Elsevier Editorial System(tm) for International Journal for Parasitology
Manuscript Draft

Manuscript Number: IJPara15_144R2

Title: Two recently sequenced vertebrate genomes are contaminated with apicomplexan species of the Sarcocystidae family

Article Type: Full Length Article

Keywords: Apicomplexa, apicoplast, Sarcocystis, apicortin, host, phylogenetic tree, bat, Nephroisospora

Corresponding Author: Prof. Ferenc Orosz, PhD

Corresponding Author's Institution:

First Author: Ferenc Orosz, PhD

Order of Authors: Ferenc Orosz, PhD

Manuscript Region of Origin: HUNGARY

Abstract: The paper enlightens a general problem, namely, that host genomes can easily be contaminated with parasite ones thus careful isolation of genetic material and careful bioinformatics analysis are needed in all cases. I show for two recently published genomes that they are contaminated with apicomplexan parasites which belong to the Sarcocystidae family. Sequences of the characteristic apicomplexan organelle, apicoplast, were used as queries in BLASTN search against nucleotide sequences of various animal groups looking for possible contaminations. I found that the draft genomes of a bird, Colinus virginianus (Halley et al., 2014, PLoS ONE 9, e90240), and a bat, Myotis davidii (Zhang et al., 2013, Science, 339, 456-460.) contain at least 6 and 17 contigs, respectively, originated from the apicoplast of an apicomplexan species, and other genes specific of this phylum can also be found in the published genomes. Obviously, the sources of the genetic material, the muscle and the kidney of the animals, respectively, contained the parasitic cysts. Phylogenetic analyses using 18S ribosomal RNA and internal transcribed spacer 1 genes show that the parasite contaminating C. virginianus is a species of Sarcocystis related to ones known to cycle between avian and mammalian hosts; in the case of M. davidii it belongs to the Nephroisospora genus, the only member of which, N. eptesici, has been recently identified from the kidney of big brown bats.

Cover Letter

A. Loukas
Editor-in-Chief
International Journal for Parasitology
Queensland Tropical Health Alliance, Fac. of Medicine, Health & Molecular Sciences,
James Cook University, McGregor Rd, Smithfield, Cairns, Qld 4878, Queensland, Australia

Dear Professor Loukas,

I would like to resubmit a research article entitled "**Two recently sequenced vertebrate genomes are contaminated with apicomplexan species of the Sarcocystidae family**" to your journal for considering of publication. It is the revised version of my manuscript, IJPara15_144.

Budapest, July 9, 2015

Sincerely yours

Ferenc Orosz, PhD
Institute of Enzymology
Research Centre for Natural Sciences,
Hungarian Academy of Sciences

*Response to Reviewers

Jan Slapeta

Deputy Editor

International Journal for Parasitology

Dear Editor,

Thank you for your favourable criticism. Here are my answers to you and to the Reviewers.

Answers to the Editor:

"Authors must remove any reference to a formal new name."

I removed it from the abstract and the main text, too.

"Reviewer #3 provided an annotated manuscript that should be used to prepare the further revised version."

I used it and accepted his suggestions without exception.

Answers to Reviewer 2

"I think merging the manuscripts was a good idea, and think the result is acceptable."

Thank you for your favourable criticism.

Answers to Reviewer 4

"I think that this paper is an interesting one, and should be published in the International Journal for Parasitology."

Thank you for your favourable criticism.

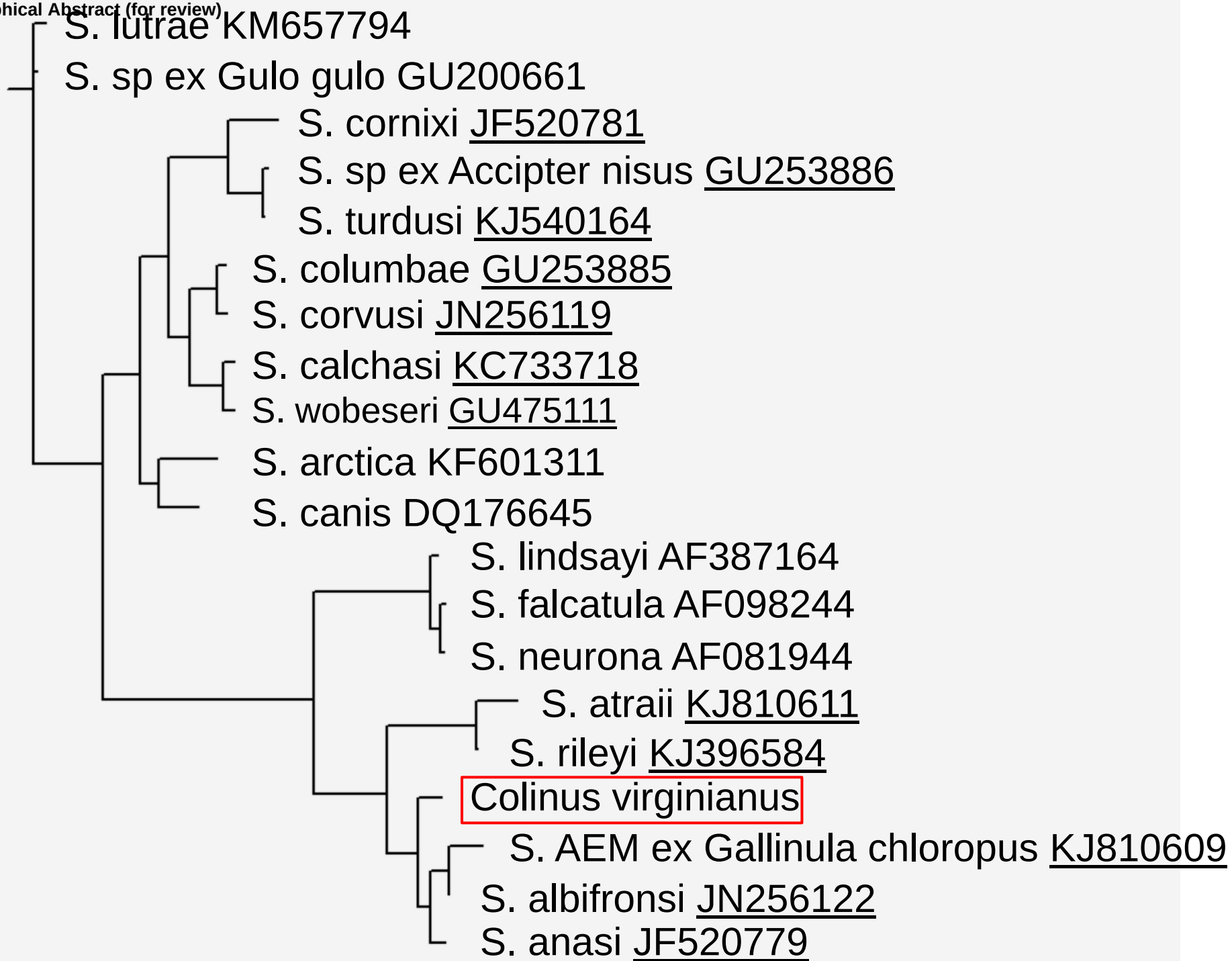
"I have only one substantial suggestion. At the end of the manuscript the author says: "which can be named tentatively as Nephroisospora myotisi." I think that the author should delete this phrase, which also occurs in the Abstract. This could be interpreted as formally naming the species, which should not be done without considerably more information than is provided here."

I removed it from the abstract and the main text, too.

I have made several suggestions for improving the clarity of the English expression.

Special thanks for the corrections of the grammatical and style shortcomings. Of course, all of them were accepted.

*Graphical Abstract (for review)



Highlights

The draft genome of *Colinus virginianus* is contaminated by a *Sarcocystis* species

The parasite is related to ones known to cycle between avian and mammalian hosts

The draft genome of *Myotis davidii*, a bat, is contaminated by a Sarcocystidae species

The parasite is the second (tentative) member of the *Nephroisospora* genus

Several contigs of the hosts originated from the apicomplexan organelle, apicoplast

Two recently sequenced vertebrate genomes are contaminated with apicomplexan species of the Sarcocystidae family

Ferenc Orosz

Institute of Enzymology, Research Centre for Natural Sciences, Hungarian Academy of Sciences

Address:

Ferenc Orosz, Ph.D., D.Sc.
Institute of Enzymology,
Research Centre for Natural Sciences,
Hungarian Academy of Sciences
Magyar tudósok körútja 2.
H-1117 Budapest, Hungary
Phone: (36)-1-3826714

E-mail: orosz.ferenc@ttk.mta.hu

Note: Supplementary Data are associated with this article′

Abstract

The paper highlights a general problem, namely, that host genome sequences can easily be contaminated with parasite ones thus careful isolation of genetic material and careful bioinformatics analysis are needed in all cases. I show for two recently published genomes that they are contaminated with sequences of apicomplexan parasites which belong to the Sarcocystidae family. Sequences of the characteristic apicomplexan organelle, the apicoplast, were used as queries in BLASTN search against nucleotide sequences of various animal groups looking for possible contaminations. I found that the draft genomes of a bird, *Colinus virginianus* (Halley et al., 2014, PLoS ONE 9, e90240), and a bat, *Myotis davidii* (Zhang et al., 2013, Science, 339, 456-460.) contain at least 6 and 17 contigs, respectively, originating from the apicoplast of an apicomplexan species, and other genes specific to this phylum can also be found in the published genomes. Obviously, the sources of the genetic material, the muscle and the kidney of the animals, respectively, contained the parasitic cysts. Phylogenetic analyses using *18S ribosomal RNA* and *internal transcribed spacer 1* genes show that the parasite contaminating *C. virginianus* is a species of *Sarcocystis* related to ones known to cycle between avian and mammalian hosts. In the case of *M. davidii* it belongs to the *Nephroisospora* genus, the only member of which, *N. eptesici*, has been recently identified from the kidney of big brown bats.

Keynotes: Apicomplexa, apicoplast, *Sarcocystis*, apicortin, host, phylogenetic tree, bat, *Nephroisospora*

1. Introduction

The raw data from a genome sequencing project sometimes contains DNA from contaminating organisms, which may be introduced during sample collection or sequence preparation. In some instances, these contaminants remain in the sequence even after assembly and deposition of the genome into public databases. As a consequence, searches of these databases may yield erroneous and confusing results (Merchant et al., 2014). Human DNA is a common contaminant, from the scientists who handle the samples during the process of extraction through sequencing (Longo et al., 2011). Computational filters applied to the raw sequencing reads are usually effective at removal of human DNA and other common laboratory contaminants such as *E. coli*, but other contaminants may be more difficult to identify. In the present paper I highlight an additional problem, namely, that host genomes can easily be contaminated with parasite ones, and thus careful isolation of genetic material and careful bioinformatics analysis are needed in all cases.

Apicomplexan parasites cause serious illnesses in humans and domestic animals. Most members of the phylum Apicomplexa are obligate parasites, with some members being causative agents for diseases in vertebrates. Species in the genus *Plasmodium* cause malaria, from which over 1 million people die each year. Members of the apicomplexan families Babesiidae, Theileriidae, Eimeriidae, Sarcocistidae, and Cryptosporiidae are responsible for numerous infectious diseases in wild and domesticated animals, such as coccidiosis and babesiosis, resulting in significant economic burden for animal husbandry.

One of the apicomplexan families, belonging to the class Coccidia, is the Sarcocystidae. The members of the genera *Besnoitia*, *Hammondia*, and *Sarcocystis* have obligatory two-host predator-prey life cycle: asexual stages (sarcocysts) develop in the muscles of the intermediate hosts (prey); ingestion of muscle sarcocysts through predation or scavenging by the definitive host (predator) propagates the life cycle, and sexual multiplication takes place in its small intestine that results in sporocyst shedding in feces (Dubey et al., 1988). Other families of the Sarcocystidae, such as *Toxoplasma*, *Neospora*, and *Cystoisospora* can complete their life cycle using only one host (Wünschmann et al., 2010).

A new member of the Sarcocystidae family has been recently identified, namely, *Nephroisospora eptesici*, from the big brown bat (*Eptesicus fuscus*) (Wünschmann et al., 2010). It is the only known member

of the *Nephroisospora* genus, which is similar to *Besnoitia*, *Toxoplasma* and *Hammondia* species. Bats are known to host almost every kind of apicomplexan parasite (except gregarines which are known to parasitize only invertebrates): Sarcocystidae (Cabral et al., 2013; Dodd et al., 2014), Eimeriidae (McAllister et al., 2011; Afonso et al., 2014), Cryptosporidae (Wang et al., 2013), Haemosporidia (Duval et al., 2012; Schaer et al., 2013).

In searches of genomes and sequence data that have become recently available, I found that a sequence similar to a characteristic apicomplexan protein, apicortin (Orosz, 2009, 2011), is present in the whole genome shotgun (WGS) sequence of a bird, *Colinus virginianus* (bobwhite), the draft sequence of which has recently been published (Halley et al., 2014). This is a very surprising finding since in higher level (Eumetazoa) animals no apicortins have been found to date. It can be either the result of horizontal gene transfer, or it is, more probably, a contamination. Thus I decided to systematically investigate this problem: sequences of the characteristic apicomplexan organelle, the apicoplast, were used as queries in BLASTN search against nucleotide sequences of various animal groups, looking for possible contaminations. I found that, indeed, the latter case is valid; moreover, further vertebrate genomes are contaminated with apicomplexan sequences. Additionally, based on the contamination of a recently published bat genome (Zhang et al., 2013), I suggest the existence of a second member of the *Nephroisospora* genus hosted also by a bat, *Myotis davidii*.

2. Methods

2.1. Database similarity search and phylogenetic analysis

Accession numbers of protein and nucleotide sequences refer to the NCBI GenBank database. The database search was started with an NCBI blast search using the sequences of known apicortin proteins as queries. BLASTP or TBLASTN analyses (Altschul et al., 1997) were performed on protein or nucleotide sequences available at the NCBI website. Then the whole nucleotide sequences of known apicomplexan apicoplasts were used as queries. BLASTN analysis (Altschul et al., 1997) was performed on nucleotide sequences, including expressed sequenced tags, TSAs (Transcriptome Shotgun Assembly) and WGSs,

available at the NCBI website. In further analyses, the *18S ribosomal RNA* and the *internal transcribed spacer 1* (*ITS-1*) genes of Sarcocystidae were used as queries against the *C. virginianus* and *M. davidii* nucleotide sequences using BLASTN.

Multiple alignments of protein and nucleotide sequences were carried out by the Clustal Omega program (Sievers et al., 2011) and were manually refined. Multiple sequence alignments used for constructing phylogenetic trees are provided in Supplementary Data S1. The alignments were subjected to Bayesian phylogenetic analysis with the software MrBayes v.3.1.2 (Ronquist and Huelsenbeck, 2003). Default priors and the GTR model (Tavaré, 1986) including a proportion of invariant sites and a gamma-shaped distribution of variable sites with four rate categories (GTR+$\Gamma_{(4)}$+I) were used. Four chains were run up to $2.4x10^6$ generations, with a sampling frequency of 0.01, and the first 25% of the generations were discarded as burn-in. The tree was drawn using the program Drawgram.

The Phylip (Phylogeny Inference Package, version 3.696) program package (Felsenstein, 2008) was used to build a Maximum Likelihood (ML) tree with bootstrap values. One thousand datasets were generated using the program Seqboot from the original data, i.e. the multiple alignments done by Clustal Omega. This was followed by running the program Dnaml (DNA Maximum Likelihood) on each of the datasets in the group, using the same rate heterogeneity model as above ($\Gamma_{(4)}$+I). The values for the gamma distribution were taken from the Bayesian analysis. A consensus tree (from all the 1000 trees) was generated using the program Consense. The trees were drawn using the program Drawgram.

3. Results and Discussion

3.1. Apicortin is present in the C. virginianus WGS sequence

C. virginianus putative apicortin is very similar to the apicortins of the Sarcocystidae, *T. gondii, N. caninum* and *Hammondia hammondii* (Fig. 1). However, it shows the highest identity with a WGS sequence of *Sarcocystis neurona* (Fig. 1), the draft (not fully annotated) genome of which has been reported very recently (Blazejewski et al., 2015). *S. neurona,* an apicomplexan pathogen that cycles in nature between its

definitive host, Virginia opossum (*Didelphis virginiana*), and a broad range of mammals and birds (orders Passeriforme and Psittaciformes) as intermediate hosts, causes equine protozoal myeloencephalitis, a neurologic disease of horses (Dame et al., 1995). Although reports showing that *C. virginianus* is a host for *S. neurona* are not yet known, the geographical identity of its habitats with that of *D. virginiana*, the definitive host for *S. neurona*, makes it reasonable that similarly to other birds, *C. virginianus* is also an intermediate host of the parasite. (*C. virginianus* belongs to the order Galliformes, which are also known to be *Sarcocystis* hosts (Odening, 1998).) However, although the similarity between the *C. virginianus* and *S. neurona* potential apicortin-coding sequences is very high (the identity is above 90%), it is not high enough to suggest that the sequence found in *C. virginianus* genome is a contamination originated from *S. neurona*. Rather, it is probably a contamination from another species of the *Sarcocystis* genus, the genome of which has not been sequenced yet.

3.2. Genes of apicoplast origin are present in the C. virginianus WGS sequence

Most apicomplexan parasites possess an apicoplast, a plastid with no photosynthetic ability, which is essential for cell survival (Arisue and Hashimoto, 2015). The apicoplast has its own genome that mainly encodes the transcription and translation related genes necessary for plastid gene expression. There are six independent entries at the NCBI web page for the query "Sarcocystis + apicoplast" as nucleotide sequences, five of them for RNA polymerase beta subunit-like (RPOb) gene from various *Sarcocystis* species, and one of them for small subunit ribosomal RNA gene from *Sarcocystis muris*. Using these sequences as BLASTN queries against avian nucleotide sequences of the NCBI GenBank database, there are no hits but those of the *C. virginianus* WGS sequence. Any hit would be very improbable; however, the identities are extremely high, 94-98% (Table 1). Examples of hits are shown in Supplementary Data S2. All the hits are in duplicate, since Halley et al. (2014) produced a simple *de novo* (i.e. no scaffolding) and a scaffolded *de novo* assembly, and both of them were deposited in GenBank.

The complete genemap of several apicoplast genomes are available as listed in Arisue and Hashimoto (2015). These genomes commonly encode rRNAs, tRNAs, ribosomal proteins, bacterial-type RNA

polymerase subunits, EF-Tu and ClpC protein. The closest relative of *Sarcocystis* species is *T. gondii*, belonging to the same family, Sarcocystidae. Thus I carried out a BLASTN search in the NCBI GenBank databases against avian nucleotide sequences using the complete 35 kb sequence of *T. gondii* apicoplast (GenBank U87145). Again, there were no hits but those of the *C. virginianus* WGS sequence; moreover, with high identity values (Table 2, Supplementary Data S3).

The apicoplast genomes of *P. falciparum*, *E. tenella* and *T. gondii* have an inverted repeat region, which contains duplicated SSU and LSU rRNA genes as well as duplicated tRNA genes (Arisue and Hashimoto, 2015). Genes are bi-directionally encoded in the genome, with half of the circle being mainly transcribed in a clockwise direction and the other half generally transcribed counter clockwise. The presence of the inverted repeat region is the reason that the "Total score" is about twice the "Max. score" in the first two lines of Table 2: both the clockwise and counter clockwise *T. gondii* sequences give hits on the *C. virginianus* contigs in Plus/Minus and Plus/Plus directions (cf. AWGU01003450.1 and AWGT01002297.1 in Supplementary Data S3).

Although there are no available complete apicoplast sequences for *Sarcocystis* spp. in the database yet, a significant part of the *S. neurona* apicoplast sequence can be identified by using the *T. gondii* apicoplast sequence as a query. The JAQE01002351.1 (24 001 bp) of the *S. neurona* WGS sequence covers about two-thirds of the *T. gondii* apicoplast sequence, while the JAQE01002350.1 (5698 bp) corresponds to the inverted repeat region. Using these sequences as queries against avian nucleotide sequences, only the *C. virginianus* WGS sequence produced hits, with even higher cover and identity values than the *T. gondii* query (Table 3, Supplementary Data S4, 5). Beside the contigs identified in Tables 1 and 2 (DeNovo_contigs 9595, 3450, 45101, 73205), further hits were found; the full sequences of the six contigs, where the score is higher than 1000, seem to originate from *Sarcocystis* contamination (cf. Supplementary Data S4, 5).

3.3. A general method for fast identification of genome contaminations of apicomplexan origin

The available apicoplast genomes include *T. gondii* (GenBank U87145; KE138841), *E. tenella* (Cai et al., 2003), *Theileria parva* (Gardner et al., 2005), *B. bovis* (Brayton et al., 2007), *Leucocytozoon caulleryi*

(GenBank NC_022667), *Plasmodium falciparum* (Wilson et al., 1996) and several other *Plasmodium* species. Thus, I carried out a BLASTN search (Altschul et al., 1997) in the NCBI GenBank databases against nucleotide sequences (including ESTs, TSAs and WGSs) of various animal groups (mammals, birds, insects etc.) using the complete 29-35 kb sequences of the above mentioned apicoplasts as queries, looking for possible contaminations. As expected, there were very few hits; they included, beside *C. virginianus*, several contigs of the *M. davidii* and *Ovis aries musimon* WGS sequences; moreover, with high identity values. The highest identities were found when the *T. gondii* apicoplast was used as a query. In the case of the unfinished genome of *O. aries musimon* (contig 124943, 5098 bp, GenBank CBYI010124943.1), it was obvious that the source of the contamination is *T. gondii* since the sequence identity is practically 100% (Supplementary Data S6). For *M. davidii*, the 17 contigs with 299–4986 bp length are shown in Table 4 and Supplementary Data S7. Since *T. gondii* is a member of the Sarcocystidae family of coccidian parasites, it is a reasonable suggestion that the source of contamination found in *M. davidii* also belongs to this family.

3.4. Which apicomplexan species may cause the contaminations?

The question arises whether the Sarcocystidae species causing the contamination of the bird and bat genomes can be identified. Unfortunately, no complete *Sarcocystis* genomes except that of *S. neurona* have been sequenced; however, the sequences of two characteristic genes, often used for phylogenetic analysis, are known in many cases. These are the *18S ribosomal RNA* and the *internal transcribed spacer 1* (*ITS-1*) genes. I used these genes of *Sarcocystis* species as queries in BLASTN search against the *C. virginianus* WSG sequences, and identified the best hits. The *18S ribosomal RNA* gene is rather conservative; e.g. other avian sequences, including *C. virginianus* itself, show higher than 70% identity with the *Sarcocystis* genes. However, there are no hits above 80% identity, and thus the difference is high enough to differentiate between contamination and endogenous sequences. On the contrary, the *ITS-1* gene differs much more among the various species.

In the other case (*M. davidii*), not the genus but only the family can be identified on the basis of the BLAST search using apicoplast sequences as queries. Thus the highly divergent *ITS-1* sequences cannot be used for search, and only the *18S ribosomal RNA* genes are suitable for the analysis.

Table 5 and Supplementary Table S1 show the identity of the *18S ribosomal RNA* gene of various *Sarcocystis* species with the best hit in the *C. virginianus* WGS sequence. (The hits can be found on the AWGU01001200 and AWGT01000758 contigs.) The identity is always higher than 90%, and in some cases higher than 99%. The highest identity values are produced, not surprisingly, by *Sarcocystis* species of avian hosts (Table 5). The highest identities (99.8%) were found in the case of *S. albifronsi* (GenBank: EU502868.2) and *S. anasi* (GenBank: EU553477.2). The difference is only 3 nucleotides in both cases. The third most similar sequence is that of the *S. rileyi* (GenBank: KJ396583.1; 99.7%), where the difference is 6 nucleotides. It is worth noting that the sources of the isolation, the intermediate hosts of the parasites, belong to the order Anseriformes (e.g. ducks and geese) in these three cases. Anseriformes and Galliformes form the clade Galloanserae. *C. virginianus* belongs to Odontophoridae, which is one of the families of Galliformes. Although there are examples where a species belonging to Galliformes hosts a *Sarcocystis* species (Wenzel et al., 1982; Odening, 1998; Chen et al., 2012), there are no sequence data available in these cases. Thus Anseriformes are the closest relatives of *C. virginianus* for which these data can be used.

Much fewer hits were found when the *ITS-1* genes were used as queries. Besides *S. albifronsi* (GenBank: JN256122.1; 87%) and *S. anasi* (GenBank: JF520779.2; 88%) only a sequence of *Sarcocystis* AEM-2014a (GenBank: KJ810609.1; 80%), hosted by *Gallinula chloropus*, was recognised in the BLASTN search as similar to the *C. virginianus* sequence, the 4300-3367 nucleotides of AWG1001200.1 SimpleDeNovo_contig_1200. When I used this corresponding *C. virginianus* sequence as query in BLASTN search (in general, not only against Sarcocystidae nucleotide sequences), only the above mentioned three *Sarcocystis* nucleotide sequences were obtained as hits.

A phylogenetic tree of *Sarcocytis* spp was constructed by Bayesian analysis using the available *ITS-1* nucleotide sequences, with *S. lutrae* sequence as an outgroup (Fig. 2). The corresponding nucleotide sequence of SimpleDeNovo_contig_1200 of *C. virginianus* was also involved. The tree fits to the known phylogeny (Kutkiené et al., 2012; Prakas et al., 2013); e.g., *S. wobeseri* is sister to *S. calchasi*, they are sister to *S. columbae* and all of them are sister to *S. cornixi* + *S. sp* ex *Accipter nisus*. Similarly, it was also found

that *S. anasi* and *S. albifronsi* are sister to *S. riley*, and they are sister to (*S. neurona* +*S. falcatula*) + *S. lindsayi* whose definitive hosts are opossums, while a wide range of birds are known as intermediate hosts (cf. Fig. 2). (It was suggested by Dame et al (1995) that *S. neurona* is identical with *S. falcatula*. The sequence identity of the *ITS-1* region is 97.3%, one of the highest values of the known sequences.) The position of the sequence contaminating *C. virginianus* is clearly within the clade represented by the above mentioned Anseriformes-hosted *Sarcocystis* species (*S. anasi*, *S. albifronsi*, *S. riley*). Both the other two species in this clade (*S. atraii* and *S. AEM-14a*) were isolated from hosts belonging to the order Gruiformes.

There are at least 30 *Sarcocystis* species that are known to be hosted by birds (Odening, 1998; Kutkiené et al., 2012**)**. Granivorous, insectivorous, and omnivorous birds serve as intermediate hosts, and carnivorous birds are usually definitive hosts. Most of the avian orders are known to be infected by *Sarcocystis* species. Unfortunately, the majority of investigations into bird *Sarcocystis* lack molecular data or are characterized to the genus only ("*Sarcocystis* sp.") (Kutkiené et al., 2012**)**. As Fig. 2 shows, in accordance with the results of others (Kutkiené et al., 2012; Prakas et al., 2013), *Sarcocystis* species from birds form two groups. In one of them, the "upper group" on Fig. 2, predatory birds are thought to be definitive hosts (Olias et al., 2011; Prakas et al., 2013). In the "lower group" with *S anasi* and others, mammalian definitive hosts are known: Arctic fox (*Alopex lagopus*) for *S. albifronsi* (Kutkiené et al., 2012); and striped skunk (*Mephitis mephitis*), red fox (*Vulpes vulpes*) and raccoon dogs (*Nyctereutes procyonoides*) for *S. rileyi* (Odening, 1998; Prakas et al., 2015). Thus, the *Sarcocystis* species hosted by *C. virginianus* as intermediate host probably has a mammalian definitive host as well. It is known that some *Sarcocystis* species are not host specific and can parasitize a wide range of hosts. For example, *S. neurona* is hosted not only by horses and various birds as originally thought (Dame et al. 1995), but its host range expands to raccoons, cats, skunks, a variety of mustelids, pinnipeds, cetaceans and more recently sea otters, harbour seals, and harbour porpoises (Blazejewski et al., 2015). In the case of bird-hosted parasites, *S. wobeseri* is hosted by mallard duck (*Anas platyrhynchos*) and white-fronted goose (*Anser albifrons*) from the order Anseriformes as well as by herring gull (*Larus argentatus*) from the order Charadriiformes (Kutkiené et al., 2010; Prakas et al., 2011). Thus it could be possible that a known *Sarcocystis* species is shared by *C. virginianus* and another bird of the order Galliformes or even by a bird of Anseriformes, Gruiformes or Charadriiformes. However, we can state at this moment only that there is no *Sarcocystis* species with a

known genome or partial sequence which can be identified as being responsible for the contamination of the *C. virginianus* genome. The contamination could be due to the fact that for isolation of the genomic DNA, skeletal muscle derived from the legs of a bobwhite was utilized (Halley et al., 2014), and sarcocysts develop in the muscles of the intermediate hosts.

3.5. In silico identification of the second member of the *Nephroisospora* genus

In the case of *M. davidii*, using the *18S ribosomal RNA* genes of Sarcocystidae as queries against the *M. davidii* WGS sequence, the hits can be found on contig111512 (GenBank: ALWT01111512.1). The identity is always higher than 95%, and in one case it is even higher than 99%. The highest identity (99.3%; 14 bp difference) was found in the case of *N. eptesici* (GenBank: EU334134.1).

The identities of the *18S ribosomal RNA* genes of various Sarcocystidae species with the 211-2019 nucleotides of contig111512 of the *M. davidii* WGS sequence are shown in Table S6. The genes listed there were used for the construction of a phylogenetic tree of *18S ribosomal RNA* genes of Sarcocystidae family by Bayesian and maximum likelihood analysis involving some Eimeriidae genes and using *Goussia balatonica* as outgroup (Fig. 3). The corresponding nucleotide sequence of contig111512 of *M. davidii* was also involved. (I used two phylogenetic methods since some branches of the tree were weakly supported.) In accordance with previous phylogenies (Morrison et al., 2004; Wünschmann et al., 2010; Matsubayashi et al., 2011), the Sarcocystidae genes are well-separated from Eimeriidae (*Eimeria, Lankesterella, Goussia*) genes; and within the Sarcocystidae clade *Sarcocystis* (including *Frenkelia*) is sister to all other genera in both trees. The *Cystoisospora/Isospora* (Barta et al., 2005) and the *Besnoitia* genera as well as the *Toxoplasma*/*Hammondia*/*Neospora* group form evidently three different clades, but their mutual relationships are not clear since this part of the tree is weakly supported in the cases of both methods. (According to the above mentioned three analyses, *Cystoisospora/Isospora* forms the sister to *Besnoitia* + *Toxoplasma*/*Hammondia*/*Neospora*.) However, most importantly, the sequence from *M. davidii* forms a clade with the *N. eptesici* gene in both trees with maximal support. *N. eptesici* is the only known member of the *Nephroisospora* genus, from the most common bat, *E. fuscus* (Wünschmann et al., 2010). It was found in

297 the kidney of the big brown bat and it was concluded that it is closely related to *Besnoitia*, *Hammondia*,
298 *Neospora* and *Toxoplasma*. Similarly to *Toxoplasma* and *Neospora*, it can complete its life cycle using only
299 one host. Moreover, the entire cycle is completed in the kidney of the single host (Wünschmann et al., 2010).
300 The source of the genetic material used for genome sequencing of *M. davidii* was not given by Zhang et al.
301 (2013). However, in the GenBank web-page
302 (http://www.ncbi.nlm.nih.gov/Traces/wgs/?val=ALWT&display=contigs&search=ALWT01000000), where
303 the nucleotide sequences were deposited by the authors of the paper, it is stated that the kidney (beside
304 spleen and small intestine) of the animal was used. The only nucleotide sequence established for *N. eptesici*
305 is the *18S ribosomal RNA* gene, and thus no more data are available for comparison. However, the
306 phylogenetic position of the tentative species, the bat host and the renal localization of the parasite make it
307 reasonable to suggest that the contamination found in *M. davidii* originates from a new species of the
308 *Nephroisospora* genus.

309 Finally, I hope that this paper highlights the general problem, namely, that hosts genomes can be
310 easily contaminated with parasite ones thus careful isolation of genetic material and careful bioinformatic
311 analysis are needed in all cases.

312

313 **Acknowledgments:** The author thanks Dr. Judit Oláh for the careful reading of the manuscript as well as the
314 unknown reviewers of the paper for their suggestions.

315

316 **References**

317

318 Afonso, E., Baurand, P.E., Tournant, P., Capelli, N.. 2014. First amplification of *Eimeria hessei* DNA from
319 the lesser horseshoe bat (*Rhinolophus hipposideros*) and its phylogenetic relationships with *Eimeria*
320 species from other bats and rodents. Exp. Parasitol. 139, 58-62. doi: 10.1016/j.exppara.2014.02.013.

321 Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped
322 BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res.
323 25, 3389–3402.

324 Arisue, N., Hashimoto, T., 2015. Phylogeny and evolution of apicoplasts and apicomplexan parasites.
325 Parasitol. Int. 64, 254-259. doi: 10.1016/j.parint.2014.10.005.

326 Barta, J.R., Schrenzel, M.D., Carreno, R., Rideout. B.A., 2005. The genus *Atoxoplasma* (Garnham 1950) as a
327 junior objective synonym of the genus *Isospora* (Schneider 1881) species infecting birds and
328 resurrection of *Cystoisospora* (Frenkel 1977) as the correct genus for *Isospora* species infecting
329 mammals. J. Parasitol. 91, 726–727.

330 Blazejewski, T., Nursimulu, N., Pszenny, V., Dangoudoubiyam, S., Namasivayam, S., Chiasson, M.A.,
331 Chessman, K., Tonkin, M., Swapna, L.S., Hung, S.S., Bridgers, J., Ricklefs, S.M., Boulanger, M.J.,
332 Dubey, J.P., Porcella, S.F., Kissinger, J.C., Howe, D.K., Grigg, M.E., Parkinson, J., 2015. Systems-
333 based analysis of the *Sarcocystis neurona* genome identifies pathways that contribute to a
334 heteroxenous life cycle. MBio. 6, e02445-14. doi: 10.1128/mBio.02445-14.

335 Brayton, K.A., Lau, A.O., Herndon, D.R., Hannick, L., Kappmeyer, L.S., Berens, S.J., Bidwell, SL., Brown,
336 W.C., Crabtree, J., Fadrosh, D., Feldblum, T., Forberger, H.A., Haas, B.J., Howell, J.M., Khouri, H.,
337 Koo, H., Mann, D.J., Norimine, J., Paulsen, I.T., Radune, D., Ren, Q., Smith, R.K. Jr., Suarez, C.E.,
338 White, O., Wortman, J.R., Knowles, D.P. Jr., McElwain, T.F., Nene, V.M., 2007. Genome sequence
339 of *Babesia bovis* and comparative analysis of apicomplexan hemoprotozoa. PLoS Pathog. 3, 1401–
340 1413. doi:10.1371/journal.ppat.0030148

341 Cabral, A.D., Gama, A.R., Sodré, M.M., Savani, E.S., Galvão-Dias, M.A., Jordão, L.R., Maeda, M.M., Yai,
342 L.E., Gennari, S.M., Pena, H.F., 2013. First isolation and genotyping of *Toxoplasma gondii* from bats
343 (Mammalia: Chiroptera).Vet. Parasitol. 193, 100-104. doi: 10.1016/j.vetpar.2012.11.015.

344 Cai, X., Fuller, A.L., McDougald, L.R., Zhu, G., 2003. Apicoplast genome of the coccidian *Eimeria tenella*.
345 Gene 321, 39–46.
346 Chen, X., He, Y., Liu, Y., Olias, P., Rosenthal, B.M.., Cui, L., Zuo, Y., Yang, Z., 2012. Infections with
347 *Sarcocystis wenzeli* are prevalent in the chickens of Yunnan Province, China, but not in the flocks of
348 domesticated pigeons or ducks. Exp. Parasitol. 131, 31–34.
349 Dame, J.B., MacKay, R.J., Yowell, C.A., Cutler, T.J., Marsht, A., Greiner, E.C., 1995. *Sarcocystis falcatula*
350 from passerine and psittacine birds: synonymy with *Sarcocystis neurona* agent of equine protozoal
351 myeloencephalitis. J. Parasitol. 81, 930-935.
352 Dodd, N.S., Lord, J.S., Jehle, R., Parker, S., Parker, F., Brooks, D.R., Hide, G., 2014. *Toxoplasma gondii*:
353 prevalence in species and genotypes of British bats (*Pipistrellus pipistrellus* and *P. pygmaeus*). Exp.
354 Parasitol. 139, 6-11. doi: 10.1016/j.exppara.2014.02.007.
355 Dubey, J.P., Speer, C.A., Fayer, R., 1988. Sarcocystosis of animals and man. CRC Press, Boca Raton, FL
356 Duval, L., Mejean, C., Maganga, G.D., Makanga, B.K., Mangama Koumba, L.B., Peirce, M.A., Ariey, F.,
357 Bourgarel, M., 2012. The chiropteran haemosporidian *Polychromophilus melanipherus*: a worldwide
358 species complex restricted to the family Miniopteridae. Infect. Genet. Evol. 12, 1558-1566. doi:
359 10.1016/j.meegid.2012.06.006.
360 Felsenstein, J., 2008. PHYLIP (Phylogeny Inference Package) version 3.696. Department of Genome
361 Sciences and Department of Biology University of Washington, Seattle, WA.
362 http://evolution.genetics.washington.edu/phylip.html
363 Gardner, M.J., Bishop, R., Shah, T., de Villiers, E.P., Carlton, J.M., Hall, N., Ren, Q., Paulsen, I.T., Pain, A.,
364 Berriman, M., Wilson, R.J., Sato, S., Ralph, S.A., Mann, DJ., Xiong, Z., Shallom, S.J., Weidman, J.,
365 Jiang, L., Lynn, J., Weaver, B., Shoaibi, A., Domingo, A.R., Wasawo, D., Crabtree, J., Wortman, J.R.,
366 Haas, B., Angiuoli, S.V., Creasy, T.H., Lu, C., Suh, B., Silva, JC., Utterback, T.R., Feldblyum, T.V.,
367 Pertea, M., Allen, J., Nierman, W.C., Taracha, E.L., Salzberg, S.L., White, O.R., Fitzhugh, H.A.,
368 Morzaria, S., Venter, J.C., Fraser, C.M., Nene, V., 2005. Genome sequence of *Theileria parva*., a
369 bovine pathogen that transforms lymphocytes. Science 309, 134–137.
370 Halley, Y.A., Dowd, S.E, Decker, J.E., Seabury, P.M., Bhattarai, E., Johnson, C.D., Rollins, D., Tizard, I.R.,
371 Brightsmith, D.J., Peterson, M.J., Taylor, J.F., Seabury, C.M., 2014. A draft *de novo* genome assembly
372 for the northern bobwhite (*Colinus virginianus*) reveals evidence for a rapid decline in effective

373 population size beginning in the late Pleistocene. PLoS One 9, e90240.
374 doi:10.1371/journal.pone.0090240
375 Kutkiené, L., Prakas, P., Sruoga, A., Butkauskas, D. 2010. The mallard duck (*Anas platyrhynchos*) as
376 intermediate host for *Sarcocystis wobeseri* sp. nov. from the barnacle goose (*Branta leucopsis*)
377 Parasitol. Res. 107, 879-888.
378 Kutkiené, L., Prakas, P., Sruoga, A., Butkauskas, D., 2012. Description of *Sarcocystis anasi* sp. nov. and
379 *Sarcocystis albifronsi* sp. nov. in birds of the order Anseriformes. Parasitol. Res. 110, 1043–1046.
380 Longo, M.S., O'Neill, M.J., O'Neill, R.J., 2011. Abundant human DNA contamination identified in non-
381 primate genome databases. PLoS ONE 6, e16410; doi: 10.1371/journal.pone.0016410.
382 Matsubayashi, M., Carreno, R.A., Tani, H., Yoshiuchi, R., Kanai, T., Kimata, I., Uni, S., Furuya, M., Sasai,
383 K., 2011. Phylogenetic identification of *Cystoisospora* spp. from dogs, cats, and raccoon dogs in
384 Japan. Vet. Parasitol. 176, 270–274.
385 McAllister, C.T., Burt, S., Seville, R.S., Robison, H.W., 2011. A new species of *Eimeria* (Apicomplexa:
386 Eimeriidae) from the eastern pipistrelle, *Perimyotis subflavus* (Chiroptera: Vespertilionidae), in
387 Arkansas. J. Parasitol. 97, 896-898. doi: 10.1645/GE-2761.1.
388 Merchant, S., Wood, D.E., Salzberg, S.L., 2014. Unexpected cross-species contamination in genome
389 sequencing projects. PeerJ 2, e675; doi: 10.7717/peerj.675
390 Morrison, D.A., Bornstein, S., Thebo, P., Wernery, U., Kinne, J., Mattsson, .J. G., 2004. The current status of
391 the small subunit rRNA phylogeny of the coccidia (Sporozoa). Int. J. Parasitol. 34, 501-514.
392 Odening, K. 1998. The present state of species-systematics in *Sarcocystis* Lankester, 1882 (Protista,
393 Sporozoa, Coccidia). Syst. Parasitol. 41, 209–233.
394 Olias, P., Olias, L., Krücken, J., Lierz, M., Gruber, A.D., 2011. High prevalence of *Sarcocystis calchasi*
395 sporocysts in European Accipiter hawks. Vet. Parasitol. 175, 230–236.
396 Orosz, F., 2009. Apicortin a unique protein with a putative cytoskeletal role shared only by apicomplexan
397 parasites and the placozoan *Trichoplax adhaerens*. Infect. Genet. Evol. 9, 1275–1286.
398 Orosz, F., 2011. Apicomplexan apicortins possess a long disordered N-terminal extension. Infect. Genet.
399 Evol. 11, 1037–1044.

Prakas, P., Kutkiené, L., Butkauskas, D., Sruoga, A., Zalakevičius, M., 2013. Molecular and morphological investigations of *Sarcocystis corvusi* sp. nov. from the jackdaw (*Corvus monedula*). Parasitol Res. 112, 1163-1167. doi: 10.1007/s00436-012-3247-5.

Prakas, P., Kutkiené, L., Sruoga, A., Butkauskas, D., 2011. *Sarcocystis sp*. from the herring gull (*Larus argentatus*) identity to *Sarcocystis wobeseri* based on cyst morphology and DNA results. Parasitol. Res. 109, 1603-1608. doi: 10.1007/s00436-011-2421-5.

Prakas, P., Liaugaudaitė, S., Kutkiené, L., Sruoga, A., Švažas, S., 2015. Molecular identification of *Sarcocystis rileyi* sporocysts in red foxes (*Vulpes vulpes*) and raccoon dogs (*Nyctereutes procyonoides*) in Lithuania. Parasitol. Res. 114, 1671-1676. doi: 10.1007/s00436-015-4348-8

Ronquist, F., Huelsenbeck, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixture models. Bioinformatics 19, 1572–1574.

Schaer, J., Perkins, S.L., Decher, J., Leendertz, F.H., Fahr, J., Weber, N., Matuschewski, K., 2013. High diversity of West African bat malaria parasites and a tight link with rodent *Plasmodium* taxa. Proc. Natl. Acad. Sci. U.S.A. 110, 17415-17419. doi: 10.1073/pnas.1311016110.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J.D., Higgins, D.G., 2011. Fast scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. 7, 539. doi: 10.1038/msb.2011.75

Tavare, S. 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. Lect. Math. Life Sci. 17, 57-86.

Wang, W., Cao, L., He, B., Li, J., Hu, T., Zhang, F., Fan, Q., Tu, C., Liu, Q., 2013. Molecular characterization of *Cryptosporidium* in bats from Yunnan province, southwestern China. J. Parasitol. 99, 1148-1150. doi: 10.1645/13-322.1.

Wenzel, R., Erber, M., Boch, J., Schellner, H.P., 1982. Sarcosporidia infections in domestic fowl, pheasant and coot. Berl. Munch. Tierarztl. Wochenschr. 95, 188–193.

Wilson, R.J., Denny, P.W., Preiser, P.R., Rangachari, K., Roberts, K., Roy, A., Whyte, A., Strath, M., Moore, D.J., Moore, P.W., Williamson, D.H., 1996. Complete gene map of the plastid-like DNA of the malaria parasite *Plasmodium falciparum*. J. Mol. Biol. 261, 155–172.

Wünschmann, A., Wellehan, J.F. Jr., Armien, A., Bemrick, W.J., Barnes, D., Averbeck, G.A., Roback, R., Schwabenlander, M., D'Almeida, E., Joki, R., Childress, A.L., Cortinas, R., Gardiner, C.H., Greiner,

E.C., 2010. Renal infection by a new coccidian genus in big brown bats (*Eptesicus fuscus*). J. Parasitol. 96, 178-183.

Zhang, G., Cowled, C., Shi, Z., Huang, Z., Bishop-Lilly, KA., Fang, X., Wynne, J.W., Xiong, Z., Baker, M.L., Zhao W., Tachedjian, M., Zhu, Y., Zhou, P., Jiang, X., Ng, J., Yang, L., Wu, L., Xiao, J., Feng, Y., Chen, Y., Sun, X., Zhang, Y., Marsh, G.A., Crameri, G., Broder, C.C., Frey, K.G., Wang, L.F., Wang, J., 2013. Comparative analysis of bat genomes provides insight into the evolution of flight and immunity. Science, 339, 456-460.

Legends to Figures

Fig. 1. Multiple alignment of hypothesized apicortin sequences of Sarcocystidae and *C. virginianus*. Identical and similar amino acids in all cases are labelled by asterisks and colons, respectively. Grey background indicates whether the amino acids are identical only in *Sarcocystis neurona* and *C. virginianus*. Hammondia: *H. hammondi*, XP_008888750; Neospora: *N. caninum*, NCLIV_029060; Toxoplasma: *T. gondii* XP_002364910; Colinus: *C. virginianus* AWGU01108821; Sarcocystis: *S. neurona* JAQE01000498 (JXWP01000002).

Fig. 2. Bayesian phylogenetic tree of the *Sarcocystis* spp. based on sequences of the ITS-1 region. The tree was rooted with *S. lutrae*. Posterior probability support was maximal in all cases. The GenBank accession numbers of ITS-1 genes are given for each taxon and underlined if it was isolated from bird-hosted species. For *Colinus virginianus* the 4300-3367 nucleotides of AWG1001200.1 SimpleDeNovo_contig_1200 were used.

Fig. 3. Phylogenetic trees of the Sarcocystidae *18S rRNA* genes. Three genes of members of the Eimeriidae were also included, using *G. balatonica* as the outgroup. The GenBank accession numbers are given in Supplementary Data S9. For *M. davidii* the 211-2019 nucleotides of contig111512 (GenBank: ALWT01111512.1) were used. The position of *N. eptesici* gene and *M. davidii* contig are indicated in both cases. (**A**) Bayesian analysis. The branch lengths indicate the inferred amount of evolutionary change, according to the scale bar shown. Bayesian posterior probability values are shown for the main branches of the tree. Black circles represent full support for the node. (**B**) Maximum Likelihood tree. Confidence of the tree topology obtained is shown by maximum likelihood bootstrap values calculated from 1000 replicates.

458 **Table 1. Hits in the *C. virginianus* WGS sequence when using as queries the**

459 **available nucleotide sequences from *Sarcocystis* apicoplasts**

Species	GenBank	gene	*C. virginianus*	Identity
S. campestris	GQ851963	RPOb	AWGU01009595 AWGT01005489	94%
S. canis	KC191642	RPOb	AWGU01009595 AWGT01005489	94%
S. falcatula	GQ851962	RPOb	AWGU01009595 AWGT01005489	94%
S. neurona	GQ851961	RPOb	AWGU01009595 AWGT01005489	95%
S. sp.	KC191641	RPOb	AWGU01009595 AWGT01005489	95%
S. muris	AF255924	small subunit ribosomal RNA	AWGU01003450 AWGT01002297	98%

460 RPOb - RNA polymerase beta subunit-like gene. The AWGU and AWGT sequence IDs stand

461 for simple *de novo* (i.e. no scaffolding) and scaffolded *de novo* contigs, respectively.

462

463 **Table 2. Hits in the *C. virginianus* WGS sequence when using as a query the *T. gondii* apicoplast**

464 **complete genome (GenBank U87145.2)**

Description	Max score	Total score	Query cover	E value	Identity	Accession
SimpleDeNovo_contig_3450	6835	13510	30%	0.0	89%	AWGU01003450.1
ScaffoldedDeNovo_contig_2297	6835	13510	30%	0.0	89%	AWGT01002297.1
SimpleDeNovo_contig_9595	6630	6630	23%	0.0	82%	AWGU01009595.1
ScaffoldedDeNovo_contig_5489	6630	6630	23%	0.0	82%	AWGT01005489.1
SimpleDeNovo_contig_45101	3236	4781	23%	0.0	77%	AWGU01045101.1
ScaffoldedDeNovo_contig_42584	3236	4781	23%	0.0	77%	AWGT01042584.1
SimpleDeNovo_contig_73205	2663	2663	11%	0.0	80%	AWGU01073205.1
ScaffoldedDeNovo_contig_42585	2663	2663	11%	0.0	80%	AWGT01042585.1
SimpleDeNovo_contig_47460	198	397	0%	1e-46	94%	AWGU01047460.1
ScaffoldedDeNovo_contig_11590	198	397	0%	1e-46	94%	AWGT01011590.1
ScaffoldedDeNovo_contig_54661	91.6	183	0%	2e-14	95%	AWGT01054661.1
SimpleDeNovo_contig_66517	71.3	142	0%	3e-08	95%	AWGU01066517.1
ScaffoldedDeNovo_contig_21304	71.3	142	0%	3e-08	95%	AWGT01021304.1
SimpleDeNovo_contig_137357	69.4	138	0%	1e-07	98%	AWGU01137357.1
ScaffoldedDeNovo_contig_33333	69.4	138	0%	1e-07	98%	AWGT01033333.1
SimpleDeNovo_contig_257380	63.9	127	0%	4e-06	95%	AWGU01257380.1

465

Table 3. Hits in the *C. virginianus* WGS sequence when using as queries the *S. neurona* potential apicoplast nucleotide sequences

Description	MaxScore	TotalScore	Query cover	E value	Identity	Accession
Query: JAQE01002351 (24001 bp)						
SimpleDeNovo_contig_45101	15005	15005	42%	0.0	93%	AWGU01045101.1
SimpleDeNovo_contig_9595	12759	12759	34%	0.0	95%	AWGU01009595.1
SimpleDeNovo_contig_73205	5775	5775	16%	0.0	93%	AWGU01073205.1
SimpleDeNovo_contig_206802	1341	1341	3%	0.0	93%	AWGU01206802.1
SimpleDeNovo_contig_78021	1194	1194	3%	0.0	95%	AWGU01078021.1
SimpleDeNovo_contig_70976	154	154	0%	2e-33	89%	AWGU01070976.1
SimpleDeNovo_contig_68381	121	121	0%	2e-23	99%	AWGU01068381.1
SimpleDeNovo_contig_282453	110	110	0%	4e-20	97%	AWGU01282453.1
SimpleDeNovo_contig_270464	86.1	86.1	0%	6e-13	95%	AWGU01270464.1
SimpleDeNovo_contig_106377	84.2	84.2	0%	2e-12	92%	AWGU01106377.1
SimpleDeNovo_contig_56616	75.0	75.0	0%	1e-09	91%	AWGU01056616.1
SimpleDeNovo_contig_168981	67.6	67.6	0%	2e-07	95%	AWGU01168981.1
SimpleDeNovo_contig_6309	56.5	56.5	0%	5e-04	86%	AWGU01006309.1
Query: JAQE01002350 (5698 bp)						
SimpleDeNovo_contig_3450	9788	10121	99%	0.0	99%	AWGU01003450.1
SimpleDeNovo_contig_151619	244	443	5%	2e-61	95%	AWGU01151619.1
SimpleDeNovo_contig_135140	137	270	4%	4e-29	88%	AWGU01135140.1
SimpleDeNovo_contig_158980	110	110	1%	9e-21	94%	AWGU01158980.1
SimpleDeNovo_contig_257380	69.4	69.4	0%	2e-08	98%	AWGU01257380.1
SimpleDeNovo_contig_133852	69.4	69.4	0%	2e-08	95%	AWGU01133852.1
SimpleDeNovo_contig_57220	69.4	124	1%	2e-08	98%	AWGU01057220.1
SimpleDeNovo_contig_39406	60.2	60.2	0%	9e-06	100%	AWGU01039406.1
SimpleDeNovo_contig_226848	56.5	56.5	0%	1e-04	97%	AWGU01226848.1
SimpleDeNovo_contig_110794	54.7	54.7	0%	4e-04	94%	AWGU01110794.1

ScaffoldedDeNovo_contig hits are not shown here. See Supplementary Data S4 and S5 for them.

Table 4. Hits in the *M. davidii* WGS sequence when using as a query the *T. gondii* apicoplast complete genome (GenBank U87145.2)

Contig	Max score	Total score	BLAST E value	Identity	GenBank Accession	Length, bp
314352	3493	11197	0.0	88%	ALWT01314352.1	4986
309146	1395	1395	0.0	87%	ALWT01309146.1	1290
307364	1375	1375	0.0	89%	ALWT01307364.1	1162
304579	924	924	0.0	84%	ALWT01304579.1	1027
295304	922	922	0.0	89%	ALWT01295304.1	749
297403	880	880	0.0	87%	ALWT01297403.1	796
293953	822	822	0.0	87%	ALWT01293953.1	721
297566	817	817	0.0	85%	ALWT01297566.1	801
293275	769	769	0.0	86%	ALWT01293275.1	709
274256	490	490	3e-132	86%	ALWT01274256.1	450
257569	370	370	4e-96	89%	ALWT01257569.1	305
266684	348	348	2e-89	88%	ALWT01266684.1	376
256340	339	339	1e-86	87%	ALWT01256340.1	297
293156	337	1348	4e-86	84%	ALWT01293156.1	706
256707	287	287	4e-71	84%	ALWT01256707.1	299
262086	276	276	8e-68	84%	ALWT01262086.1	338
261058	267	267	5e-65	82%	ALWT01261058.1	331

473 **Table 5. *Sarcocystis* 18S ribosomal RNA genes showing highest identity to *C. colinus* contamination**

Species	GenBank	Identity, %	bp	Source of isolation (host)	Order (birds)
S. albifronsi	EU502868	99.8	1792	*Anser albifrons*	Anseriformes
S. anasi	EU553477	99.8	1792	*Anas platyrhynchos*	Anseriformes
S. arctica	KF601301	99.0	1803	*Vulpes lagopus*	
S. arctosi	EF564590	99.3	1484	*Ursus arctos*	
S. atraii	KJ810606	99.3	1493	*Fulica atra*	Gruiformes
S. calchasi	GQ245670	99.3	1804	*Columba livia*	Columbiformes
S. canis	DQ146148	99.5	994	*Canis canis*	
S. columbae	HM125054	99.3	1765	*Columba palumbus*	Columbiformes
S. cornixi	EU553478	98.7	1795	*Corvus cornix*	Passeriformes
S. corvusi	JN256117	99.3	1792	*Corvus monedula*	Passeriformes
S. dispersa	AF120115	98.6	1610	*Tyto alba*	Strigiformes
S. lutrae	KM657770	99.3	1804	*Lutra lutra*	
S. mucosa	AF109679	99.2	1824		
S. muris	SARRR16S	98.0	1809		
S. neurona	U07812	98.9	1803	*Bos taurus*	
S. neurona	HQ709144	99.1	763	*Martes pennant*	
S. neurona	JXWP01000699	98.4	1384	*Enhydra lutris nereis*	
S. rileyi	KJ396583	99.7	1803	*Somateria mollissima*	Anseriformes
S. rileyi Europa	HM185742	99.7	1792	*Anas platyrhynchos*	Anseriformes
S. sp	KM362427	98.9	1669	*Canis familiaris*	
S. sp	KF309699	99.3	1749	*Eothenomys miletus pocok*	
S. sp	GU253884	99.4	1630	*Accipiter nisus*	Falconiformes

S. sp.	JQ733511	99.4	1801	*Phalacrocorax carbo*	Pelecaniformes
S. sp	JQ733508	99.4	1802	*Larus marinus*	Charadriiformes
S. sp.	AF513487	98.9	1594	*Sorex araneus*	
S. sp.	KF278953	99.2	1448	*Didelphis virginiana*	
S. tupaia	FJ827486	99.1	1311	*Tupaia belangeri chinensis*	
S. turdusi	JF975681	99.3	1793	*Turdus merula*	Passeriformes
S. wobeseri	HM159419 GQ922885 EU502869	99.4	1792	*Larus argentatus* *Anas platyrhynchos* *Branta leucopsis* *Anser albifrons*	Charadriiformes Anseriformes

474

475 **Table 6.** : **Sarcocystidae 18S ribosomal RNA genes showing highest identity to *M. davidii***
476 **contamination**

Species	Max score	Query cover	E value	Identity	GenBank
Nephroisospora eptesici	3212	98%	0.0	99%	EU334134
Besnoitia besnoiti	3134	100%	0.0	98%	EU789637
Toxoplasma gondii	3116	100%	0.0	98%	M97703
Neospora caninum	3112	99%	0.0	98%	U16159
Cystoisospora timoni	3081	99%	0.0	98%	AY279205
Besnoitia jellisoni	3077	99%	0.0	98%	AF291426
Isospora belli	3075	99%	0.0	98%	DQ060661
Hammondia hammondi	3035	97%	0.0	98%	AF096498
Hammondia triffittae	3031	97%	0.0	98%	GQ984222
Hammondia heydorni	3031	97%	0.0	98%	JX220986
Cystoisospora belli	3005	97%	0.0	98%	AB268326
Isospora suis	2990	99%	0.0	97%	U97523
Cystoisospora sp.	2968	96%	0.0	98%	AB519674
Sarcocystis sp. ex Phalacrocorax carbo	2944	99%	0.0	96%	JQ733511
Sarcocystis sp. ex Columba livia	2944	100%	0.0	96%	GQ245670
Sarcocystis arctica	2942	99%	0.0	96%	KF601301
Sarcocystis mucosa	2940	99%	0.0	96%	AF109679
Sarcocystis lutrae	2933	99%	0.0	96%	KM657769
Sarcocystis sp. ex Corvus monedula	2933	99%	0.0	96%	JN256117
Isospora felis	2933	97%	0.0	97%	L76471
Sarcocystis rileyi	2931	99%	0.0	96%	KJ396583
Sarcocystis sp. ex Larus marinus	2929	99%	0.0	96%	JQ733508
Cystoisospora ohioensis	2929	96%	0.0	97%	GU292304
Sarcocystis anasi	2922	99%	0.0	96%	EU553477
Sarcocystis albifronsi	2916	99%	0.0	96%	EU502868
Sarcocystis sp. ex Columba palumbus	2872	97%	0.0	96%	HM125054
Besnoitia darlingi	2826	91%	0.0	98%	GU479631
Besnoitia oryctofelisi	2822	91%	0.0	98%	GU479632
Besnoitia bennetti	2804	90%	0.0	98%	AY665399
Frenkelia glareoli	2678	90%	0.0	96%	AF009245
Frenkelia microti	2639	90%	0.0	96%	AF009244
Hyaloklossia lieberkuehni	2580	87%	0.0	96%	AF298623
Isospora ohioensis	2357	77%	0.0	97%	AF029303
Eimeria variabilis	2102	98%	0.0	88%	GU479674
Lankesterella minima	1925	93%	0.0	87%	AF080611
Goussia balatonica	2396	94%	0.0	92%	GU479650

477 Query: *Myotis davidii*, contig111512, 211-2019 nucleotides (GenBank: ALWT01111512)

478 Subject: Sarcocystidae nucleotide sequences

479 The last three species belong to the Eimeridiiae, the nearest relative of Sarcocystidae.

480

Supplementary data

Supplementary Data S1. Multiple sequence alignments used for constructing phylogenetic trees of Fig. 2 and Fig. 3.

Supplementary Data S2. BLASTN search used *S. neurona* RNA polymerase beta subunit (RPOb) (GenBank: GQ851961.1) and *Sarcocystis muris* small subunit ribosomal RNA gene (GenBank: AF255924.1) genes as queries against *C. virginianus* WGS nucleotide sequences.

Supplementary Data S3. Hits in the *C. virginianus* WGS sequence when using as a query the *T. gondii* apicoplast complete genome (U87145.2)

Supplementary Data S4. Hits in the *C. virginianus* WGS sequence when using as a query a *S. neurona* potential apicoplast nucleotide sequence (GenBank JAQE01002351).

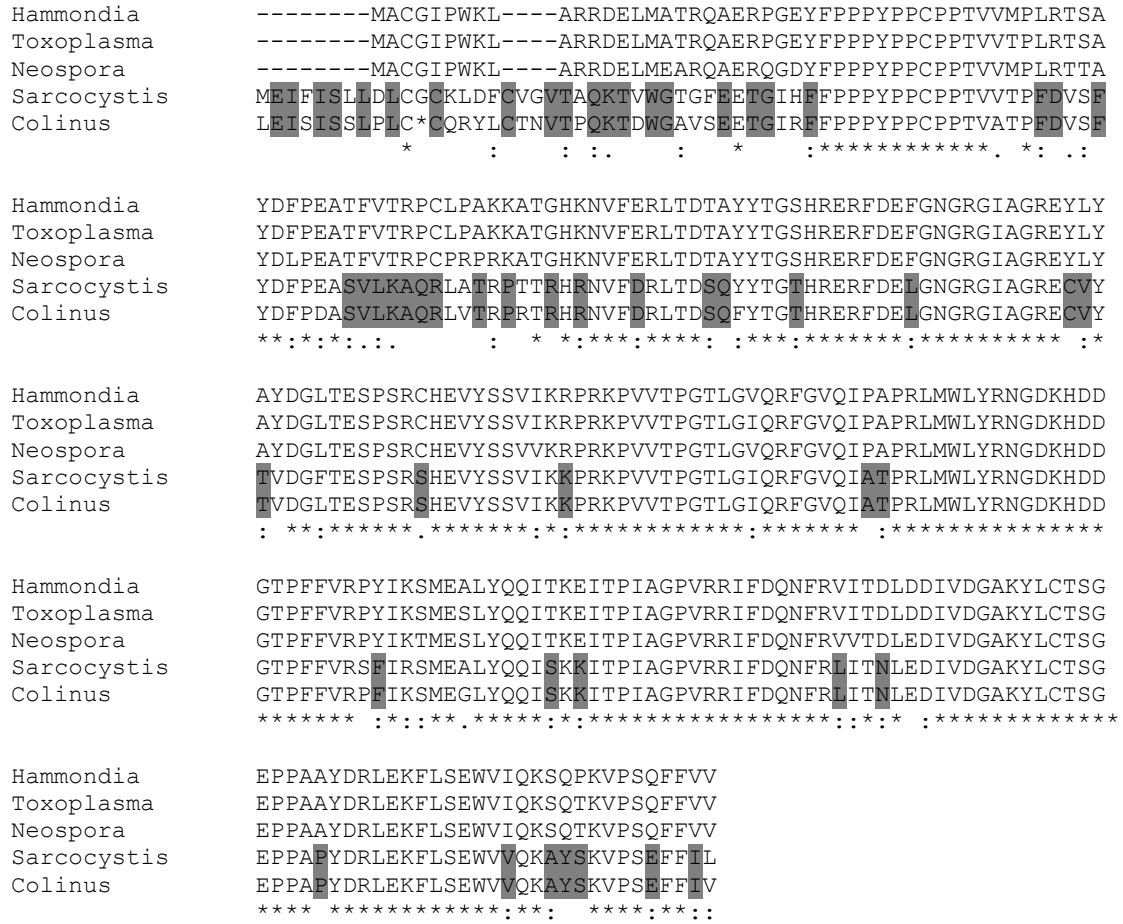
Supplementary Data S5. Hits in the *C. virginianus* WGS sequence when using as a query a *S. neurona* potential apicoplast nucleotide sequence (GenBank JAQE01002350).

Supplementary Data S6. Hits in the *Ovis aries musimon* WSG sequence when using as a query the *T. gondii* apicoplast complete genome

Supplementary Data S7. Sequences producing significant alignments by BLASTN search. Query: *T. gondii* apicoplast, complete genome (GenBank: U87145.2). Subject: *M. davidii*, WGS sequence.

Supplementary Table S1. Further *Sarcocystis* 18S ribosomal RNA genes

Figure


Hammondia --------MACGIPWKL----ARRDELMATRQAERPGEYFPPPYPPCPPTVVMPLRTSA
Toxoplasma --------MACGIPWKL----ARRDELMATRQAERPGEYFPPPYPPCPPTVVTPLRTSA
Neospora --------MACGIPWKL----ARRDELMEARQAERQGDYFPPPYPPCPPTVVMPLRTTA
Sarcocystis MEIFISLLDLCGCKLDFCVGVTAQKTVWGTGFEETGIHFFPPPYPPCPPTVVTPFDVSF
Colinus LEISISSLPLC*CQRYLCTNVTPQKTDWGAVSEETGIRFFPPPYPPCPPTVATPFDVSF
Hammondia YDFPEATFVTRPCLPAKKATGHKNVFERLTDTAYYTGSHRERFDEFGNGRGIAGREYLY
Toxoplasma YDFPEATFVTRPCLPAKKATGHKNVFERLTDTAYYTGSHRERFDEFGNGRGIAGREYLY
Neospora YDLPEATFVTRPCPRPRKATGHKNVFERLTDTAYYTGSHRERFDEFGNGRGIAGREYLY
Sarcocystis YDFPEASVLKAQRLATRPTTRHRNVFDRLTDSQYYTGTHRERFDELGNGRGIAGRECVY
Colinus YDFPDASVLKAQRLVTRPRTRHRNVFDRLTDSQFYTGTHRERFDELGNGRGIAGRECVY
Hammondia AYDGLTESPSRCHEVYSSVIKRPRKPVVTPGTLGVQRFGVQIPAPRLMWLYRNGDKHDD
Toxoplasma AYDGLTESPSRCHEVYSSVIKRPRKPVVTPGTLGIQRFGVQIPAPRLMWLYRNGDKHDD
Neospora AYDGLTESPSRCHEVYSSVVKRPRKPVVTPGTLGVQRFGVQIPAPRLMWLYRNGDKHDD
Sarcocystis TVDGFTESPSRSHEVYSSVIKKPRKPVVTPGTLGIQRFGVQIATPRLMWLYRNGDKHDD
Colinus TVDGLTESPSRSHEVYSSVIKKPRKPVVTPGTLGIQRFGVQIATPRLMWLYRNGDKHDD
Hammondia GTPFFVRPYIKSMEALYQQITKEITPIAGPVRRIFDQNFRVITDLDDIVDGAKYLCTSG
Toxoplasma GTPFFVRPYIKSMESLYQQITKEITPIAGPVRRIFDQNFRVITDLDDIVDGAKYLCTSG
Neospora GTPFFVRPYIKTMESLYQQITKEITPIAGPVRRIFDQNFRVVTDLEDIVDGAKYLCTSG
Sarcocystis GTPFFVRSFIRSMEALYQQISKKITPIAGPVRRIFDQNFRLITNLEDIVDGAKYLCTSG
Colinus GTPFFVRPFIKSMEGLYQQISKKITPIAGPVRRIFDQNFRLITNLEDIVDGAKYLCTSG
Hammondia EPPAAYDRLEKFLSEWVIQKSQPKVPSQFFVV
Toxoplasma EPPAAYDRLEKFLSEWVIQKSQTKVPSQFFVV
Neospora EPPAAYDRLEKFLSEWVIQKSQTKVPSQFFVV
Sarcocystis EPPAPYDRLEKFLSEWVVQKAYSKVPSEFFIL
Colinus EPPAPYDRLEKFLSEWVVQKAYSKVPSEFFIV

Figure

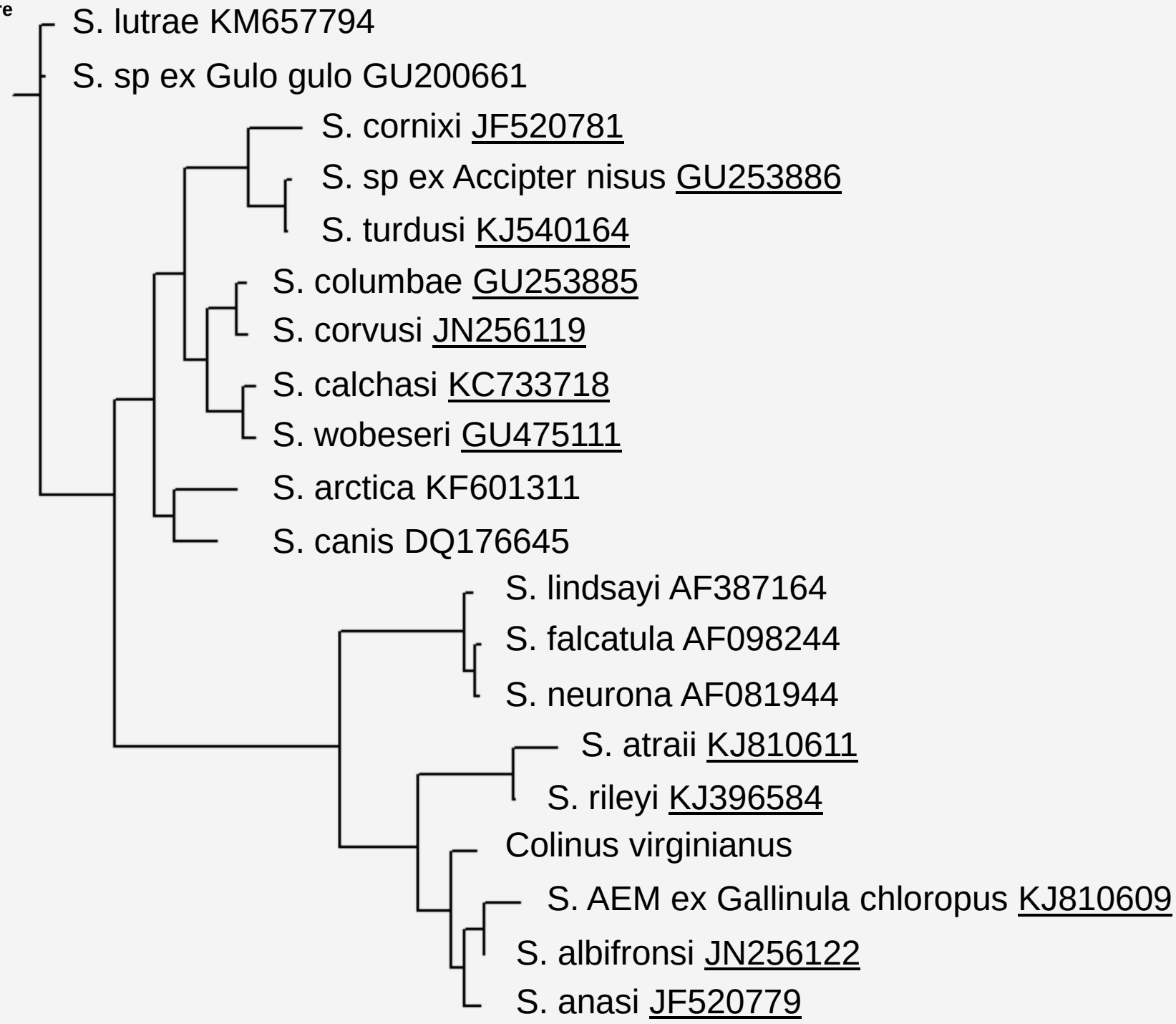
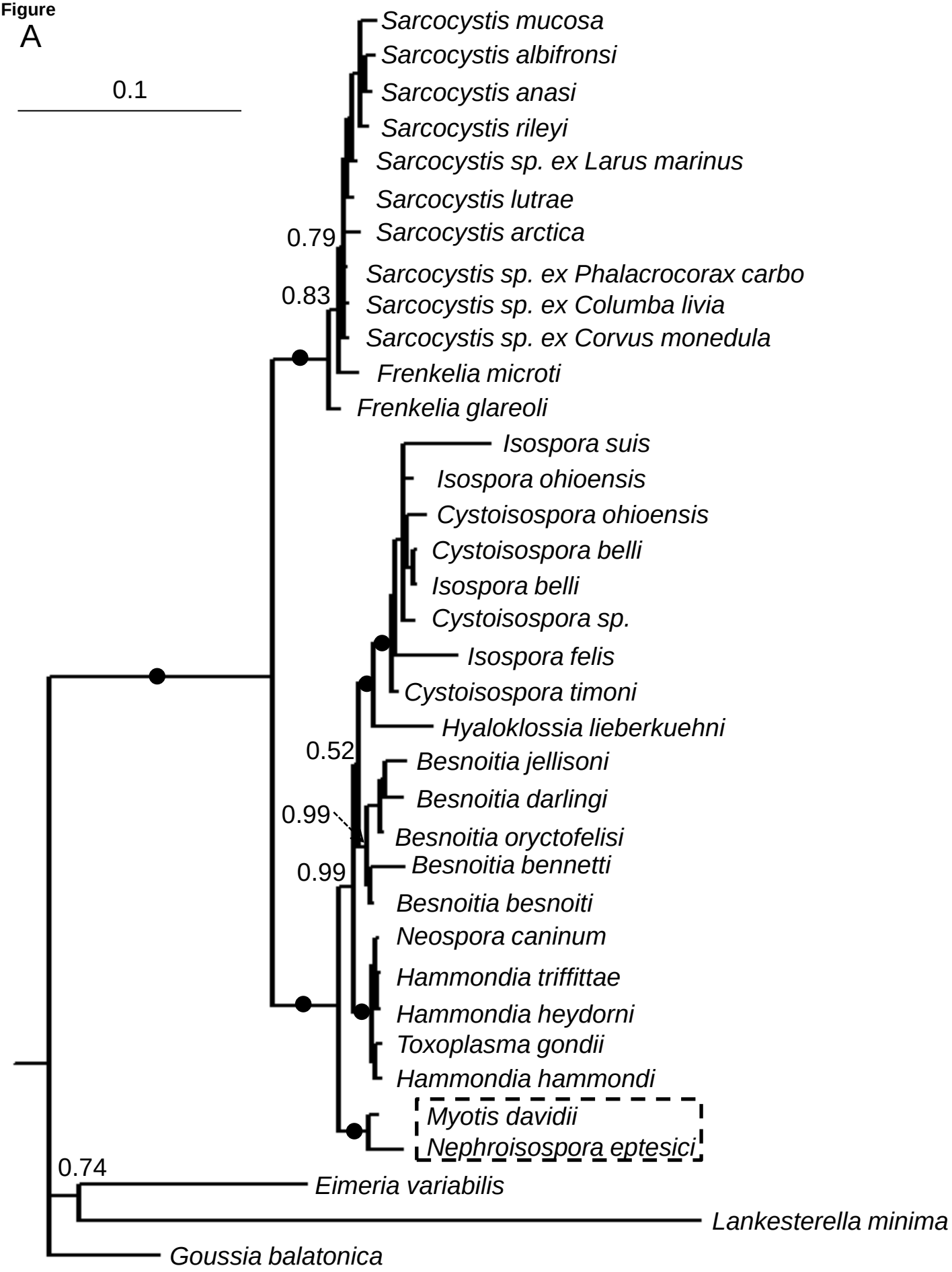
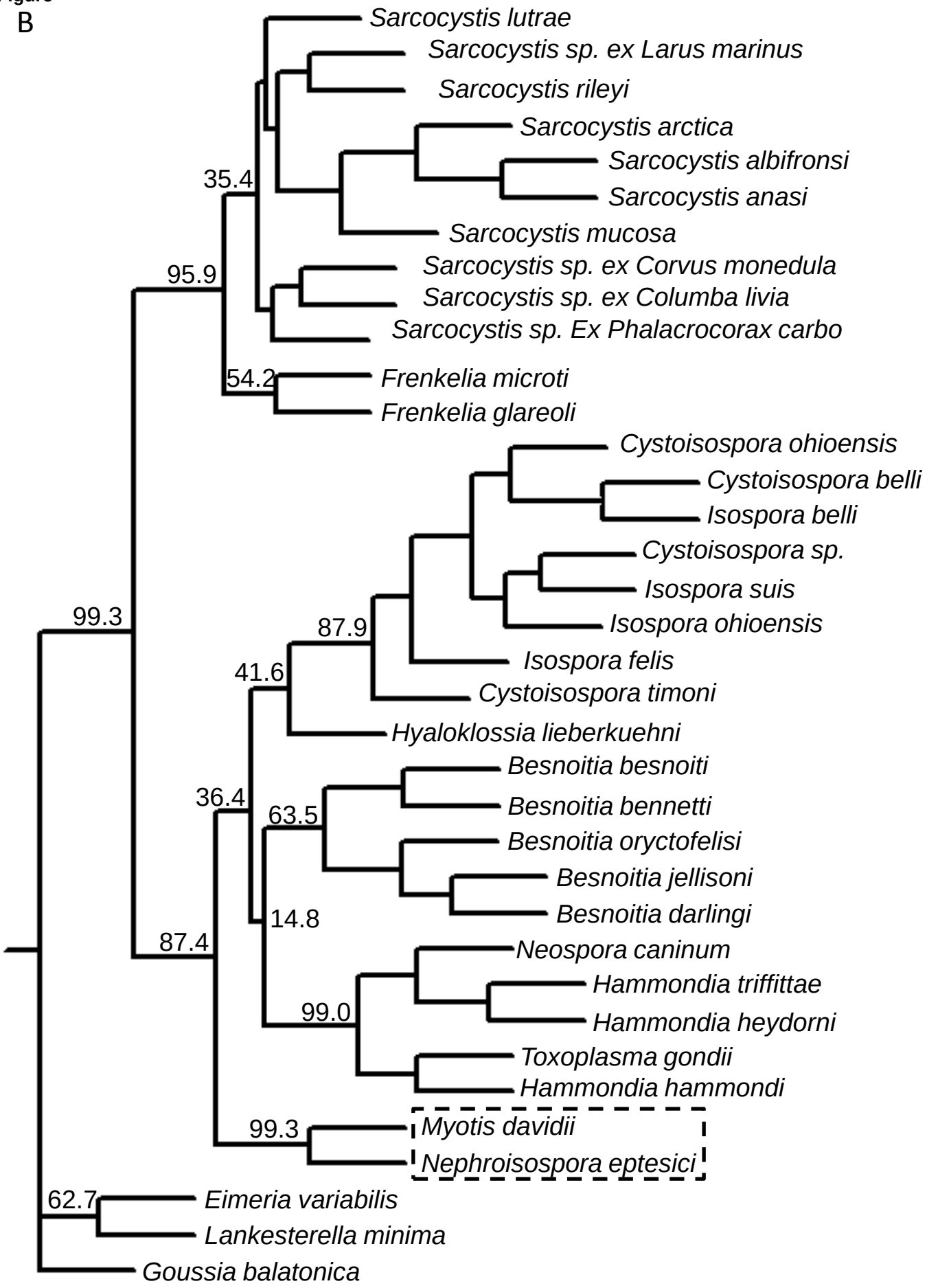

S. lutrae KM657794
S. sp ex Gulo gulo GU200661
S. cornixi JF520781
S. sp ex Accipter nisus GU253886
S. turdusi KJ540164
S. columbae GU253885
S. corvusi JN256119
S. calchasi KC733718
S. wobeseri GU475111
S. arctica KF601311
S. canis DQ176645
S. lindsayi AF387164
S. falcatula AF098244
S. neurona AF081944
S. atraii KJ810611
S. rileyi KJ396584
Colinus virginianus
S. AEM ex Gallinula chloropus KJ810609
S. albifronsi JN256122
S. anasi JF520779

Figure
A


0.1
Sarcocystis mucosa
Sarcocystis albifronsi
Sarcocystis anasi
Sarcocystis rileyi
Sarcocystis sp. ex Larus marinus
Sarcocystis lutrae
0.79
Sarcocystis arctica
Sarcocystis sp. ex Phalacrocorax carbo
0.83
Sarcocystis sp. ex Columba livia
Sarcocystis sp. ex Corvus monedula
Frenkelia microti
Frenkelia glareoli
Isospora suis
Isospora ohioensis
Cystoisospora ohioensis
Cystoisospora belli
Isospora belli
Cystoisospora sp.
Isospora felis
Cystoisospora timoni
Hyaloklossia lieberkuehni
0.52
Besnoitia jellisoni
Besnoitia darlingi
0.99
Besnoitia oryctofelisi
0.99
Besnoitia bennetti
Besnoitia besnoiti
Neospora caninum
Hammondia triffittae
Hammondia heydorni
Toxoplasma gondii
Hammondia hammondi
Myotis davidii
Nephroisospora eptesici
0.74
Eimeria variabilis
Lankesterella minima
Goussia balatonica

Figure

B


Sarcocystis lutrae
Sarcocystis sp. ex Larus marinus
Sarcocystis rileyi
Sarcocystis arctica
Sarcocystis albifronsi
Sarcocystis anasi
Sarcocystis mucosa
Sarcocystis sp. ex Corvus monedula
Sarcocystis sp. ex Columba livia
Sarcocystis sp. Ex Phalacrocorax carbo
Frenkelia microti
Frenkelia glareoli
Cystoisospora ohioensis
Cystoisospora belli
Isospora belli
Cystoisospora sp.
Isospora suis
Isospora ohioensis
Isospora felis
Cystoisospora timoni
Hyaloklossia lieberkuehni
Besnoitia besnoiti
Besnoitia bennetti
Besnoitia oryctofelisi
Besnoitia jellisoni
Besnoitia darlingi
Neospora caninum
Hammondia triffittae
Hammondia heydorni
Toxoplasma gondii
Hammondia hammondi
Myotis davidii
Nephroisospora eptesici
Eimeria variabilis
Lankesterella minima
Goussia balatonica
35.4
95.9
54.2
99.3
87.9
41.6
36.4
63.5
14.8
87.4
99.0
99.3
62.7

Multi-media supplement

Click here to download Multi-media supplement: SupplementaryDataS1.docx

Multi-media supplement

Click here to download Multi-media supplement: SupplementaryDataS2.docx

Multi-media supplement

Click here to download Multi-media supplement: SupplementaryDataS3.docx

Multi-media supplement

Click here to download Multi-media supplement: SupplementaryDataS4.docx

Multi-media supplement

Click here to download Multi-media supplement: SupplementaryDataS5.docx

Multi-media supplement

Click here to download Multi-media supplement: SupplementaryDataS6.docx

Multi-media supplement

Click here to download Multi-media supplement: SupplementarydataS7.docx

Multi-media supplement

Click here to download Multi-media supplement: SupplementaryTableS1.docx