Contents lists available at ScienceDirect

International Journal for Parasitology

journal homepage: www.elsevier.com/locate/ijpara

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, Budapest, Hungary

ARTICLE INFO

Article history:
Received 9 April 2015
Received in revised form 9 July 2015
Accepted 11 July 2015
Available online xxxx

Keywords:
Apicomplexa
Apicoplast
Sarcocystis
Apicortin
Host
Phylogenetic tree
Bat
Nephroisospora

ABSTRACT

This paper highlights a general problem, namely that host genome sequences can easily be contaminated with parasite sequences, thus careful isolation of genetic material and careful bioinformatics analysis are needed in all cases. Two recently published genomes are shown here to be 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 searches against nucleotide sequences of various animal groups looking for possible contamination. Draft genomes of a bird, *Colinus virginianus* (Halley et al., 2014), and a bat, *Myotis davidii* (Zhang et al., 2013) were found to contain at least six 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 rRNA 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, *Nephroisospora eptesici*, has been recently identified from the kidney of big brown bats (*Eptesicus fuscus*).

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 in removing human DNA and other common laboratory contaminants such as *Escherichia 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 those of parasites, 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 cycles: 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).

[*] Address: Institute of Enzymology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Magyar tudósok körútja 2., H-1117 Budapest, Hungary. Tel.: +36 1 3826714.
E-mail address: orosz.ferenc@ttk.mta.hu

http://dx.doi.org/10.1016/j.ijpara.2015.07.002
0020-7519/© 2015 Australian Society for Parasitology Inc. Published by Elsevier Ltd. All rights reserved.

A new member of the Sarcocystidae family has recently been 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* spp. 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 recently become 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, due to contamination. Thus it was decided to systematically investigate this problem and sequences of the characteristic apicomplexan organelle, the apicoplast, were used as queries in BLASTN searches against nucleotide sequences of various animal groups, looking for possible contamination. 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, also hosted by a bat, *Myotis davidii*.

2. Materials and methods

2.1. Database similarity search and phylogenetic analysis

Accession numbers of protein and nucleotide sequences refer to the National Center for Biotechnology Information (NCBI) (USA) 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 (ESTs), Transcriptome Shotgun Assembly (TSAs) and WGSs available at the NCBI website. In further analyses, the 18S rRNA and the internal transcribed spacer 1 (ITS1) 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 using 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 (Tavare, 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 2.4×10^6 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 (version 3.695).

The Phylip (Phylogeny Inference Package, version 3.696; http://evolution.genetics.washington.edu/phylip.html) 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 (version 3.695), from the original data i.e. the multiple alignments done by Clustal Omega. This was followed by running the program Dnaml (version 3.695) (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 1000 trees) was generated using the program Consense (version 3.695). The trees were drawn using the program Drawgram.

3. Results and discussion

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

Colinus virginianus putative apicortin is very similar to the apicortins of the Sarcocystidae, *Toxoplasma gondii, Neospora caninum* and *Hammondia hammondi* (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). *Sarcocystis 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 neurological 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 to assume that, similarly to other birds, *C. virginianus* is also an intermediate host of the parasite. (*Colinus 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 the *C. virginianus* genome is contamination which originated from *S. neurona*. Rather, it is probably contamination from another species of the *Sarcocystis* genus, the genome of which has not yet been sequenced.

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 those for RNA polymerase beta subunit-like (RPOb) genes from various *Sarcocystis* spp., and one of those for a ssrRNA gene from *Sarcocystis muris*. Using these sequences as BLASTN queries against avian nucleotide sequences of the NCBI GenBank database, there are no hits other than those of the *C. virginianus* WGS sequence. Any hit would be highly 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 those 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 proteins. The closest relative of *Sarcocystis* spp. is *T. gondii*, belonging to the same