An unexpected tenant: contamination in a *Physeter catodon* (Physeteridae, Artiodactyla) genome indicates undescribed species of *Sarcocystis* Lankester, 1882 (*Sarcocystidae, Eucoccidiorida*) in the marine environment

Joaquim Filipe Faria and David James Harris

1 Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Porto, Portugal;
2 CIBIO Research Centre in Biodiversity and Genetic Resources, InBIO, Universidade do Porto, Campus de Vairão, Vairão, Vila do Conde, Portugal

**Abstract:** Data mining animal of genomes has been used before to identify endoparasites, and may be a particularly useful tool to surpass some difficulties faced by studies in the marine environment. We detected a species of *Sarcocystis* Lankester, 1882, contamination in the sperm whale (*Physeter catodon* Linnaeus) reference genome available in the GenBank database. We identified and extracted multiple gene fragments and placed the sequences in a phylogenetic framework. Our results indicate that the sequences of *Sarcocystis* sp. found in the genome do not correspond to any currently described species, despite a few other similar sequences having been identified in fur seals (*Pinnipedia*) and another sperm whale. Including data from previous studies, we suggest there is enough evidence to support the occurrence of at least four species of *Sarcocystis* in marine mammals. We also demonstrate that the term “*S. canis-like*” has been used for samples not closely related to *Sarcocystis canis* Dubey et Speer, 1991.

**Keywords:** Endoparasite, sperm whale, marine mammal, *Sarcocystis neurona*, *Sarcocystis canis*

Detecting infectious agents is the first step towards disease management in both the terrestrial and marine environment, but the complexity of the marine environment makes this task particularly intricate. Increasing disease reports in marine mammals may indicate deteriorating ecosystems, through both climate change and pollution, but may also reflect improvements in diagnostic methodologies, especially the use of molecular screening. DNA amplification of endoparasites using PCR directly from host tissues enables both accurate parasite identification and placement within a phylogenetic framework.

*Sarcocystis* Lankester, 1882 is a genus of parasitic protists (Apicomplexa), the majority of which infect mammals, and which typically have a two-host, predator-prey life cycle including an intermediate and a definitive host. Most have a narrow intermediate host range, although notable exceptions include *Sarcocystis neurona* Dubey, Davis, Speer, Bowman, Lahunta, Granstrom, Topper, Hamir, Cummings et Suter, 1991, and *Sarcocystis canis* Dubey et Speer, 1991 (Dubey 2016). *Sarcocystis neurona* was first identified as causing neurological diseases in horses, but has since been reported from numerous marine mammals, including seals, sea-lions and sea-otters (Dubey et al. 2001, Mylincenko et al. 2008, Girard et al. 2016). Since the definitive host of *S. neurona* is the Virginia opossum (*Didelphis virginiana* Kerr), a land-sea transmission pathway has been proposed, with *S. neurona* cysts remaining viable in the environment for months (Burgess et al. 2020).

In contrast, “*S. canis-like*” parasites were detected in an Indopacific bottlenosed dolphin (*Tursiops aduncus* Ehrenberg) (Calero-Bernal et al. 2017), although most diagnoses in marine mammals have been based solely on the morphology of cysts found in muscle tissues, such as the *S. canis-like* infection of a Hawaiian monk seal (*Neomona-chus schauinslandi* Matschie) (Yantis et al. 2003).

Reisfeld et al. (2019) identified *Sarcocystis* in subantarctic fur seals (*Arctocephalus tropicalis* Gray) using ITS1 sequences that were similar to sequences from a parasite of a sperm whale (*Physeter catodon* Linnaeus), but there were limited data available for this marker to assess relationships further. Various other studies of marine mammals have reported infections with *Sarcocystis coccidia*, but without establishing the species (e.g., Cuvertoret-Sanz et al. 2020).

It is clear that additional molecular data are needed to better ascertain the parasites found in marine mammals, and to determine how related they are to those from terres-
trial hosts. Targeted screening of host tissues revealed endoparasites that have occasionally been identified through datamining approaches of animal genomes from public databases. For example, Borner and Burmester (2017) identified extensive genes from a piroplasmid parasite from the assembly of the genome of platypus Ornithorhynchus anatinus Shaw, highlighting how this can be a valuable source of information regarding poorly known parasites.

During searches of public databases as part of ongoing molecular screening approaches of Sarcocystis parasites using the 18S rRNA gene, we unexpectedly identified similarities with the reference genome of the sperm whale (P. catodon), indicating that this genome included parasite contaminant sequences. Given the limited current knowledge regarding these parasites from marine mammals, we here identified and extracted sequences from several genes, and estimated a phylogeny to determine how related these were to either S. neurona or S. canis.

Potential genes of species of Sarcocystis from the contaminated P. catodon reference genome were identified using BLAST similarity searches (Altschul et al. 1990). We searched for genes that were widely used in phylogenetic assessments of Sarcocystis spp. and other apicomplexan parasites, so that comparative data were available for three nuclear regions (28S rRNA, 18S rRNA and ITS1), two mitochondrial (cox1 and cytB) and one apicoplast gene (rpoB). Sequences of Sarcocystis spp. from these genes were used as query for the blastn searches in the RefSeq Genome Database, with the search narrowed to the P. catodon (taxid: 9755) genome. Sequences from P. catodon were then aligned against comparable sequences from known species of Sarcocystis, downloaded from GenBank, using MUSCLE (Edgar 2004) and with Gblocks (Castresana 2000) to identify and remove poorly aligned regions.

Haplotype networks for individual genes were built using TCS v1.21 (Clement et al. 2000), with gaps treated as missing data. The connection limit was lowered to 90% for the 28S and cox1 genes. Results were visualised in tcsBU (Múrias dos Santos et al. 2016).

Phylogenies for the 28S rRNA, 18S rRNA, ITS1, cox1 and rpoB genes were estimated using MrBayes v3.2.7 (Ronquist and Huelsenbeck 2003), and a species tree for the same genes was estimated in BEAST2 v2.6.3 (Bouckaert et al. 2019) using the StarBeast2 template for the species tree. The most appropriate model of molecular evolution for each gene was determined with the AIC, using PartitionFinder2 (Lanfear et al. 2016) and MEGAX v10 (Kumar et al. 2018). Estimates employed random starting trees, ran for 10 × 10^7 generations, and sampled every 1000. A 10–25% burn-in was applied to the obtained trees, and remaining data were used to estimate posterior nodal probabilities and produce a summary phylogeny. Results were visualised with FigTree v1.4.4 (Rambaut 2018).

BLAST searches using sequences from the 18S rRNA, 28S rRNA, ITS1, cox1, cytB and rpoB genes from species of Sarcocystis known to infect marine mammals against the P. catodon reference genome in NCBI GenBank database yielded matches with high query cover and percent identity (Table 1), both in unplaced genomic scaffold se-

---

**Fig. 1.** Haplotype networks and BI phylogenies for the 18S rRNA, 28S rRNA, ITS1, cox1 and rpoB genes. Haplotypes marked with a star do not connect at the defined connection limit. Additional labels reflect GenBank numbers (see text for details). Probability support values are represented in the phylogeny for nodes scoring under 100%.
sequences and in the annotated genome. The ITS1 sequences were only identified when using BLAST with specific species of Sarcocystis, due to the high diversity of this region (Rosenthal et al. 2008).

Bayesian inference analyses for the contaminated P. catodon genome fragments corresponding to the 18S rRNA, 28S rRNA, ITS1 cox1 and rpoB gene regions of Sarcocystis sp. (aligned lengths 1,757, 1,504, 388, 1,053 and 886 base pairs, respectively), both individually and concatenated, allowed phylogenetic estimates of relationships to be determined. Sequences of Sarcocystis spp. for the cytB gene are scarce; therefore, we did not analyse the data obtained from this gene in a phylogenetic framework.

Haplotype networks show that most species can be distinguished, although haplotype sharing between closely related species is seen in 18S rRNA, cox1 and possibly rpoB (Fig. 1). Sarcocystis neurona and Sarcocystis falcata Stiles, 1893 consistently group together. The Sarcocystis sequences identified from the datamining of the P. catodon genome have unique haplotypes for all genes. “Sarcocystis canis-like” 18S rRNA sequences from seal species (Pinnipedia) and a polar bear (Ursus maritimus Phipps) cluster with S. canis in the haplotype network, but the two cox1 sequences identified as S. canis by Calero-Bernal et al. (2017) from the Indopacific bottlenosed dolphin (KX721496, KX721497) do not.

Unidentified Sarcocystis spp. rpoB sequences from a sea otter, Enhydra lutris (Linnaeus) (AY164998), and Magellanic penguin, Spheniscus magellanicus (Forster) (MG493472), appear near or share haplotypes with S. neurona and S. falcata, suggesting they could also be of those species haplogroup.

For the network of ITS1 sequences, most species formed unconnected haplogroups, with the exceptions of S. canis with Sarcocystis spp. from bears, S. neurona with S. falcata, and the sequence recovered from the P. catodon genome with three other unidentified Sarcocystis spp. collected from two subantarctic fur seals (A. tropicalis MH918015, MW264422) and another sperm whale (HQ184185). These last three matched with over 99% identity in the BLAST searches using the recovered sequence as query, while the next best match scored only 87% for S. canis from an American black bear (Ursus americanus Pallas).

Estimation of a species tree (Fig. 2) using the combined five genes indicates that Sarcocystis sp. from the whale genome is a sister taxon to the pair of Sarcocystis arctica and S. canis, despite being distinct from both. This group of three species is related to other species of Sarcocystis, including S. lari and S. haliei from white tailed sea eagles (Haliaeetus albicilla (Linnaeus)) and great cormorants (Phalacrocorax carbo (Linnaeus)), and Sarcocystis lutrae from otters (Lutra lutra (Linnaeus)), foxes (Vulpes vulpes (Linnaeus)) and some mustelid species. Sarcocystis neurona and S. falcata seem to be closely related sister-species, and together are sister taxa to the remaining taxa of Sarcocystis analysed.

Our datamining approach recovered multiple genes of Sarcocystis sp. from the whale genome. Given that we were able to identify all the genes most widely used in phylogenetic analyses of Sarcocystis, and for long fragments (up to 2,472 bp), it seems likely that sequences from many genes of these parasites have incidentally been included in this genomic assessment. Our approach again indicates how useful datamining can be to obtain genetic data from poorly-known endoparasites.

To the best of our knowledge, prior to this study, coccidia of Sarcocystis from marine mammals were assigned to S. neurona, S. canis, S. canis-like, or were unidentified.
Table 1. Matches from genomic sequences of Physeter catodon Linnaeus when using nucleotide sequences from species of Sarcocystis known to infect marine mammals as query in BLAST searches.

| Gene  | Sarcocystis species | GenBank accession no. | P. catodon accession no. | Query cover (base pairs) | Percent identity |
|-------|---------------------|-----------------------|--------------------------|--------------------------|-----------------|
| 18S rRNA S. arctica | KF601301 | NW_021151540 | 1803 | 99% |
| S. canis | DQ146148 | NW_021151540 | 994 | 99% |
| S. neurona | KF854255 | NW_021151540 | 1437 | 99% |

28S rRNA

| Gene  | Sarcocystis species | GenBank accession no. | P. catodon accession no. | Query cover (base pairs) | Percent identity |
|-------|---------------------|-----------------------|--------------------------|--------------------------|-----------------|
| S. arctica | KF601312 | NW_021154098 | 1533 | 99% |

S. canis

| Gene  | Sarcocystis species | GenBank accession no. | P. catodon accession no. | Query cover (base pairs) | Percent identity |
|-------|---------------------|-----------------------|--------------------------|--------------------------|-----------------|
| S. canis | n/a | | | | |
| S. neurona | AF092927 | NW_021154098 | 1922 | 97% |

cox1

| Gene  | Sarcocystis species | GenBank accession no. | P. catodon accession no. | Query cover (base pairs) | Percent identity |
|-------|---------------------|-----------------------|--------------------------|--------------------------|-----------------|
| S. arctica | KX022114 | NC_041214 | 1021 | 99% |
| S. canis | KX721495 | NC_041214 | 916 | 99% |
| S. neurona | CO747338 | NC_041214 | 848 | 96% |

cytB

| Gene  | Sarcocystis species | GenBank accession no. | P. catodon accession no. | Query cover (base pairs) | Percent identity |
|-------|---------------------|-----------------------|--------------------------|--------------------------|-----------------|
| S. arctica | n/a | | | | |
| S. canis | n/a | | | | |
| S. neurona | KP871703 | NC_041214 | 580 | 97% |

rpoB

| Gene  | Sarcocystis species | GenBank accession no. | P. catodon accession no. | Query cover (base pairs) | Percent identity |
|-------|---------------------|-----------------------|--------------------------|--------------------------|-----------------|
| S. arctica | MF986311 | NC_041214 | 570 | 99% |
| S. canis | KC191642 | NC_041214 | 462 | 98% |
| S. neurona | GQ851961 | NC_041214 | 909 | 96% |

ITS1

| Gene  | Sarcocystis species | GenBank accession no. | P. catodon accession no. | Query cover (base pairs) | Percent identity |
|-------|---------------------|-----------------------|--------------------------|--------------------------|-----------------|
| S. arctica | KF601306 | NW_021151540 | 210 | 97% |
| S. canis | MW136927 | NW_021151540 | 407 | 86% |
| S. neurona | AH009986 | | | |

All of the evidence collected here suggests that Sarcocystis sp. from P. catodon is not S. neurona and it is most closely related to S. canis. Gibson et al. (2011) in a wide study of 161 marine mammals from North America’s Pacific coast identified S. neurona within most positive samples, but also recorded a novel, unidentified species of Sarcocystis from P. catodon. Based on ITS1 sequences they proposed this was most similar to S. canis from a polar bear. However, at that time, only three species of Sarcocystis were available for comparisons with this marker. The ITS1 sequences from P. catodon are 99% identical to those of samples of Sarcocystis from subantarctic fur seals, but are not closely related to S. canis. Given the high mutation rate of the ITS1 region, the percent similarity between the sequences may suggest that the same species of Sarcocystis occurs in both these seals and P. catodon. This clearly warrants further investigation.

Another aspect arising from our results is misidentification of some coccidian as “S. canis-like”. In the case of the Indopacific bottlenosed dolphins studied by Calero-Bernal et al. (2017), coccidia found were identified as “S. canis-like” based on cox1 gene sequence similarity. With inclusion of additional sequences, it is clear that these sequences are not most closely related to S. canis (Fig. 1), nor are they closely related to any of the other species of Sarcocystis reported from marine mammals.

Our results suggest at least two other forms occur in marine mammals in addition to S. canis in seals and bears, and S. neurona in sea otters, sea lions and harbour seals. One unidentified species in Indopacific bottlenosed dolphins was previously considered as “S. canis-like” (Calero-Bernal et al. 2017), and the other unidentified species occurs in sperm whales and subantarctic fur seals, as indicated by the similarity of the ITS sequences (Reisfeld et al. 2019). The markers we analysed show different mutation rates, with 18S rRNA and cox1 lower, 28S rRNA and rpoB more intermediate, and ITS1 with a notably higher mutation rate. Whereas the genes with the lower rates do not always show differences between closely related species, the faster ITS1 is difficult to align between distantly related forms. However, the combination of these five markers resolved most relationships with notable support in our species-tree approach.

Our datamining technique demonstrates that extensive parasite genetic data can be extracted from contaminated host genomic studies. This may be particularly useful for studying parasites from marine mammals, which are currently poorly-known. At the same time the markers used may be useful for identifying species of Sarcocystis from other studies. Assessments of Sarcocystis coccidian from marine mammals should consider that multiple species are circulating in these hosts, and not presume that they all correspond to S. neurona or “S. canis-like” lineages.

Acknowledgements. JFF is supported by a PhD scholarship from FCT (2021/05681/BD).

REFERENCES

Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. 1990: Basic local alignment search tool. J. Mol. Biol. 215: 403–410.

Borner J., Burmester T. 2017: Parasite infection of public databases: a data mining approach to identify apicomplexan contam-
infections in animal genome and transcriptome assemblies. BMC Genomics 18: 100.

Burgess T.L., Tinker M.T., Miller M.A., Smith W.A., Bodkin J.L., Murray M.J., Nichol L.M., Saarinen J.A., Larson S., Tomoleoni J.A., Conrad P.A., Johnson C.K. 2020: Spatial epidemiological patterns suggest mechanisms of land-sea transmission for Sarcocystis neurona in a coastal marine mammal. Sci. Rep. 10: 3683.

Calero-Bernal R., Mauroo N.E., Hui S.W., Kuiken T., van de Bilt D.W.G., de Jong A.W., Osterhaus A.D.M.E., Sims L., Gendron-Fitzpatrick A., Carmena D., Cerqueira-Cezar C.K., Rosenthal B.M., Dubey J.P. 2017: Acute fatal sarcocystosis hepatitis in an Indo-Pacific bottlenose dolphin (Tursiops aduncus) in Hong Kong. Vet. Parasitol. 235: 64–68.

Castresana J. 2000: Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17: 540–552.

Clement M., Posada D., Crandall K.A. 2000: TCS: A computer program to estimate gene genealogies. Mol. Ecol. 9: 1657–1659.

Cuvertoret-Sanz M., López-Figueroa C., O’Byrne A., Canturlia, Martí-Garcia B., Pintado E., Pérez L., Ganges L., Cobos A., Abarca M., Raga J., Van Bressem M., Domingo M. 2020: Causes of cetacean stranding and death on the Catalan coast (western Mediterranean Sea), 2012–2019. Dis. Aquat. Org. 142: 239–253.

Dubey J.P. 2016: Toxoplasmosis of Animals and Humans. Toxoplasmosis of Animals and Humans. CRC Press, New York, 564 pp.

Dubey J.P., Lindsay D.S., Saville W.J.A., Reed S.M., Granstrom D.E., Speer C.A. 2001: A review of Sarcocystis neurona and equine protozoal myeloencephalitis (EPM). Vet. Parasitol. 95: 89–131.

Edgar R.C. 2004: MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucl. Acids Res. 32: 1792–1797.

Gibson A.K., Raverty S., Lambourn D.M., Huggins J., Marigo J., Ruoppolo V., Catão-Dias J. 2019: Toxoplasmosis and treatment of Sarcocystis neurona infection. Vet. Parasitol. 272: 183–194.

Múrias Dos Santos A., Cabezas M.P., Tavares A.I., Xavier R., Branco M. 2016: TsBU: A tool to extend TCS network layout and visualization. Bioinformatics 32: 627–628.

Mylniczenko N. D., Kearns K. S., Melli A. C. 2008: Diagnosis and treatment of Sarcocystis neurona in a captive harbor seal (Phoca vitulina). J. Zoo. Wildl. Med. 39: 228–235.

Rambaut A. 2018: FigTree v1.4.4, http://tree.bio.ed.ac.uk/software/figtree/.

Reiffelder L., Scartistán C., Ferreira Machado E., Sánchez-Sarmiento A., Costa-Silva S., Ewbank A., Navas-Suárez P., Guerra J., Barrel J., Réssio R., Favero C., Gestal S., Kolesnikovas C., Martínez J., Ruoppolo V., Caetano-Dias J. 2019: Toxoplasmosis and Sarcocystis spp. infection in wild pinnipeds of the Brazilian coast. Dis. Aquat. Org. 136: 235–241.

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

Rosenthal B.M., Dunams D.B., Pritt B. 2008: Restricted genetic diversity in the ubiquitous cattle parasite, Sarcocystis cruzi. Infect. Genetics Evol. 8: 588–592.

Yantis D., Moeller R., Braun R., Gardiner C.H., Aguirre A., Dubey J.P. 2003: Hepatitis associated with a Sarcocystis canis–like protozoan in a Hawaiian monk seal (Monachus schauinslandi). J. Parasitol. 89: 1258–1260.