software/figtree/.
142. Katoh, K., and Standley, D.M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution *30*, 772–780. 10.1093/molbev/mst010.
143. Letunic, I., Khedkar, S., and Bork, P. (2021). SMART: recent updates, new developments and status in 2020. Nucleic Acids Res *49*, D458–D460. 10.1093/nar/gkaa937.
144. Wickham, H. (2009). Ggplot2: elegant graphics for data analysis (Springer).
145. Birney, E. (2004). GeneWise and Genomewise. Genome Research *14*, 988–995. 10.1101/gr.1865504.

CellPress

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Bacterial and virus strains		
Biological samples		
Cyst sample from Lobster #7 (DNA)	This paper	JS-470
Cyst sample from Lobster #11 (DNA)	This paper	JS-482
Cyst sample from Lobster #12 (DNA)	This paper	JS-488
Cyst sample from Lobster #12 (DNA)	This paper	JS-489
Cyst sample from Lobster #26 (RNA)	This paper	JS-555
Cyst sample from Lobster #34 (RNA)	This paper	JS-575c
Chemicals, peptides, and recombinant proteins		
Penicillin/Streptomycin	Gibco, Life Technologies, USA	Ref. 15140-122
Gentamycin	Interchim, Montluçon, France	Ref. 31008a
Microporous Specimen Capsules	Electron Microscopy Science	Ref. 70187-20
Gold	JEOL Fine Coater	JFC-1200
Ethanol	VWR International	Ref. 20281-321
uranyl acetate	Electron Microscopy Science	Ref. 22400, lot 950724
Glutaraldehyde	Electron Microscopy Science	Ref. 16220
Sodium cacodylate	Electron Microscopy Science	Ref. 12300
Critical commercial assays		
SMART-Seq v4 Ultra Low Input RNA Kit	Takara	
Macherey Nagel basic RNA Isolation kit	Macherey Nagel	Ref. 740955.10
Macherey Nagel Nucleozol-based RNA Isolation kit	Macherey Nagel	Ref; 74040.200; Ref. 740406.10
NextSeq 500 Mid Output Kit		
Deposited data		
Raw and analyzed data	This paper	SRA: PRJNA734792

Experimental models: Cell lines		
Experimental models: Organisms/strains		
Porospora cf. gigantea A	This paper	N/A
Porospora cf. gigantea B	This paper	N/A
Oligonucleotides		
5′- GGAAGGCAGCAGGCGCGC - 3′	Schrevel et al., 2016	LWA1
5′- GACGGGCGGTGTGTAC - 3′	Lara et al., 2007	EukP3
5′- CCGCTAAGGAGTGTGTAACAAC - 3′	Simdyanov et al., 2015	28d5
5′- ACTCCTYRGTCCGTGTTTCA - 3′	Simdyanov et al., 2015	28r3.2
5′- TACTTGTYBRCTATCG - 3′	Simdyanov et al., 2015	28r2
5′- GGTGGTGCATGGCCAAACTT - 3′	Modified from Simdyanov et al., 2015	d6new
5′- GCTAAGGAGTGTGTAACAAC - 3′	Modified from Simdyanov et al., 2015	28d5short
5′- TAATTTGCCGACTTCCCTCA - 3′	Modified from Simdyanov et al., 2015	28r7new
5′- ACATTCCTTGGGTTACCC - 3′	This study	PIF5F
5′- TAACGACCCGAAAATCGG - 3′	This study	PIF6F
5′- CATGCTAACACAAGGGGG - 3′	This study	PIF7F
5′- CCGACAGTTTAACTAAAACC - 3′	This study	PIF8F
5′- GAGATCATATCGACGCGG- 3′	This study	PIF9F
5′- CATCAGTGCGACGATACC - 3′	This study	PIF5R
5′- GTTTGAGAATCAGTCGAGG - 3′	This study	PIF6R
5′- CTTTCGACTTCCGACAGC - 3′	This study	PIF7R

5′- TTGTTTGCTATCGGTATAGG - 3′	This study	PIF8R
5′- AAATCTCAAGAGAGATGGAG- 3′	This study	PIF9R
5′- GCTAAGGATCGATAGGCC - 3′	This study	PIF10R
Recombinant DNA		
Software and algorithms		
FastQC v0.11.5	Andrews, 2010	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Trim Galore v0.4.4	Krueger, 2015	https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
SPAdes V3.9.1	Bankevich et al, 2012	https://github.com/ablab/spades
Trinity	Grabherr et al, 2011; Haas et al, 2013	https://github.com/trinityrnaseq/trinityrnaseq/wiki
Augustus v3.3	Stanke et al, 2004	http://bioinf.uni-greifswald.de/augustus/
BLAST	Altschul et al, 1990	https://blast.ncbi.nlm.nih.gov/Blast.cgi
Bowtie2	Langmead et al, 2018	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
Samtools	Li et al., 2009	http://samtools.sourceforge.net/
Bedtools	Quinla and Hall, 2010	https://bedtools.readthedocs.io/en/latest/
Mummer 3.0	Kurtz et al, 2004	http://mummer.sourceforge.net/
QUAST v5.0	Gurevich et al, 2013	http://bioinf.spbau.ru/quast
GMAP	Wu and Watanabe, 2005	http://research-pub.gene.com/gmap/
PASA program	Haas et al, 2003	https://github.com/PASApipeline/PASApipeline/wiki
SNAP version 2017-11-15	Korf, 2004	https://github.com/KorfLab/SNAP
RepeatMasker v4.0.8	Smit et al, 2015	https://www.repeatmasker.org/
BUSCO v4.0.6	Seppey et al, 2019	https://busco.ezlab.org/
OrthoMCL v2.0.9	Li et al, 2003	https://orthomcl.org/orthomcl/app
TimeTree	Kumar et al, 2017	http://www.timetree.org/
MacSE	Ranwez et al, 2011	https://bioweb.supagro.inra.fr/macse/
PAML 4	Yang, 2007	http://abacus.gene.ucl.ac.uk/software/paml.html
Infernal	Nawrocki et al, 2013	http://eddylab.org/infernal/

iSeGWalker	Karadjian, Hassanin et al. 2016	-
Gblocks v0.91b	Castresana, 2000	http://molevol.cmima.csic.es/castresana/Gblocks/Gblocks_documentati
MEGA X	Kumar et al, 2018	https://www.megasoftware.net/
RaxML v8.2.12	Stamatakis, 2014	https://cme.h-its.org/exelixis/web/software/raxml/
MrBayes v3.2.3	Ronquist et al, 2012	https://nbisweden.github.io/MrBayes/download.html
FigTree v1.4.4	Rambaut	http://tree.bio.ed.ac.uk/software/figtree/
Inkscape		www.inkscape.org
mafft	Katoh and Standley, 2013	https://mafft.cbrc.jp/alignment/server/
R tidyverse package	Wickham et al, 2009	https://www.tidyverse.org/
Genewise	Birney et al, 2004	https://www.ebi.ac.uk/Tools/psa/genewise/
Biorender		biorender.com
Other		

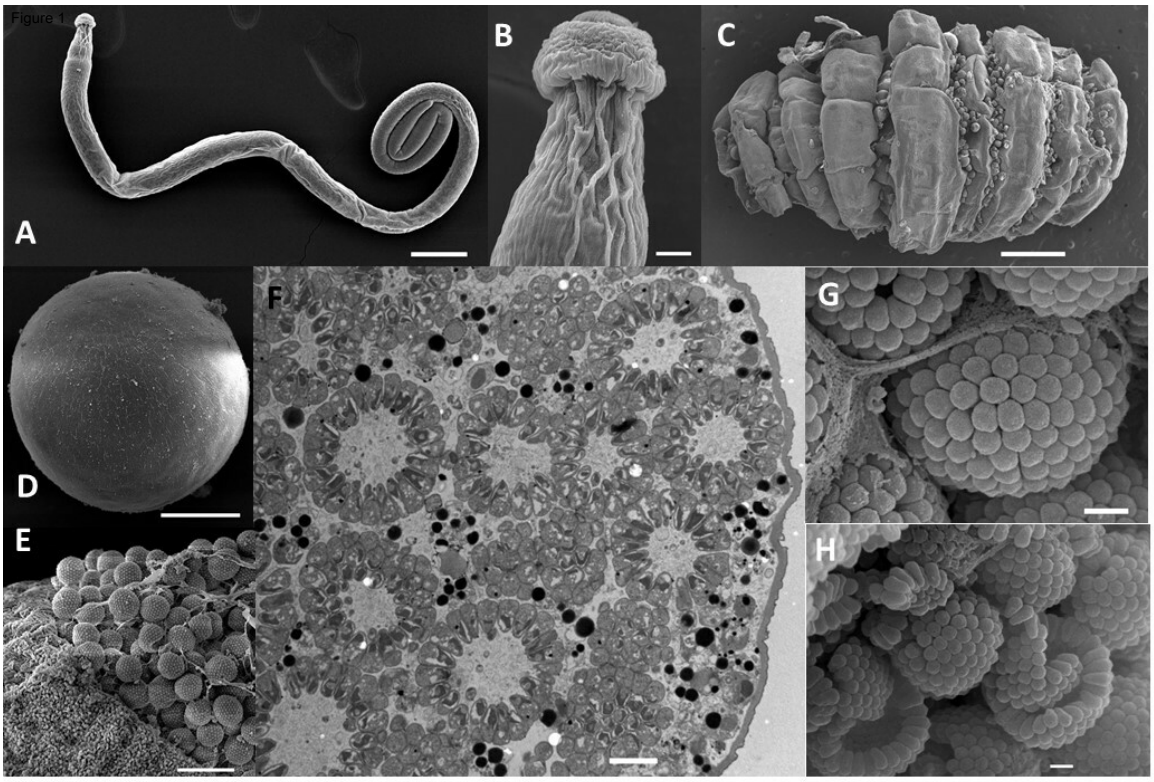
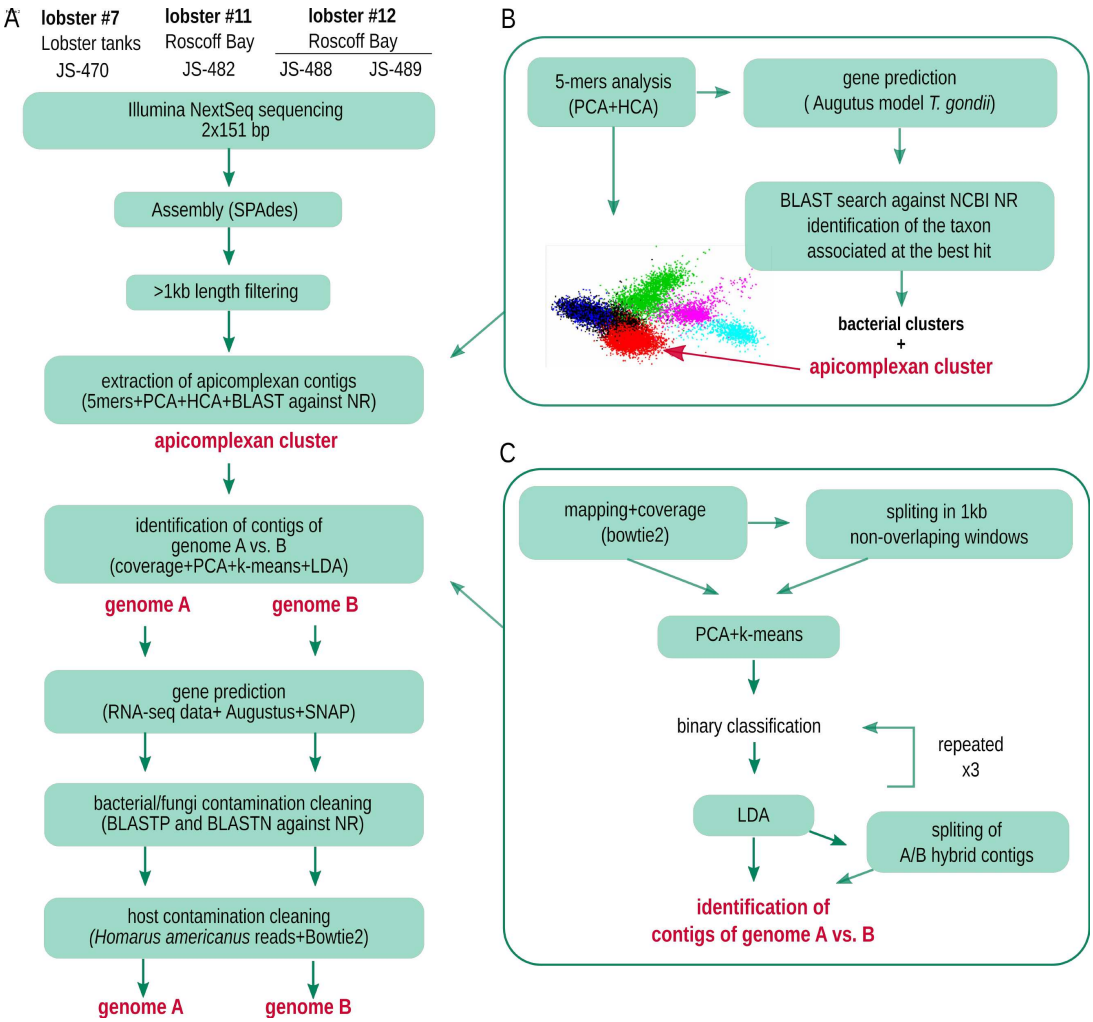
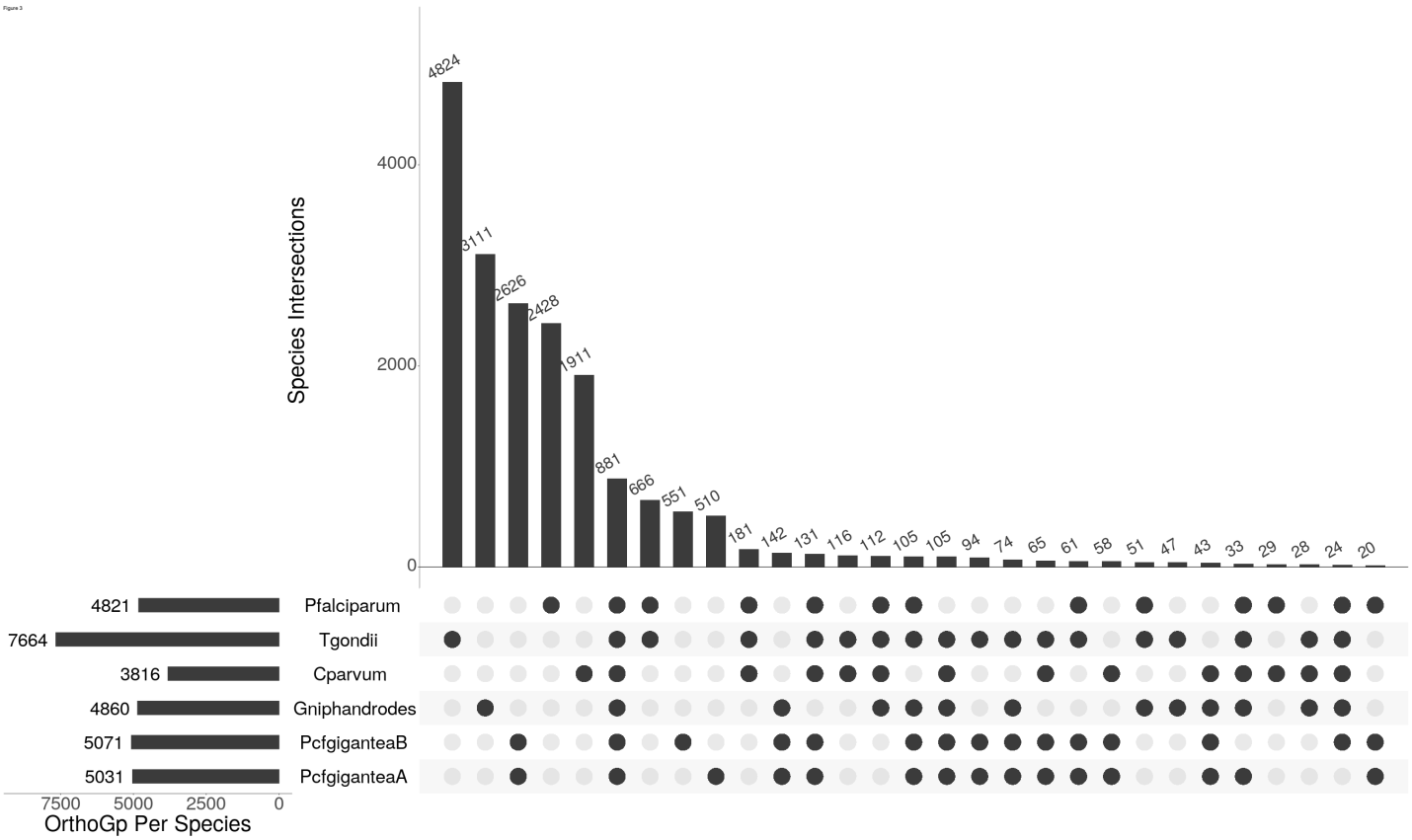


Figure 1
A
B
C
D
E
F
G
H


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
(Augutus 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
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
4000
2000
0
4821
7664
3816
4860
5071
5031
Pfalciparum
Tgondii
Cparvum
Gniphandrodes
PcfgiganteaB
PcfgiganteaA
7500
5000
2500
0
OrthoGp Per Species

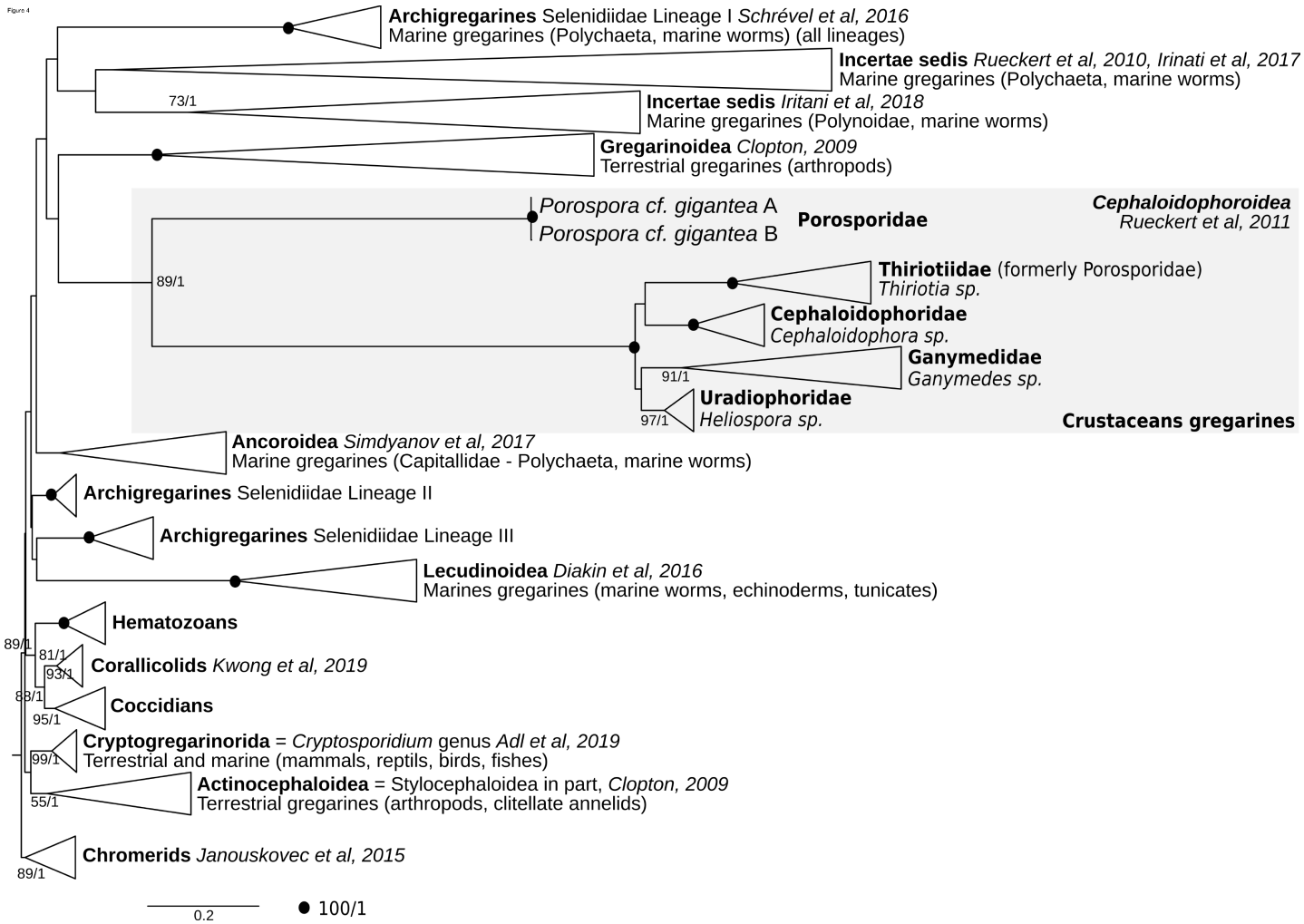
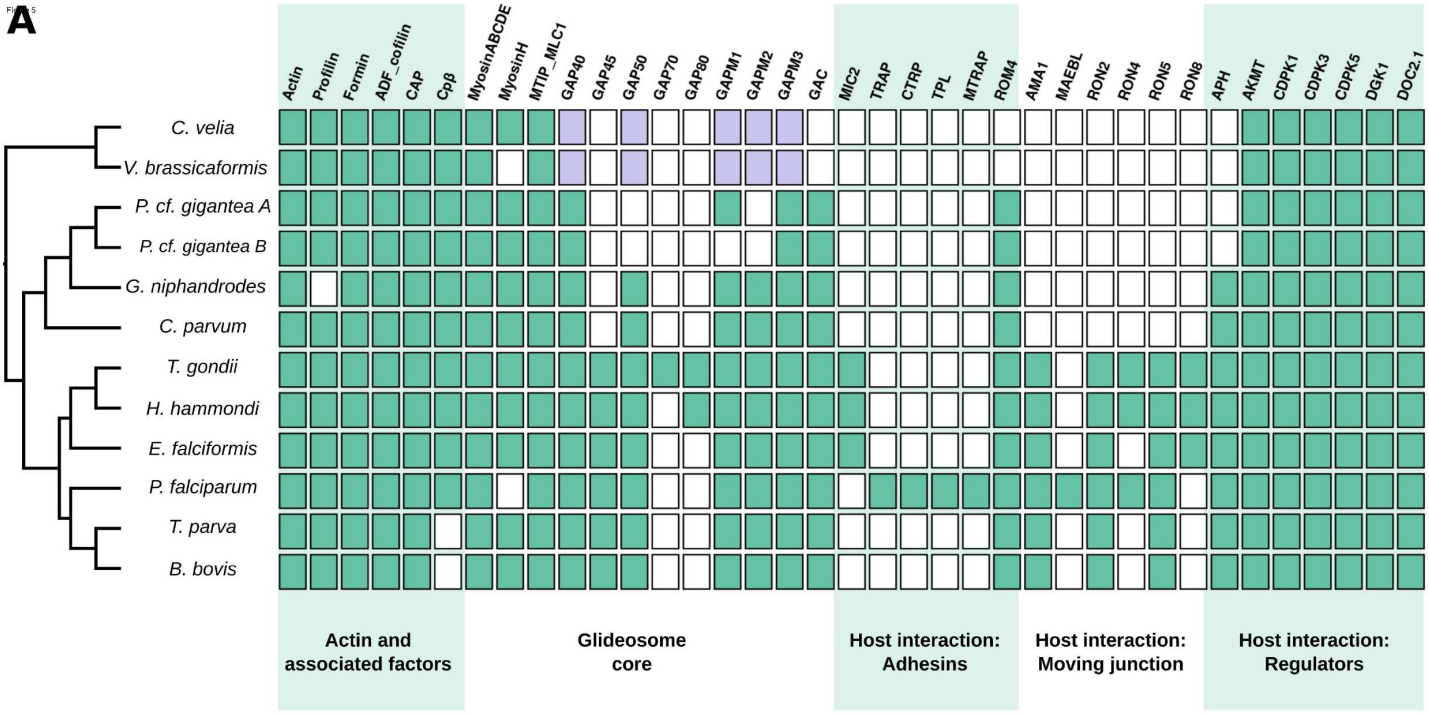


Figure 4
Archigregarines Selenidiidae Lineage I Schrével et al, 2016
Marine gregarines (Polychaeta, marine worms) (all lineages)
Incertae sedis Rueckert et al, 2010, Irinati et al, 2017
Marine gregarines (Polychaeta, marine worms)
73/1
Incertae sedis Iritani et al, 2018
Marine gregarines (Polynoidae, marine worms)
Gregarinoidea Clopton, 2009
Terrestrial gregarines (arthropods)
Porospora cf. gigantea A
Porospora cf. gigantea B
Porosporidae
Cephaloidophoroidea
Rueckert et al, 2011
89/1
Thiriotiidae (formerly Porosporidae)
Thiriotia sp.
Cephaloidophoridae
Cephaloidophora sp.
Ganymedidae
Ganymedes sp.
91/1
Uradiophoridae
97/1
Heliospora sp.
Crustaceans gregarines
Ancoroidea Simdyanov et al, 2017
Marine gregarines (Capitallidae - Polychaeta, marine worms)
Archigregarines Selenidiidae Lineage II
Archigregarines Selenidiidae Lineage III
Lecudinoidea Diakin et al, 2016
Marines gregarines (marine worms, echinoderms, tunicates)
Hematozoans
89/1
81/1
93/1
Corallicolids Kwong et al, 2019
88/1
95/1
Coccidians
Cryptogregarinorida = Cryptosporidium genus Adl et al, 2019
99/1
Terrestrial and marine (mammals, reptils, birds, fishes)
Actinocephaloidea = Stylocephaloidea in part, Clopton, 2009
55/1
Terrestrial gregarines (arthropods, clitellate annelids)
Chromerids Janouskovec et al, 2015
89/1
0.2
100/1


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

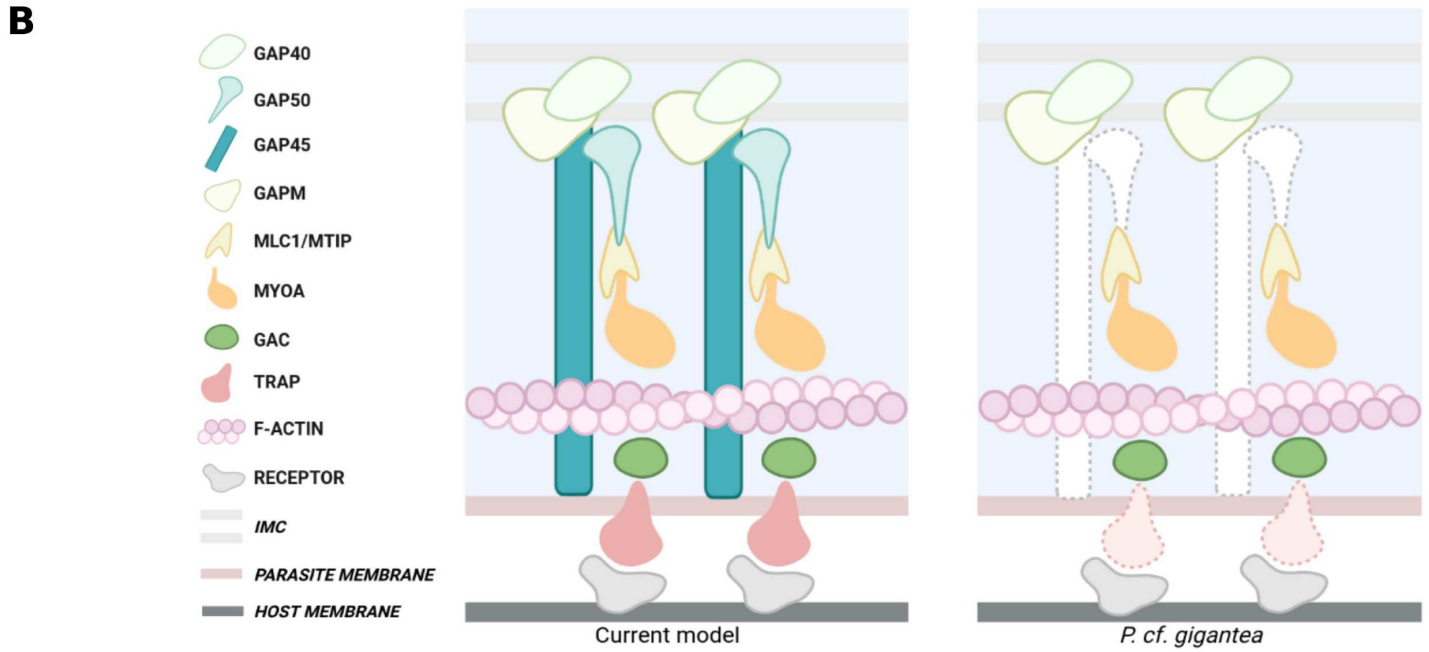

B
GAP40
GAP50
GAP45
GAPM
MLC1/MTIP
MYOA
GAC
TRAP
F-ACTIN
RECEPTOR
IMC
PARASITE MEMBRANE
HOST MEMBRANE
Current model
P. cf. gigantea

Figure 6

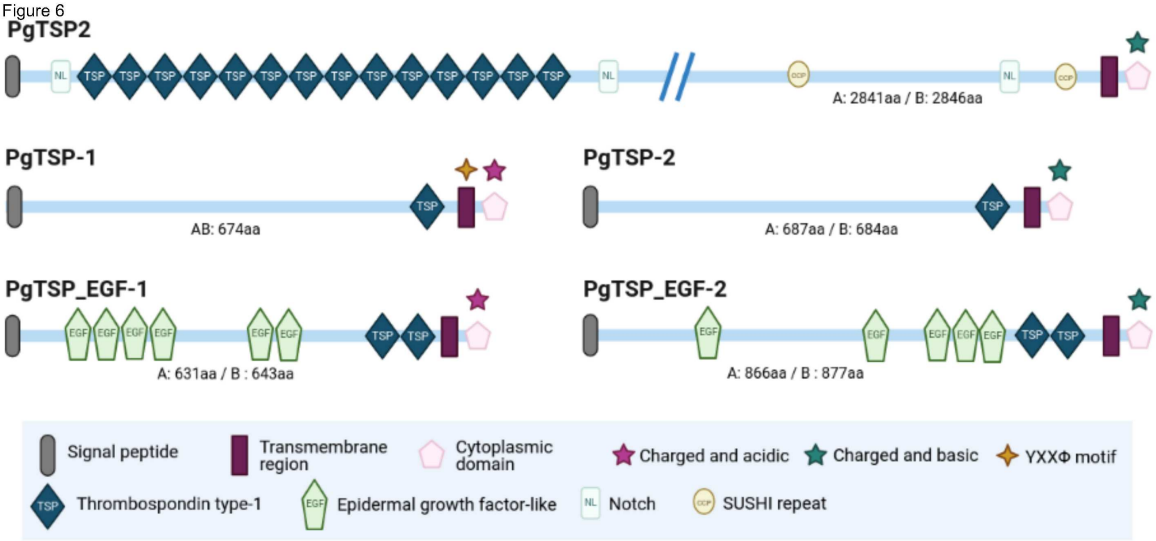

PgTSP2
A: 2841aa / B: 2846aa
PgTSP-1
AB: 674aa
PgTSP-2
A: 687aa / B: 684aa
PgTSP_EGF-1
A: 631aa / B : 643aa
PgTSP_EGF-2
A: 866aa / B : 877aa
Signal peptide
Transmembrane region
Cytoplasmic domain
Charged and acidic
Charged and basic
YXXΦ motif
Thrombospondin type-1
Epidermal growth factor-like
Notch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

Table 1

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)

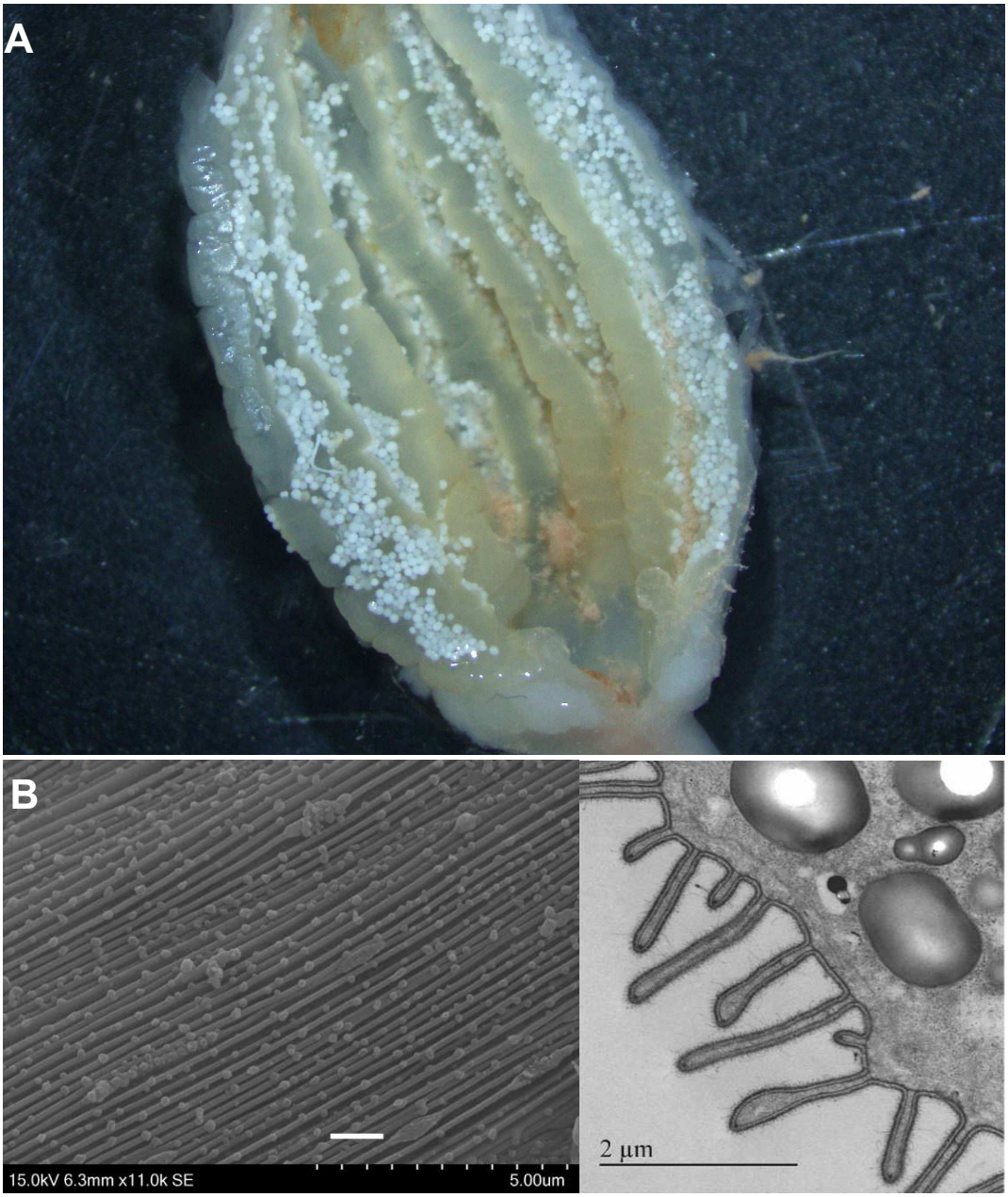


Figure S1. Additional microscopy figures, related to Figure 1. A. Photonic image of the rectal ampulla of Lobster#12, longitudinally opened and heavily packed with *Porospora cf. gigantea* cysts in chitinous folds. The length of the rectal ampulla is about 3 cm. B. Morphological evidence for epicytic folds. Zoom on epicytic folds for trophozoite#9, Lobster#12. Scale=1µm. SEM imaging (left); TEM imaging (right).

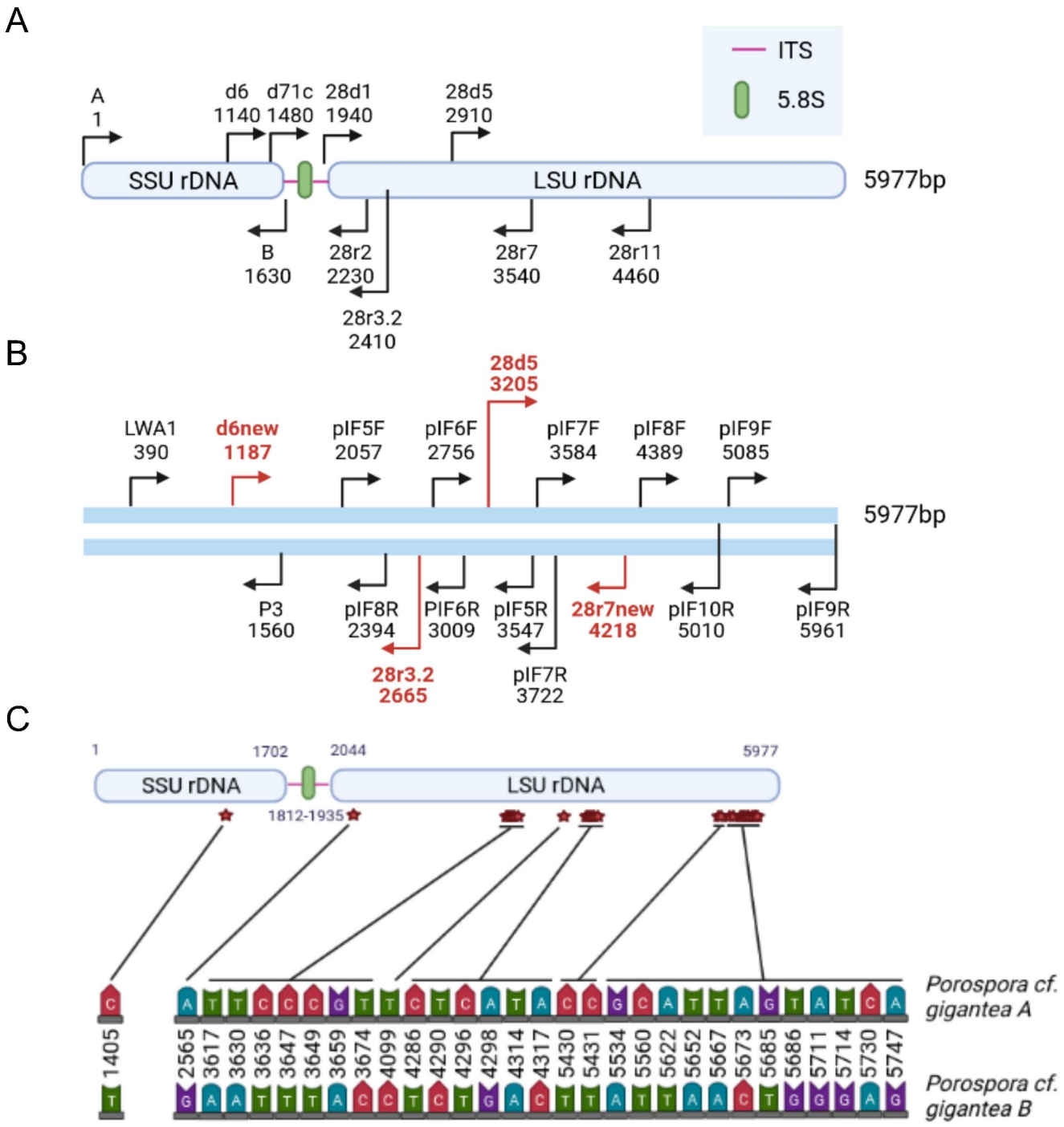


Figure S2. Complete ribosomal locus reconstruction for *Porospora cf. gigantea*, related to Figure 2. A. Complete ribosomal locus for *Cephaloidophora cf. communis* and *Heliospora cf. longissima* from Simdyanov et al. (2015)[S1]. B. Complete ribosomal locus for *Porospora cf. gigantea* A using primers based on Simdyanov et al. (2015)[S1] (red) and novel primers (black) to experimentally amplify and sequence the complete 5977bp locus. See also supp. Table 3 for primer sequences. C. Distribution of the 30 polymorphic positions between A and B complete ribosomal loci.

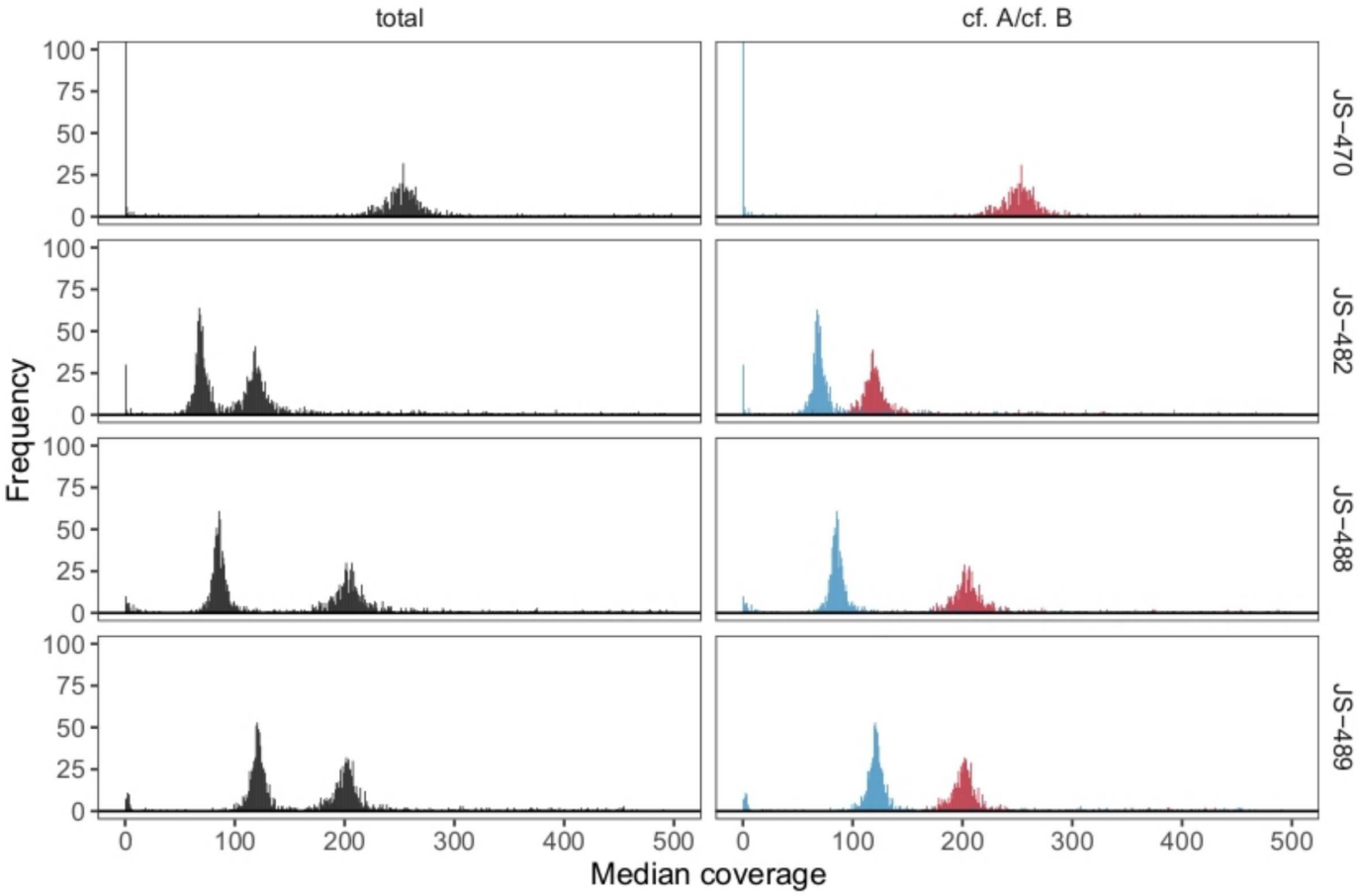


Figure S3. Distribution of the median coverage in each individual library calculated for each contig from the raw assembly, related to Figure 2. Total coverage are presented in black (left side). After genomic attribution of each contig, plot is presented again in red for *Porospora cf. gigantea* A and in blue for *Porospora cf. gigantea* B (right side).

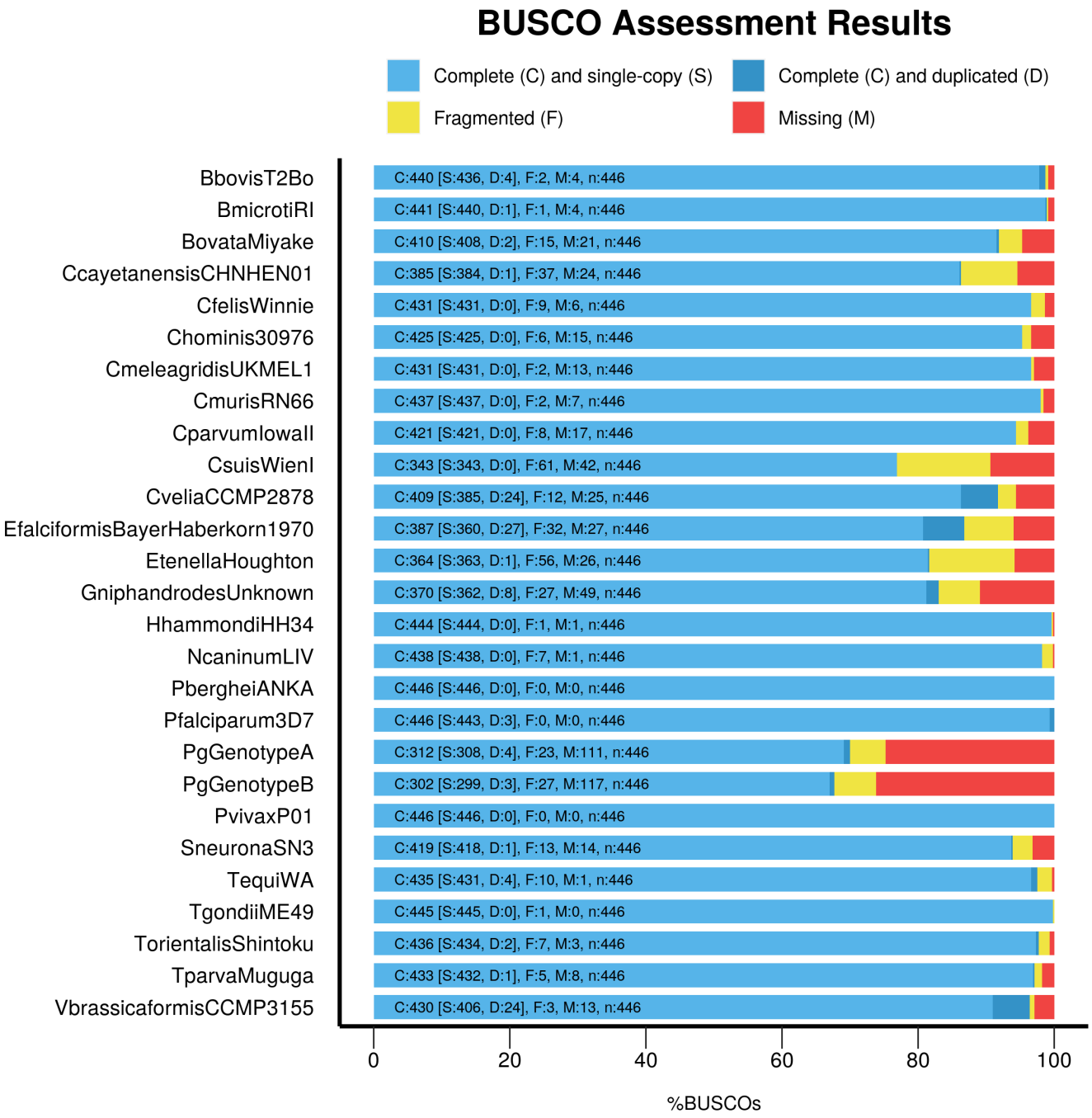


Figure S4. BUSCOs assessment results for the proteomes of both *P. cf. gigantea* and a selection of 25 reference species (geneset apicomplexa_odb10), Related to Figure 2.

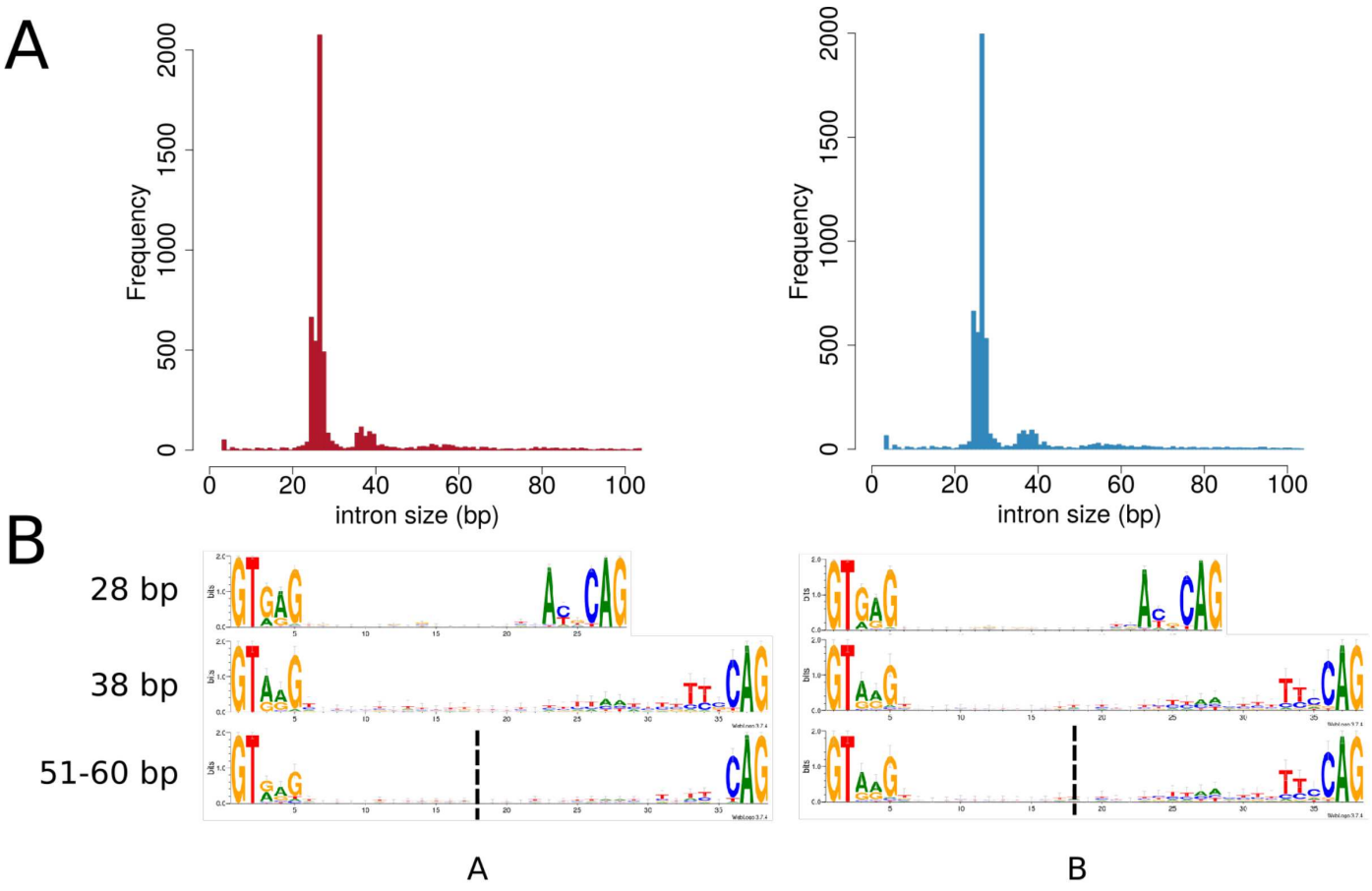


Figure S5. Introns of both P. *cf. gigantea* genomes, Related to Table 1 and Figure 2. A. Length distribution B. Consensus for the major class of short introns (28 bp long) and the two other alternative classes. Data for *P. cf. gigantea* A (left side) and B (right side).

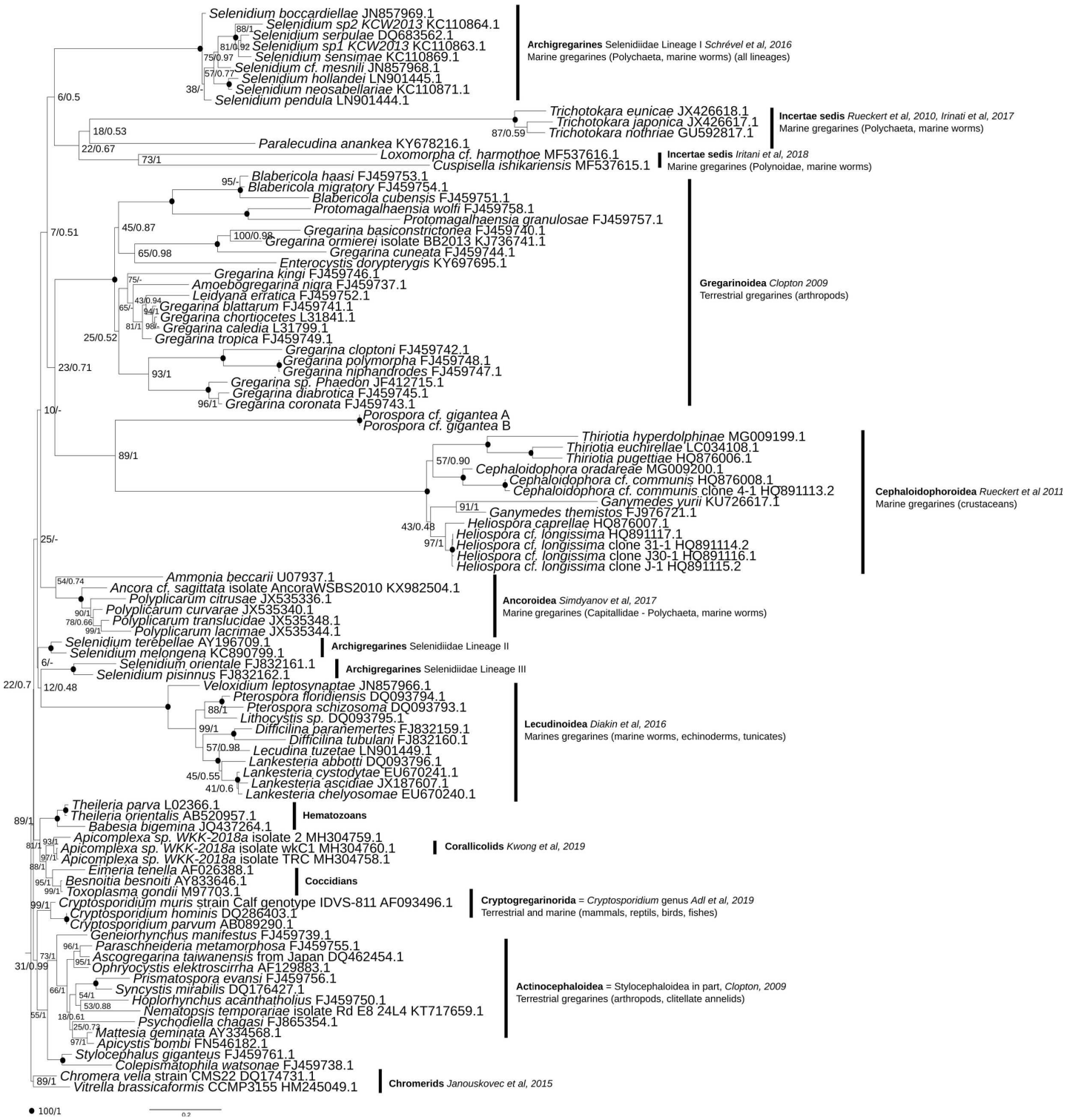


Figure S6. Gregarines/apicomplexan phylogeny, Related to Figure 4. Phylogenetic tree built using 100 18S rDNA sequences 1614 sites in order to situate *P. cf. gigantea* A and B among other known gregarines and apicomplexan clades. Chromerid sequences were used as outgroup, as they are considered as the sister group of all other apicomplexans[S2]. Evolutionary history was inferred by maximum likelihood and bayesian inference using a GTR+G+I model. Topologies were identical according to both methods. Black spots indicate 100/1 supports. Supports <70/0.7 are not shown. Families and associated literature are indicated).

Figure S7. Environmental phylogeny, related to Figure 4. Phylogenetic tree built using 189 18S rDNA sequences for 1135 sites in order to situate two *P. cf. gigantea* A and B among other crustacean gregarines and environmental sequences. Considering that Gregarinoidea sequences were placed as sister group of other crustaceans' gregarines in the gregarines/apicomplexan phylogeny, as well as in recent literature[S3,S4,S5], they were used as outgroup. Evolutionary history was inferred by maximum likelihood and bayesian inference using a GTR+G+I model. Topologies are identical according to both methods. Black spots indicate 100/1 supports. Supports <70/0.7 are not shown. Geographical provenance of all environmental sequences are indicated and their localization is highlighted in bold.

Species	Strain	Gene count (a)	Contigs (a)	Total length (Mb)(b)	GC (%) (b)	Publication (a)
Cryptosporidium hominis	30976	3994	53	9.059	30.13	Guo et al, 2016[S6]
Cryptosporidium muris	RN66	3981	75	9.242	28.47	x
Cryptosporidium meleagridis	UKMEL1	3806	57	8.973	30.97	Ifeonu et al, 2016[S7]
Cryptosporidium parvum (1)	**Iowall**	**4020**	**8**	**9.102**	**30.22**	**Abramhasen et al, 2004**[S8]
Chromera velia (1)	**CCMP2878**	**30806**	**5953**	**193.884**	**49.11**	**Woo et al, 2015**[S2]
Vitrella brassicaformis (1)	**CCMP3155**	**23503**	**1064**	**72.700**	**58.09**	**Woo et al, 2015**[S2]
Gregarina niphandrodes (1)	**Unknown**	**6606**	**468**	**14.008**	**53.78**	**x**
Cyclospora cayetanensis	CHN_HEN01	7592	2297	44.034	51.84	Liu et al, 2016[S9]
Cystoisospora suis	WienI	11767	7880	81.642	49.32	Palmieri et al, 2017[S10]
Eimeria falciformis	BayerHaberkorn1970	6037	753	43.672	49.86	Heitlinger et al, 2014[S11]
Eimeria tenella	Houghton	8634	4664	51.859	51.33	Reid et al, 2014[S12]
Hammondia hammondi	HH34	8177	3676	64.338	52.83	Walzer et al, 2013[S13]
Neospora caninum	LIV	7266	66	59.103	54.82	Reid et al, 2012[S14]
Sarcocystis neurona	SN3	7089	873	124.411	51.41	Blazejewski et al, 2015[S15]
Toxoplasma gondii (1)	**ME49**	**8920**	**2075**	**65.590**	**52.30**	**Lorenzi et al, 2016**[S16]
Babesia bovis	T2Bo	3781	14	8.179	41.59	Brayton et al, 2007[S17]
Babesia microti	RI	3685	6	6.434	36.17	Cornillot et al, 2012[S18]
Babesia ovata	Miyake	5108	91	14.453	49.27	Yamagishi et al, 2017[S19]
Theileria equi	WA	5397	12	11.674	39.48	Kappmeyer et al, 2012[S20]
Theileria orientalis	Shintoku	4058	6	9.010	41.55	Hayashida et al, 2012[S21]
Theileria parva	Muguga	4167	10	8.353	34.04	Gardner et al, 2005[S22]
Cytauxzoon felis	Winnie	4389	357	9.108	31.81	Tarigo et al, 2013[S23]
Plasmodium berghei	ANKA	5245	21	18.780	22,04	Otto et al, 2014[S24]
Plasmodium falciparum (1, 2)	**3D7**	**5712**	**16**	**23.332**	**19.34**	**Gardner et al, 2002**[S25]
Plasmodium vivax	P01	6830	242	29.052	39.78	Auburn et al, 2016[S26]
Plasmodium reichenowi (2)	G01	5909	48	24.471	24.47	Otto et al, 2014[S24]

(1) subset of 6 species (4 apicomplexan + 2 chromerids) used in some comparative analyses, including search for orthogroups and genomic metrics
(2) species used to date the divergence of *P. cf. gigantea* A and B
(a) data from VEupathDB release 41[S27]
(b) data obtained with QUAST[S28]

Table S1. Metrics of 25 apicomplexan and chromerids genomes, considered representative for comparative analyses. Related to Table 1 and Figure 3.

Lobster Specimen	Sampling date	Host from Tanks/Bay	Host sex	Host Lenght (cm)	Host Weight (g)	Cysts load in host rectal ampulla	Trophozoites Load in host gut lumen
#1	24/05/2016	Tanks	male	25	355	< 10	none
#2	24/05/2016	Tanks	male	29	645	10-100	none
#3	24/05/2016	Tanks	female	26	450	10-100	none
#4	25/05/2016	Tanks	female	29	620	< 10	none
#5	25/05/2016	Tanks	male	25	420	10-100	none
#6	26/05/2016	Tanks	male	29	745	< 10	< 10
#7	26/05/2016	Tanks	male	25	375	10-100	< 10
#8	27/05/2016	Tanks	female	27	445	10-100	none
#9	27/05/2016	Tanks	male	26	490	10-100	none
#10	30/05/2016	Tanks	male	26	470	none	none
#11	30/05/2016	Bay	female	25	420	10-100	none
#12	31/05/2016	Bay	male	24	465	100-1000	>10
#13	31/05/2016	Bay	female	24	435	100-1000	none
#14	18/10/2016	Tanks	male	27	485	~200	< 10
#15	19/10/2016	Tanks	male	26	685	none	none
#16	19/10/2016	Tanks	female	27	535	none	none
#17	20/10/2016	Bay	male	23	455	100-300	< 10
#18	20/10/2016	Bay	male	25	450	10-100	< 10
#19	24/10/2016	Bay	female	25	510	10-100	none
#20	24/10/2016	Tanks	male	23	405	10-100	>10
#21	25/10/2016	Tanks	male	27	550	none	none
#22	26/10/2016	Tanks	female	30	895	10-100	>10
#23	26/10/2016	Tanks	male	29	510	10-100	none
#24	03/10/2017	Tanks	female	27	505	10-100	>10
#25	03/10/2017	Tanks	female	28	580	10-100	>10
#26	04/10/2017	Bay	male	34	815	10-100	none
#27	05/10/2017	Bay	male	26	515	100-500	>200
#28	06/10/2017	Tanks	female	30	635	< 10	none
#29	06/10/2017	Tanks	male	26	560	10-100	none
#30	09/10/2017	Tanks	male	27	655	none	none
#31	09/10/2017	Tanks	male	28	710	none	none
#32	09/10/2017	Tanks	female	26	450	10-100	none
#33	11/10/2017	Tanks	male	26	470	10-100	>10
#34	12/10/2017	Tanks	male	27	510	10-100	none
#35	17/07/2015	Bay				100-300	none

Table S2. Sampling of the lobster specimen. Related to Figure 1.

Trophozoite specimen	Host specimen (origin)	Length (µm)	Width ± SD (µm) (n=number of measures)
#1	H0 (Bay)	1796	32.8±4.5 (n=13)
#2	H0 (Bay)	>983	34.2±3.9 (n=14)
#3	H12 (Bay)	none	45.2±3.6 (n=3)
#4	H6 (Tank)	1424	51.5±8.4 (n=25)
#5	H12 (Bay)	none	66 to 23µm
#6	H12 (Bay)	none	none
#7	H12 (Bay)	>1043	71.5±10.1
#8	H12 (Bay)	1858	43.3±7.7 (n=13)
#9	H12 (Bay)	2585	55.5±4.5 (n=6)
#10	H12 (Bay)	none	41
#11	H20 (Tank)	2000	36.2±3.1 (n=6)
#12	H20 (Tank)	2177	37.6±7.3 (n=6)
#13	H20 (Tank)	2222	30.6±1.9 (n=6)
#14	H20 (Tank)	>1062	51.0±6.6 (n=6)
#15	H20 (Tank)	>681	31.5±1.9 (n=6)
Mean value			**41.8+10.4 (n=104)**

(a) mean values for 15 trophozoites from indicated hosts specimen
(b) The sign > corresponds to truncated trophozoites that could not be measured in full.

Table S3. Length and width of trophozoites, related to Figure 1. All values are based on SEM images.

Cyst specimen	Host specimen (origin)	Diameter (µm) (a)
#1	H#12 (Bay)	118.7±4.5 (n=8) *
#2	H#6 (Tank)	168.4±9.1 (n=3) *
#3	H#12 (Bay)	135.3±1.6 (n=4)
#4	H#12 (Bay)	168.6±2.5 (n=4)
#5	H#12 (Bay)	157.1±5.4 (n=4)
#6	H#12 (Bay)	122.7±2.8 (n=4)
#7	H#12 (Bay)	162.0±4.0 (n=4)
#8	H#12 (Bay)	120.6±3.6 (n=4)
#9	H#4 (Tank)	137.0±1.7 (n=4)
#10	H#4 (Tank)	108.4±10.6 (n=4)
#11	H#4 (Tank)	109.8±3.9 (n=4) *
#12	H#4 (Tank)	168.6x128.4 (oval) (n=2)
#13	H#4 (Tank)	220.6±7.0 (n=4)
#14	H#4 (Tank)	252.2±3.7 (n=4)
#15	H#4 (Tank)	240.9±6.5 (n=4)
#16	H#4 (Tank)	211.0±10.9 (n=4)
#17	H#4 (Tank)	141.9±2.0 (n=4)
#18	H#4 (Tank)	118.1±1.4 (n=4)
#19	H#4 (Tank)	104.6±3.2 (n=4)
#20	H#4 (Tank)	108.3±3.1 (n=4)
#21	H#4 (Tank)	121.9±4.1 (n=4)
#22	H#4 (Tank)	129.7±6.1 (n=4)
#23	H#4 (Tank)	124.7±3.6 (n=4)
#24	H#4 (Tank)	230±9.9 (n=4)
#25	H#4 (Tank)	220.6±7.0 (n=4)
Mean		**151.1±45.3 (n=97)**

(a) Mean values measured for 25 cysts. One diameter ± standard deviation (for spherical cysts) or two measures (for oval cyst #12) are given. n, number of measures.
* cysts that were further investigated for gymnospores and zoites measures (see Supplementary Table 6).

Table S4. Diameters of cysts, related to Figure 1. All values are based on SEM images.

Origin of gymnospores and zoites	Host specimen (origin)	Gymnospore Diameter (µm)(a)	Zoite length (µm) (a)	Zoite width (µm) (a)
Cyst#1	H#12 (Bay)	4.97±0.37 (n=60)	1.17±0.07 (n=7)	0.565±0.218 (n=50)
Cyst#2	H#6 (Tanks)	5.79±0.62 (n=97)	1.04±0.11 (n=45)	0.616±0.033 (n=11)
Cyst#11	H#4 (Tanks)	6.04±0.72 (n=56)	1.09±0.07 (n=16)	0.674±0.043 (n=35)
JS-463b_0003	H#4 (Tanks)			0.660±0.045 (n=10)
JS-463b_0016 (a)	H#4 (Tanks)	5.30±0.05 (n=4)		0.643±0.042 (n=10)
JS-463b_0016 (b)	H#4 (Tanks)	5.08±0.16 (n=4)		
JS-463b_0016 (c)	H#4 (Tanks)	5.92±0.18 (n=4)	1.19±0.09 (n=3)	0.673±0.086 (n=3)
JS-463b_0020 (a)	H#4 (Tanks)	5.26±0.18 (n=4)		0.662±0.035 (n=10)
JS-463b_0020 (b)	H#4 (Tanks)	5.33±0.04 (n=4)		
JS-463b_0020 (c)	H#4 (Tanks)	6.21±0.22 (n=4)		
JS-463b_0020 (d)	H#4 (Tanks)	6.63±0.30 (n=4)		
JS-463b_0020 (e)	H#4 (Tanks)	6.24±0.18 (n=4)		
JS-463b_0020 (f)	H#4 (Tanks)	5.64±0.11 (n=4)		
JS-463b_0027 (a)	H#4 (Tanks)	5.66±0.19 (n=4)		0.646±0.050 (n=10)
JS-463b_0027 (b)	H#4 (Tanks)	4.69±0.10 (n=4)		0.656±0.029 (n=10)
JS-463b_0027 (c)	H#4 (Tanks)	5.70±0.10 (n=4)		0.643±0.039 (n=10)
JS-463b_0028	H#4 (Tanks)		1.09±0.09 (n=10)	
JS-463b_0030	H#4 (Tanks)	5.80±0.08 (n=4)		0.684±0.085 (n=10)
JS-463b_0036	H#4 (Tanks)		1.16±0.07 (n=13)	0.613±0.037 (n=10)
Mean		**5.63±0.69 (n=265)**	**1.04±0.16 (n=105)**	**0.630±0129 (n=176)**

(a) n, number of measures.

Table S5. Diameters of gymnospores and zoites, related to Figure 1. Diameters were measured for hundreds of gymnospores within cysts (3 first lines) or released from cysts (remaining lines). Whenever possible, length and apical width ± standard deviation of their constitutive zoites were also measured. All values are based on SEM images.

Video record	Length of recording (s)	Trophozoites number	Length (µm)	Speed (µm/s)
G5310002	40	T10	~2190	51.6
		T11	~1876	48.9
		T12	~2113	50.3
		T13	~2113	51.8
		T14	~1801	49.8
		T15	~1807	51.4
G5310003	34	T3	~3100	87-89
		T4	~4500	100-109
		T5	~3900	108-115
G5310004	60 and 48	T1	~3000	56-63
		T2	~3600	80-81
G5310018	20	T6	~2540	76
		T7	~4600	97-103
		T8	~4100	104
		T9	~3595	91-94

Table S6. Gliding recordings, related to Figure 1. All recordings are from trophozoites collected from Lobsters#12 and #13 on 31/05/2016. Due to the lack of scale on these videos, we used the mean width of trophozoites, as determined by using SEM images (41.8±10.4µm, see Table S4), to calibrate the other measures.

Primer name	Primer sequence	orientation	reference
LWA1	5'- GGAAGGCAGCAGGCGCGC - 3'	forward	Schrevel et al., 2016[S29]
EukP3	5'- GACGGGCGGTGTGTAC - 3'	reverse	Lara et al., 2007[S30]
28d5	5'- CCGCTAAGGAGTGTGTAACAAC - 3'	forward	Simdyanov et al., 2015[S1]
28r3.2	5'- ACTCCTYRGTCCGTGTTTCA - 3'	reverse	Simdyanov et al., 2015[S1]
28r2	5'- TACTTGTYBRCTATCG - 3'	reverse	Simdyanov et al., 2015[S1]
d6new	5'- GGTGGTGCATGGCCAAACTT - 3'	forward	Modified from Simdyanov et al., 2015[S1]
28d5short	5'- GCTAAGGAGTGTGTAACAAC - 3'	forward	Modified from Simdyanov et al., 2015[S1]
28r7new	5'- TAATTTGCCGACTTCCCTCA - 3'	reverse	Modified from Simdyanov et al., 2015[S1]
PIF5F	5'- ACATTCCTTGGGTTACCC - 3'	forward	This study
PIF6F	5'- TAACGACCCGAAAATCGG - 3'	forward	This study
PIF7F	5'- CATGCTAACACAAGGGGG - 3'	forward	This study
PIF8F	5'- CCGACAGTTTAACTAAAACC - 3'	forward	This study
PIF9F	5'- GAGATCATATCGACGCGG- 3'	forward	This study
PIF5R	5'- CATCAGTGCGACGATACC - 3'	reverse	This study
PIF6R	5'- GTTTGAGAATCAGTCGAGG - 3'	reverse	This study
PIF7R	5'- CTTTCGACTTCCGACAGC - 3'	reverse	This study
PIF8R	5'- TTGTTTGCTATCGGTATAGG - 3'	reverse	This study
PIF9R	5'- AAATCTCAAGAGAGATGGAG- 3'	reverse	This study
PIF10R	5'- GCTAAGGATCGATAGGCC - 3'	reverse	This study

Table S7. List of primers used for ribosomal locus amplification and sequencing by Sanger technology. Related to Figure 2. Related to Figure 2.

Protein name	P. cf. gigantea A	P. cf. gigantea B
Actin	gene148 gene3475 gene9783 gene7381	gene1189 gene176 gene7820
Profilin	gene4246	gene2168
Formin	gene1045 gene6583	gene2548 gene2999
ADF_cofilin	gene2779	gene8323
CAP	gene8064	gene1376
Cpβ F-actin capping protein β-subunit	gene5850	gene2019
MyosinACDE ClassXIV	gene7213 gene5479 gene4089 gene9485 gene5566	gene1418 gene3446 gene8205 gene4288 gene1172
MyosinH ClassXIV	gene5024	gene6924
MTIP_MLC1	gene3722	gene9046
GAP40	gene2452	gene4035
GAPM1-2-3	GAPM3 gene8241 GAPMx gene7900 GAPMx gene4892	GAPM3 gene1427 GAPMx gene10037
GAC	gene9350	gene3183
ROM4	gene7177 gene6011	gene2979 gene5358
AKMT	gene241	gene2337
CDPK1(Tg)/CDPK4(Pf)	gene9265	gene7530
CDPK3(Tg)/CDPK1(Pf)	gene8870	gene2741
CDPK5(Pf)/CDPK5(Tg)	gene4113	gene8961
DGK1	gene3462	gene10271
DOC2.1	gene3776	gene6607
TSP-1 (a)	gene7210	gene1404
TSP-2 (a)	gene4371	gene4603
TSP2 (a)	gene9608	gene6110-6121
TSP_EGF-1 (a)	gene2135	gene5987
TSP_EGF-2 (a)	gene951	gene3952

(a) TRAP like candidates

Table S8. *P. cf. gigantea* A and B glideosome and TRAP-like proteins identifiers, related to Figure 6.

Supplemental References

S1.

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.

S2.

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.

S3.

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.

S4.

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.

S5.

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.

S6.

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.

S7.

Ifeonu, O.O., Chibucos, M.C., Orvis, J., Su, Q., Elwin, K., Guo, F., Zhang, H., Xiao, L., Sun, M., Chalmers, R.M., et al. (2016). Annotated draft genome sequences of three species of *Cryptosporidium*: *Cryptosporidium meleagridis* isolate UKMEL1, *C. baileyi* isolate TAMU-09Q1 and *C. hominis* isolates TU502_2012 and UKH1. Pathogens and Disease 74, ftw080. 10.1093/femspd/ftw080.

S8.

Abrahamsen, M.S., Templeton, T.J., Enomoto, S., Abrahante, J.E., Zhu, G., Lancto, C.A., Deng, M., Liu, C., Widmer, G., Tzipori, S., et al. (2004). Complete Genome Sequence of the Apicomplexan , *Cryptosporidium parvum*. 304, 6.

S9.

Liu, S., Wang, L., Zheng, H., Xu, Z., Roellig, D.M., Li, N., Frace, M.A., Tang, K., Arrowood, M.J., Moss, D.M., et al. (2016). Comparative genomics reveals *Cyclospora cayetanensis* possesses coccidia-like metabolism and invasion components but unique surface antigens. BMC Genomics 17, 316. 10.1186/s12864-016-2632-3.

S10.

Palmieri, N., Shrestha, A., Ruttkowski, B., Beck, T., Vogl, C., Tomley, F., Blake, D.P., and Joachim, A. (2017). The genome of the protozoan parasite *Cystoisospora suis* and a reverse vaccinology approach to identify vaccine candidates. International Journal for Parasitology 47, 189–202. 10.1016/j.ijpara.2016.11.007.

S11.

Heitlinger, E., Spork, S., Lucius, R., and Dieterich, C. (2014). The genome of *Eimeria falciformis*, reduction and specialization in a single host apicomplexan parasite. BMC Genomics 15, 696. 10.1186/1471-2164-15-696.

S12.

Reid, A.J., Blake, D.P., Ansari, H.R., Billington, K., Browne, H.P., Bryant, J., Dunn, M., Hung, S.S., Kawahara, F., Miranda-Saavedra, D., et al. (2014). Genomic analysis of the causative agents of coccidiosis in domestic chickens. Genome Res. 24, 1676–1685. 10.1101/gr.168955.113.

S13.
Walzer, K.A., Adomako-Ankomah, Y., Dam, R.A., Herrmann, D.C., Schares, G., Dubey, J.P., and Boyle, J.P. (2013). Hammondia hammondi, an avirulent relative of *Toxoplasma gondii*, has functional orthologs of known *T. gondii* virulence genes. Proceedings of the National Academy of Sciences 110, 7446–7451. 10.1073/pnas.1304322110.
S14.
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.
S15.
Blazejewski, T., Nursimulu, N., Pszenny, V., Dangoudoubiyam, S., Namasivayam, S., Chiasson, M.A., Chessman, K., Tonkin, M., Swapna, L.S., Hung, S.S., et al. (2015). Systems-Based Analysis of the *Sarcocystis neurona* Genome Identifies Pathways That Contribute to a Heteroxenous Life Cycle. mBio 6, e02445-14. 10.1128/mBio.02445-14.
S16.
Lorenzi, H., Khan, A., Behnke, M.S., Namasivayam, S., Swapna, L.S., Hadjithomas, M., Karamycheva, S., Pinney, D., Brunk, B.P., Ajioka, J.W., et al. (2016). Local admixture of amplified and diversified secreted pathogenesis determinants shapes mosaic *Toxoplasma gondii* genomes. Nat Commun 7, 10147. 10.1038/ncomms10147.
S17.
Brayton, K.A., Lau, A.O.T., Herndon, D.R., Hannick, L., Kappmeyer, L.S., Berens, S.J., Bidwell, S.L., Brown, W.C., Crabtree, J., Fadrosh, D., et al. (2007). Genome Sequence of *Babesia bovis* and Comparative Analysis of Apicomplexan Hemoprotozoa. PLoS Pathogens 3, e148. 10.1371/journal.ppat.0030148.
S18.
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.
S19.
Yamagishi, J., Asada, M., Hakimi, H., Tanaka, T.Q., Sugimoto, C., and Kawazu, S. (2017). Whole-genome assembly of *Babesia ovata* and comparative genomics between closely related pathogens. BMC Genomics 18. 10.1186/s12864-017-4230-4.
S20.
Kappmeyer, L.S., Thiagarajan, M., Herndon, D.R., Ramsay, J.D., Caler, E., Djikeng, A., Gillespie, J.J., Lau, A.O., Roalson, E.H., Silva, J.C., et al. (2012). Comparative genomic analysis and phylogenetic position of *Theileria equi*. BMC Genomics 13, 603. 10.1186/1471-2164-13-603.
S21.
Hayashida, K., Hara, Y., Abe, T., Yamasaki, C., Toyoda, A., Kosuge, T., Suzuki, Y., Sato, Y., Kawashima, S., Katayama, T., et al. (2012). Comparative Genome Analysis of Three Eukaryotic Parasites with Differing Abilities To Transform Leukocytes Reveals Key Mediators of *Theileria*-Induced Leukocyte Transformation. mBio 3. 10.1128/mBio.00204-12.
S22.
Gardner, M.J. (2005). Genome Sequence of *Theileria parva*, a Bovine Pathogen That Transforms Lymphocytes. Science 309, 134–137. 10.1126/science.1110439.
S23.
Tarigo, J.L., Scholl, E.H., Bird, D.McK., Brown, C.C., Cohn, L.A., Dean, G.A., Levy, M.G., Doolan, D.L., Trieu, A., Nordone, S.K., et al. (2013). A Novel Candidate Vaccine for *Cytauxzoonosis* Inferred from Comparative Apicomplexan Genomics. PLoS ONE 8, e71233. 10.1371/journal.pone.0071233.
S24.
Otto, T.D., Böhme, U., Jackson, A.P., Hunt, M., Franke-Fayard, B., Hoeijmakers, W.A.M., Religa, A.A., Robertson, L., Sanders, M., Ogun, S.A., et al. (2014). A comprehensive evaluation of rodent malaria parasite genomes and gene expression. BMC Biology 12. 10.1186/s12915-014-0086-0.

S25.
Gardner, M.J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R.W., Carlton, J.M., Pain, A., Nelson, K.E., Bowman, S., et al. (2002). Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature 419, 498–511. 10.1038/nature01097.
S26.
Auburn, S., Böhme, U., Steinbiss, S., Trimarsanto, H., Hostetler, J., Sanders, M., Gao, Q., Nosten, F., Newbold, C.I., Berriman, M., et al. (2016). A new *Plasmodium vivax* reference sequence with improved assembly of the subtelomeres reveals an abundance of pir genes. Wellcome Open Research 1, 4. 10.12688/wellcomeopenres.9876.1.
S27.
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.
S28.
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.
S29.
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.
S30.
Lara, E., Berney, C., Ekelund, F., Harms, H., and Chatzinotas, A. (2007). Molecular comparison of cultivable protozoa from a pristine and a polycyclic aromatic hydrocarbon polluted site. Soil Biology and Biochemistry 39, 139–148. 10.1016/j.soilbio.2006.06.017.

Communications

2 Fev 2021 Julie Boisard*, Evelyne Duvernois-Berthet, Loïc Ponger, Isabelle Florent, *Challenges and solutions for studying divergent eukaryotic genomes of non- model and non-cultivable species*, **ALPHY: Bioinformatics and Evolutionary Genomics**, Virtual Meeting

10 Dec 2020 Julie Boisard*, Evelyne Duvernois-Berthet, Loïc Ponger, Isabelle Florent, *Caractérisation du génome et du protéome de Grégarines, modèles d'étude pour comprendre la diversification des apicomplexes et l'adaptation à la vie parasitaire*, **Rencontres Bioinformatiques du MNHN**, Paris, France

18 Jan 2019 Julie Boisard*, Evelyne Duvernois-Berthet, Loïc Ponger, Isabelle Florent, *Assemblage de novo de génomes de grégarines*, **Journée des doctorants et post-doctorants MCAM**, Paris, France

* speaker

Résumés en français

Résumé court

Les apicomplexes sont des micro-organismes eucaryotes unicellulaires ayant évolué vers un mode de vie parasitaire strict. Certains groupes d'apicomplexes comprennent des espèces à l'origine de pathologies graves telles que le paludisme (*Plasmodium* spp.), la toxoplasmose (*Toxoplasma gondii*) et la cryptosporidiose (*Cryptosporidium* spp.). Si les génomes de ces agents hautement pathogènes sont maintenant bien documentés, ce n'est pas le cas pour d'autres lignées d'apicomplexes comme les grégarines, considérées comme basales au sein des Apicomplexa, ont un faible pouvoir pathogène et surtout ne sont pas cultivables. Leur étude moléculaire représente actuellement un goulot d'étranglement majeur, alors qu'une connaissance précise de leurs génomes serait essentielle pour mieux comprendre l'histoire évolutive des parasites apicomplexes et la diversité de leurs adaptations au mode de vie parasitaire.

Au cours de cette thèse, la caractérisation du génome de 2 grégarines marines, *Porospora gigantea*, parasite du homard européen *Homarus gammarus* et *Diplauxis hatti*, parasite du ver marin Polychaeta *Perinereis cultrifera* ; et d'une grégarine terrestre, *Gregarina acridiorum*, parasite du criquet *Locusta migratoria* a été réalisée. La découverte de deux génomes coexistants correspondant à l'espèce morphologiquement décrite *P. gigantea*, tout comme un second exemple impliquant *G. acridiorum*, illustrent l'ampleur des révisions taxonomiques à venir, et la nécessité de se tourner vers des marqueurs moléculaires, probablement à l'échelle génomique, pour évaluer correctement la diversité des grégarines. Par ailleurs, les premières analyses de génomique comparative incluant des grégarines révèlent une diversité génétique insoupçonnée chez ces organismes. Une analyse des protéines du glidéosome à l'échelle des apicomplexes a également été réalisée. Ce modèle fait référence à une structure moléculaire complexe à l'origine du gliding, un mouvement caractéristique des Apicomplexa et essentiel à la manifestation de leur pathogénicité. Une étude comparative détaillée met en évidence sa conservation différentielle à l'échelle des apicomplexes, suggérant une diversité d'adaptations aux problèmes de motilité et d'invasion des cellules hôtes. Ce travail illustre l'importance de prendre en compte les apicomplexes non-modèles, non pathogènes et non cultivables pour fournir des indices nouveaux sur les capacités d'adaptation déployées par ce groupe de parasite à l'importance écologique et médicale majeure.

Résumé détaillé

Avec seulement une centaine de génomes déchiffrés pour ∼350 genres et ∼6000 espèces décrites, les Apicomplexa sont un groupe pour lequel il reste encore beaucoup à découvrir. Les apicomplexes sont loin d'avoir livré tous les secrets de la diversité des innovations moléculaires développées au cours de leur histoire évolutive. Le premier chapitre de cette thèse présente en particulier les grégarines et les raisons pour lesquelles elles constituent un groupe d'étude crucial pour comprendre l'histoire évolutive des apicomplexes.

À ce jour, l'exploration moléculaire des parasites apicomplexes a principalement concerné un très petit nombre d'espèces qui ont en commun :

1) d'infecter les humains en provoquant des maladies graves comme le paludisme ou des maladies moins graves mais répandues dans le monde entier comme la toxoplasmose et la cryptosporidiose ;

2) d'être cultivable en laboratoire, au moins pour certains stades de développement ;

3) d'avoir fait l'objet de développements méthodologiques extrêmement sophistiqués.

Dans ce panorama, les grégarines, membres à part entière du phylum Apicomplexa, ont été jusqu'à présent délaissées pour des raisons exactement corollaires :

1) elles n'infectent pas les humains ;

2) elles ne sont pas faciles à cultiver ;

3) alors qu'il existe une littérature très abondante sur leurs cycles de vie, leurs morphologies et leurs ultrastructures, elles sont quasiment inconnues aux niveaux génomique/transcriptomique et ont fait l'objet de très peu d'études biochimiques.

Mais leur avenir est désormais ouvert à l'exploration. Dans un premier temps, il convient de privilégier les espèces connues, sélectionnées soit pour leurs caractéristiques biologiques (intestinales, coelomiques, motiles, non motiles), soit pour leur position phylogénétique particulière (archigrégarines vs eugrégarines comme grande distinction). Comme indiqué précédemment, le principal obstacle à la production de génomes de grégarines est la difficulté de collecter suffisamment de matériel biologique pour le séquençage génomique. En effet, aucune espèce de grégarine n'est actuellement cultivable en laboratoire. Il existe deux possibilités pour surmonter ces problèmes : la première consiste à maintenir des hôtes infectés par des grégarines dans des conditions de laboratoire. Si cette solution ne résout pas le problème de la contamination par l'hôte et les microorganismes environnants lors de la collecte, elle garantit un accès régulier aux formes de développement visées et permet ainsi d'acquérir une quantité adéquate de matériel biologique. La seconde solution consiste à identifier des modèles biologiques capables, par leurs caractéristiques intrinsèques, de fournir un matériel biologique suffisant. Le second chapitre

de cette thèse montre comment ces deux solutions ont été exploitées et ont abouti à la sélection de trois modèles biologiques : 2 grégarines marines, *Porospora gigantea*, parasite du homard européen *Homarus gammarus* et *Diplauxis hatti*, parasite du ver marin Polychaeta *Perinereis cultrifera* ; et 1 grégarine terrestre, *Gregarina acridiorum*, parasite du criquet pèlerin *Locusta migratoria*.

Le laboratoire MCAM a entrepris le séquençage de l'ADN génomique des grégarines en s'appuyant sur une expertise unique au Muséum national d'Histoire naturelle grâce à la présence de deux experts mondiaux des grégarines, le Pr J. Schrével et le Dr I. Desportes. En 2016, lors d'une campagne à Roscoff (Bretagne, France), une grégarine marine a attiré l'attention de I. Florent et J. Schrével. Il s'agit de l'espèce *Porospora gigantea* Van Beneden, 1869, un parasite du homard européen *Homarus gammarus*. Selon la littérature, cette grégarine présente des formes kystiques particulières situées dans l'ampoule rectale de ces crustacés et spécifiques de la famille des Porosporidae. Celles-ci ont alors été observées et échantillonnées : elles contenaient des milliers de gymnospores, constituées de monocouches de zoïtes nus (sans enveloppe). Ces formes kystiques sont apparues comme une source exceptionnelle de matériel biologique pour l'isolement de l'ADN génomique à séquencer, fournissant des millions de copies du génome pour chaque kyste et un faible risque de contamination par l'hôte ou les microorganismes environnants. L'accès depuis 2012 à l'élevage en vivarium du MNHN d'une espèce de criquet, *Locusta migratoria*, infectée par une grégarine terrestre, *Gregarina acridiorum*, a permis d'isoler suffisamment de matériel biologique pour le séquençage génomique en 2017. Enfin, les données génomiques d'une seconde grégarine marine, *Diplauxis hatti*, une grégarine coelimique de l'annélide polychète *Perinereis cultrifera*, ont été acquises par le laboratoire fin 2018. *Diplauxis hatti* présente une adaptation unique de son cycle de vie à son hôte. En effet, des observations sur les populations naturelles de la Manche ont montré que la libération des spores du parasite est concomitante à la ponte des polychètes. Ainsi, en collectant des hôtes pendant leur saison de reproduction, fin mars ou début avril, J. Schrével, G. Prensier et L. Guillou ont pu recueillir les oocystes de *Diplauxis hatti*.

La collecte d'un matériel biologique suffisant n'est que le premier des défis posés par la génomique des organismes non modèles et non cultivables. En effet, en l'absence de données suffisamment proches pour servir de référence, l'élimination des contaminants ainsi que la prédiction des gènes à partir des assemblages génomiques sont d'autres obstacles. L'objectif du chapitre 2 est de présenter les principaux défis associés à l'analyse des génomes eucaryotes hautement divergents d'espèces non-modèles et non cultivables, et de proposer des solutions pour les relever. Les problématiques soulevées par chacune des grégarines pour lesquelles nous avons pu produire des données génomiques ont été abordées de manière différente, en les adaptant à leur situation spécifique. Le prélèvement directement sur l'hôte est similaire à l'échantillonnage environnemental dans la mesure où il est probable de collecter les bactéries ou les champignons environnants.

Lorsqu'un génome de référence n'est pas disponible, l'élimination de ces contaminants environnementaux constitue un défi particulier et peut être abordée par des méthodes utilisées en métagénomique, telles que la composition en kmer ou l'analyse de la similarité des séquences. Si un échantillon environnemental contient un mélange de deux génomes étroitement apparentés, la couverture médiane des contigs par les reads peut permettre de les distinguer.

Pour la prédiction des gènes d'une nouvelle espèce sans référence proche, la solution optimale est actuellement de construire un modèle de gènes en utilisant les données RNAseq. Cependant, cela suppose que l'on soit capable de générer de telles données, ce qui n'est pas toujours évident. En plus des contraintes évoquées ci-dessus, et qui restent valables, les préparations pour le séquençage de l'ARN sont plus délicates à réaliser, surtout lorsque le matériel biologique est rare. Cette situation étant fréquemment rencontrée dans notre discipline, le but a été de proposer des solutions alternatives à l'utilisation d'un modèle de gène basés sur les ARN. Elles consistent principalement à générer des modèles de gènes à partir des espèces les plus proches ou partageant des caractéristiques génomiques communes. La caractérisation des génomes de *G. acridiorum* et *D. hatti* est à ce jour la plus complète possible au regard des données disponibles. De nouvelles références pour les grégarines ont pu être définies concernant la taille et la structure du génome, ainsi que le nombre de protéines attendues. Ces prédictions de gènes ne pourront être affinées efficacement qu'avec l'apport de données ARN, mais entre-temps, la diversité génomique des grégarines se dévoile déjà plus clairement. En ce sens, l'un des résultats les plus frappants de cette recherche concerne précisément l'ampleur de cette diversité. Le modèle de gènes de *P. gigantea* était censé être un modèle de gènes suffisamment proche pour prédire les gènes d'autres grégarines. Pourtant, ce modèle s'est avéré trop divergent, même pour une grégarine marine, en raison d'un biais riche en AT dans les séquences de *D. hatti*. Les techniques de séquençage évoluent très rapidement, et il est probable que ces défis soient surmontés dans les années à venir. Récemment, la technologie Single Cell s'est avérée efficace pour produire des données transcriptomiques sur les stades trophozoites. On peut s'attendre à ce que davantage de génomes de grégarines viennent enrichir les bases de données publiques et ainsi révéler toute la diversité des apicomplexes.

Cet effort d'échantillonnage devrait permettre de réexaminer la taxonomie des grégarines qui va probablement être bouleversée par l'apport d'informations moléculaires. En effet, un autre sujet de préoccupation est le décalage entre la diversité des grégarines actuellement documentée, et celle qui commence à émerger grâce aux approches moléculaires, notamment environnementales. Dans ce contexte, il est particulièrement difficile d'estimer l'ampleur de l'échantillonnage nécessaire pour documenter correctement toute la diversité des grégarines. Les grégarines constituent un groupe hétérogène de parasites apicomplexes infectant une très grande diversité d'hôtes non vertébrés. La biodiversité des grégarines correspond actuellement à 1600-1700 espèces formellement décrites, mais

selon les experts du domaine, ce nombre pourrait être largement sous-estimé. De récentes études métagénomiques des sols terrestres et des environnements marins ont confirmé la présence et l'abondance de séquences de type grégarine dans ces environnements, qui restent cependant difficilement attribuables à des espèces formellement décrites.

Traditionnellement, la délimitation d'espèces de grégarines est basée sur des combinaisons de caractéristiques morphologiques et comportementales, y compris le mode de vie des parasites (spécificités de l'hôte et de l'aire de répartition de l'hôte), la localisation des parasites dans les hôtes (c.-à-d. intestin ou coelome, plus rarement intracellulaire), la description des stades de développement du cycle de vie (mesures morphologiques, durée des stades, microscopie SEM et TEM), l'appariement des gamontes (frontal, latéral, caudo-frontal) et les modes de déhiscence des gamétocystes. Au cours des dernières décennies, la prise en compte croissante des données moléculaires a permis de confirmer mais aussi parfois de réviser la représentation taxonomique et phylogénétique que nous avons des grégarines, révélant des synonymes pour certaines espèces autrefois considérées comme distinctes ou inversement, permettant d'identifier de nouvelles espèces cryptiques, c'est-à-dire morphologiquement indiscernables mais clairement distinctes au niveau moléculaire. Dans certains cas, des espèces auparavant décrites comme des grégarines ont été replacées dans d'autres groupes taxonomiques, après le séquençage de marqueurs moléculaires et la réalisation d'analyses phylogénétiques.

Dans le chapitre 3, deux exemples de délimitation d'espèces de grégarines sont présentés : l'un concerne la grégarine marine *P. cf. gigantea* pour laquelle nous avons assemblé deux génomes, comme exposé dans le chapitre 2. L'autre réévalue la taxonomie des grégarines terrestres infectant deux espèces de criquets : *G. garnhami* infectant *Schistocerca gregaria* et *G. acridiorum* infectant *Locusta migratoria*. Les deux études démontrent, dans des contextes et par des moyens différents, la co-infection par deux espèces cryptiques du homard européen *Homarus gammarus*, d'une part, et l'existence de différentes grégarines là où l'on pensait auparavant qu'une seule espèce infestait deux espèces de criquets, *Locusta migratoria* et *Schistocerca gregaria*, d'autre part. Ces résultats ont été permis par une analyse taxonomique intégrative minutieuse, associant des données morphologiques et moléculaires. Ils illustrent l'ampleur des révisions taxonomiques à venir concernant les grégarines, et la nécessité de se tourner vers des marqueurs moléculaires, et probablement à l'échelle génomique, pour évaluer correctement la diversité de ces organismes. La délimitation d'espèces de microorganismes unicellulaires ne peut plus s'appuyer sur les seules informations morphologiques, mais doit inclure les données moléculaires dans une approche taxonomique intégrative. Les données présentées ici confirment que la plupart des différences morphologiques et morphométriques ne permettent pas de délimiter de manière concluante des espèces étroitement apparentées, alors que les données moléculaires peuvent révéler des différences clairement mesurables.

Dans le premier exemple, concernant *P. cf. gigantea* A et B, aucun caractère mor-

phologique permettant de discriminer sans ambiguïté les deux espèces n'a pu être mis en évidence, alors qu'elles sont clairement distinctes au niveau moléculaire. En effet, en dépit d'une séquence 18S SSU rDNA presque identique (1bp de différence pour 1694 positions, soit une divergence de 0,05%), ces grégarines présentent une divergence nucléique globale de 10% au niveau génomique, plaidant en faveur d'espèces différentes. D'autre part, la deuxième étude démontre que *S. gregaria* est infecté par *G. garnhami*, alors que *L. migratoria* est infecté par *G. acridiorum*. Dans ce cas, deux caractéristiques morphologiques discriminantes ont été mises en évidence : la taille respective des oocystes de *G. garnhami* et de *G. acridiorum*, mais aussi la localisation de leurs formes trophozoïtes dans l'intestin de leur hôte respectif. Ces traits conservent cependant une certaine ambiguïté (plasticité phénotypique potentielle), et bien qu'ils appuient les données moléculaires, ce sont ces dernières qui permettent de caractériser des différences précises et mesurables au niveau du marqueur moléculaire 18S. En ce sens, de nouvelles études moléculaires sont cruciales pour déterminer la diversité des espèces de grégarines qui infectent les acridiens, au-delà des spécificités morphologiques. Néanmoins, il faut garder à l'esprit que ces études ne peuvent se contenter de la recherche d'un seul marqueur moléculaire, tel que le marqueur 18S qui est actuellement le seul disponible pour une variété de grégarines. On sait maintenant que ce marqueur, en plus d'être mal adapté à la délimitation des espèces au sein des protistes, n'est pas suffisant pour rendre compte de l'ampleur de la divergence génétique entre deux espèces. L'étude des génomes de *P. cf. gigantea* démontrent l'insuffisance de ce seul marqueur pour statuer sur la délimitation des espèces, et montre qu'une approche à l'échelle génomique est probablement nécessaire pour discriminer les espèces cryptiques, qui semblent nombreuses au sein des grégarines, comme le montre l'exemple des grégarines Orthoptera.

Après avoir présenté les origines des deux génomes *P. cf. gigantea* A et *P. cf. gigantea* B dans le chapitre 2, et décrit dans le chapitre 3 l'analyse de taxonomie intégrative qui a permis de décrire les espèces de grégarines qui leur correspondent de la manière la plus précise possible à l'heure actuelle, le chapitre 4 est consacré à la description détaillée de ces deux génomes de grégarines marines et à la contribution qu'ils apportent à l'enrichissement de nos connaissances sur les Apicomplexa. Dans une première partie de résultats, les caractéristiques structurelles des deux génomes sont présentées, et comparées à celles d'une sélection de génomes de diverses espèces d'apicomplexes. L'enjeu principal a consisté à montrer les spécificités de ces deux génomes de grégarines marines par rapport à ce qui est actuellement connu des génomes d'apicomplexes, et notamment en les comparant au génome de *G. niphandrodes*, le génome d'une grégarine terrestre dont les données non publiées sont cependant déposées publiquement. Ces deux génomes de grégarines marines sont les premiers à être véritablement décrits et prochainement publiés pour l'ensemble des grégarines et serviront de référence pour ce groupe dans toutes les analyses futures consacrées à la génomique comparative des apicomplexes.

Comme démontré dans le chapitre 3, la biodiversité des grégarines est largement sous-estimée, car le grand nombre d'espèces morphologiquement décrites contient probablement de nombreuses espèces cryptiques détectables uniquement au niveau moléculaire, parfois seulement au niveau génomique. Il est donc possible que l'étude des génomes de grégarines se révèle capable d'éclairer l'histoire évolutive des apicomplexes, mais aussi de mieux comprendre les mécanismes qui leur confèrent, au niveau moléculaire, les capacités adaptatives nécessaires pour assurer la pluralité de leur mode de vie parasitaire. L'une des principales caractéristiques des apicomplexes est leur capacité à traverser les tissus de leurs hôtes jusqu'à atteindre leur cellule cible, en s'y attachant de manière extra ou épi-cellulaire ou en l'envahissant complètement de manière intracellulaire. C'est pourquoi la première analyse fonctionnelle d'une structure clé des apicomplexes a été consacrée à l'exploration des protéines du glidéosome, présentée dans une seconde partie de résultats. Le glidéosome est une structure moléculaire complexe à l'origine d'un mouvement appelé *gliding*, une forme de motilité emblématique du groupe des Apicomplexa et essentiellle à la manifestation de leur pathogénicité. Si cette structure est actuellement bien décrite chez *T. gondii* et *P. falciparum*, on ne sait rien des protéines impliquées dans le *gliding* chez les grégarines, qui ont par ailleurs la particularité de présenter d'autres types de mouvements comme le *bending* ou le *rolling*. Ainsi, une étude détaillée des protéines impliquées dans cette structure moléculaire a été entreprise, afin d'évaluer leur conservation à l'échelle des apicomplexes d'une part, mais aussi et surtout de questionner la pertinence du modèle du glidéosome appliqué aux grégarines, et à défaut, d'imaginer des alternatives possibles en fonction de ce que nous savons de la biologie spécifique des grégarines.

Ces premières investigations des premiers génomes de grégarines marines mettent en évidence, d'une part, une diversité moléculaire d'une ampleur insoupçonnée, et d'autre part, elles offrent la possibilité de remettre en question les connaissances acquises sur les apicomplexes en tenant compte de données issues d'une lignée jusqu'ici négligée. Bien que la grande diversité des génomes apicomplexes soit déjà connue, que ce soit en termes de taille du génome, de nombre de gènes codants ou de proportion de gènes non codants, les grégarines étaient jusqu'à récemment considérées comme un groupe taxonomique contenant des organismes proches et sans grand intérêt. Pourtant, leur capacité à parasiter la plupart, sinon la totalité, des métazoaires invertébrés aurait dû faire envisager une diversité moléculaire importante, qui fournirait les ressources pour s'adapter à des contextes extrêmement variés. Ainsi, ces premières analyses permettent de commencer à mesurer l'ampleur de cette diversité. Les génomes de seulement trois grégarines, deux marines et une terrestre, sont maintenant connus et déjà leur patrimoine génétique apparaît comme très divergent, quand bien même elles font partie d'un même groupe, les Eugregarines. Mais qu'en est-il de la diversité génomique des deux autres groupes de grégarines : les archigrégarines, qui comprennent des espèces considérées comme basales à toutes les grégarines, et probablement polyphylétiques, et les néogrégarines, qui rassemblent des espèces

considérées comme plus récemment dérivées ?

Au-delà des caractéristiques structurales de ces génomes et de la diversité de leur potentiel codant, il est essentiel de documenter les architectures moléculaires de fonctions ou structures clés de la biologie des Apicomplexa. Les protéines impliquées dans le *gliding* ont été les premières examinées, menant à la mise en évidence de leur conservation différentielle à l'échelle de l'Apicomplexa. Qu'en est-il des protéines du complexe apical, et de l'arsenal de protéines impliquées dans l'invasion et la sortie de la cellule hôte par les parasites intracellulaires ? Ces protéines sont-elles conservées chez les grégarines ? Si la recherche de protéines déjà identifiées dans des lignées apicomplexes bien documentées semble être la priorité, se pose la question beaucoup plus complexe de toutes les protéines spécifiques aux grégarines, et de leur implication dans des structures ou des comportements qui leur sont propres, comme les mouvements alternatifs au gliding précedemment évoqués. L'abîme de connaissances à découvrir justifie pleinement le décryptage d'autres génomes de grégarines, car de telles études ne sauraient être menées à bien sans données de référence. Les deux génomes de la grégarine marine *P. cf. gigantea*, s'ils constituent une première étape, ne permettront pas à eux seuls de rendre compte de la diversité des grégarines. De nombreux génomes de référence, à l'échelle des multiples groupes de grégarines documentés, sont nécessaires. À la lumière de ces premiers résultats, de nombreuses questions se posent sur l'histoire évolutive des grégarines, et des apicomplexes dans leur ensemble. Les parasites hautement pathogènes tels que les espèces *T. gondii* ou *Plasmodium* ont été très bien documentés par la communauté scientifique au niveau moléculaire, mais il est temps de reconnaître que ces espèces ne représentent qu'une petite partie de la diversité des apicomplexes. En outre, les connaissances acquises sur ces parasites bien connus doivent pouvoir être questionnées afin de ne pas les généraliser trop hâtivement à tous les apicomplexes. Même s'il est légitime d'un point de vue médical d'avoir concentré les efforts scientifiques sur les parasites hautement pathogènes, il faut désormais veiller à ne pas laisser dans l'ombre les connaissances que les grégarines sont susceptibles de nous apporter sur les Apicomplexa.

Mots-clefs

Apicomplexa, grégarine marine, grégarine terrestre, assemblage génomique, prédiction de gènes, taxonomie intégrative, phylogénie moléculaire, génomique comparative, gliding, évolution du parasitisme

Abstract

Apicomplexan are unicellular eukaryotic microorganisms that have evolved towards strict parasitic lifestyle. Some apicomplexan groups include species that cause serious pathologies such as malaria (*Plasmodium* ssp.), toxoplasmosis (*Toxoplasma gondii*) and cryptosporidiosis (*Cryptosporidium* spp.). While the genomes of these highly pathogenic agents are now well documented, this is not the case for other apicomplexan lineages such as gregarines, which are considered basal within the Apicomplexa, have low pathogenicity and above all are non-cultivable. Their molecular study currently represents a major bottleneck, whereas a precise knowledge of their genomes would be essential to better understand the evolutionary history of apicomplexan parasites and the diversity of their adaptive paths to parasitic lifestyle. During this thesis the genome caracterisation of 2 marine gregarines, *Porospora gigantea*, parasite of the European lobster *Homarus gammarus* and *Diplauxis hatti*, parasite of the Polychaeta marine worm *Perinereis cultrifera*; and 1 terrestrial gregarine, *Gregarina acridiorum*, parasite of the locust *Locusta migratoria* have been carried out. The discovery of two coexisting genomes matching the morphologically described species *P. gigantea*, along with another example involving *G. acridiorum* illustrates the magnitude of the upcoming taxonomic revisions, and the need to turn to molecular markers, likely on a genomic scale, to properly assess the diversity of gregarines. Furthermore, the first comparative genomics analyses including gregarines reveal their unsuspected genetic diversity across Apicomplexa. An apicomplexan scale analyses of the glideosome proteins was also performed. This model refers to a complex molecular structure at the origin of gliding, a signature movement of Apicomplexa that is essential for the manifestation of their pathogenicity. A detailed comparative analysis highlights its differential conservation at the apicomplexan scale, suggesting a diversity of adaptations to motility and host cell invasion issues. This study illustrates the importance of considering non-model, non-pathogenic, non-cultivatable apicomplexan to provide novel clues to the adaptive capabilities displayed by this ecologically and medically major group of parasites.

Keywords

Apicomplexa, marine gregarine, terrestrial gregarine, genome assembly, gene prediction, integrative taxonomy, molecular phylogeny, comparative genomics, gliding, evolution of parasitism