UPPSALA
UNIVERSITET

Phylogenomics of Ascetosporea

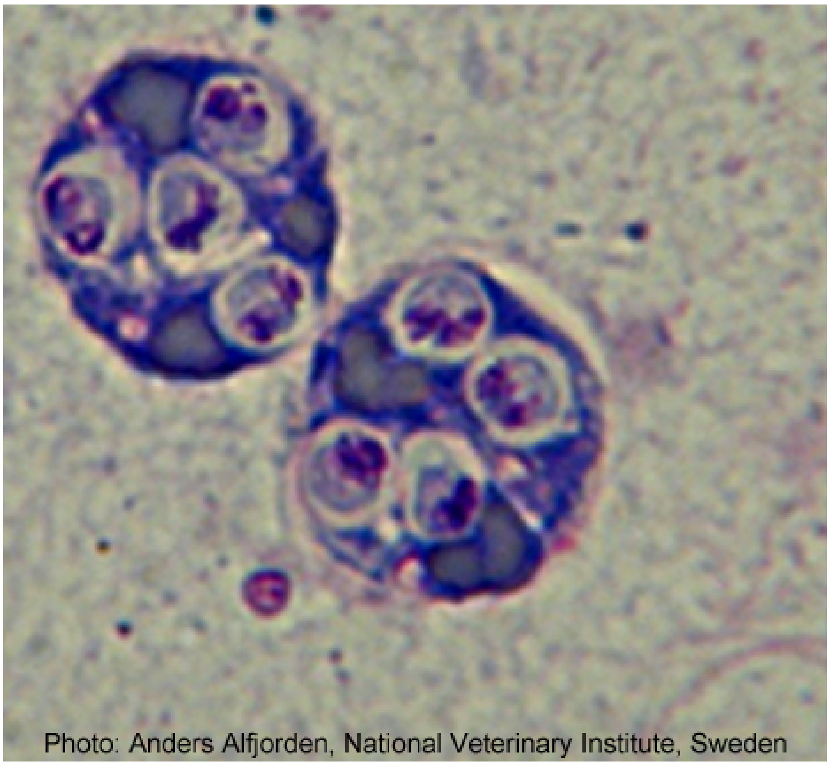


Photo: Anders Alfjorden, National Veterinary Institute, Sweden

Harshal Kunal Bhawe

Degree project in bioinformatics, 2022
Examensarbete i bioinformatik 30 hp till masterexamen, 2022
Biology Education Centre and Systematic Biology, Uppsala University
Supervisors: Fabien Burki and Ioana Onut Brännström

Abstract

Ascetosporea is a class of poorly studied unicellular eukaryotes that function as parasites of marine invertebrates. These parasites cause mass mortality events in aquaculture species such as oysters and mussels. The economic importance of these aquaculture species should lead to more attention on the genomics of *Ascetosporea* and their place on the evolutionary tree of life. With the onset of global warming and rising sea levels and temperatures, many emerging pathogens have been seen and until these are sequenced and analysed, it is difficult to make any conclusions about their relationships and evolution. As there aren't many genomes and transcriptomes available for *Ascetosporea*, their position in the larger eukaryotic tree of life remains hypothetical. To attempt to remedy this lack of information, the Burki lab has recently generated sequencing data through sample collection and sequencing for these organisms (genomes and transcriptomes).

A curated dataset of the various eukaryotic species was previously created and newly sampled and sequenced *Ascetosporean* genomes of *Paramarteilia* sp., *Marteilia pararefringens, Paramikrocytos canceri, etc.* from multiple sampling locations like Ireland, Norway, Sweden, and the UK were included. These could increase the genomic and transcriptomic data available for *Ascetosporea* and help to resolve the relationships within *Ascetosporea*. A few reasons why this group has not yet been placed on the tree of life are that the samples are from host tissue, which makes it difficult to sequence these parasites. These *Ascetosporeans* have also been seen to be very fast-evolving.

After building phylogenetic relationships with single gene trees to allow for the identification of possible contaminants and paralogs, it was seen that there was a lot of contamination in *Ascetosporea*, due to the sampling being from host tissue material (hosts are open to the environment). After cleaning and filtering the possible contaminated genes, the trees were remade and a possible link between a fungal group called *Microsporidia* and *Ascetosporea* was observed in a few genes. This was hypothesized to be lateral gene transfer between the two groups resulting from their similar lifestyles and infection of invertebrates.

There were complications like contamination and short blast hits that arose during analysis, and these could be caused by problems by fragmentation in the genome. This fragmentation could have negative effects on genome annotation predictions and consequently phylogenetic and phylogenomic analysis. Due to this and the challenging nature of collecting samples, the read coverage for the genomes is low but it can be used to perform phylogenetic and phylogenomic studies using currently available data and methods. Another expected result was that the sequenced data had contaminants, and a thorough and comprehensive search would have to be conducted on a dataset-wide level to remove any contaminants.

Decreasing ‘mussel’ mass?

Popular Science Summary

Harshal Bhawe

Seafood is widely eaten in many countries across the world, but Nordic countries in particular have large populations of blue mussels, salmon, crabs, oysters, etc. These are relatively large populations, so it is said to be more sustainable to consume than other land-based farmed populations. With the increasing effects of global warming and rising temperatures affecting the usually cold Nordic countries to a greater extent, these marine populations are getting infected by new microorganisms which previously did not influence them at such a prominent level. Of these newly surfaced microorganisms, *Rhizaria,* which is a group of protists (unicellular eukaryotes) were seen to be very common in marine aquaculture. One such parasitic species was responsible for the Denman Island Disease in Canada and caused great harm to the local oyster population.

Not much is known about these parasitic organisms, except the fact that they seemingly belong in the *Rhizarian* group. Within *Rhizaria*, there are multiple subgroups, some of which are parasitic. One of these groups is called *Ascetosporea*, which is said to be the culprit for the rising parasitic infections in marine populations. The focus of the study was this parasitic group *Ascetosporea*. Samples of infected animals were collected from the Baltic Sea, Ireland and the United Kingdom and were sequenced to produce genomes, which could be used for further research and testing.

Using this data and the known eukaryotic tree of life, which details the relationships of eukaryotes based on their ancestry and genetic makeup, an attempt was made to observe the relationships between these microorganisms. This was done in order to clarify which microorganisms *Ascetosporea* are closely related to and find out how they evolved with respect to other organisms present on the Earth. Studies conducted also resulted in the unexpected discovery of a possible link between a fungal parasite called *Microsporidia*, and *Ascetosporea*, which is a topic of research that can be explored further.

Degree project in bioinformatics, 2022
Examensarbete i bioinformatik 30 hp till masterexamen, 2022
Biology Education Centre and Systematic Biology, Department of Organismal Biology
Supervisor: Fabien Burki and Ioana Onut Brännström

Table of Contents

Abbreviations

Blast	Basic Local Alignment Search Tool
DNA	deoxyribonucleic acid
FASTA/Fasta	Fast-All Sequence Format
HSP	High Scoring Pair/s
LBA	Long Branch Attraction
LGT	Lateral Gene Transfer
RNA	ribonucleic acid
SAR	Stramenopila-Alveolata-Rhizaria
SGT	Single Gene Tree
TSAR	Telonemia-Stramenopila-Alveolata-Rhizaria

1 Introduction

Ascetosporea is a group of protists (unicellular eukaryotes), which are parasites of invertebrates (Hine *et al.* 2001, Kerr *et al.* 2018). These parasites can cause mass mortality events in natural and farmed populations, most notably in aquaculture species such as bivalves (oysters and mussels) as well as crustaceans (crabs) (Hine *et al.* 2001, Carnegie *et al.* 2003). Crustacean aquaculture plays a large role in World Food Security, especially in poorer countries, where seafood is considered a vital food source (Bondad-Reantaso *et al.* 2012) and effects on seafood populations will be felt across the world (Stentiford *et al.* 2012).

The TSAR (*Telonemia-Stramenopila-Alveolata-Rhizaria)* supergroup is one of the largest clades of the eukaryotic tree of life and contains many parasitic species including *Plasmodium* from the group *Alveolata*, which causes malaria in humans (Milner 2018) as well as oomycetes from the group *Stramenopila* which causes blight in plants (Larousse & Galiana 2017). *Ascetosporea* belongs to the *Rhizaria* group of the TSAR supergroup (Bass *et al.* 2019, Strassert *et al.* 2019). These *Ascetosporean* parasites need to be studied further but our current knowledge of the biology and evolution of these species is extremely limited. The main hurdle in resolving the relationships of these species is that there are almost no reference genomes or transcriptomes for any of the species and groups present in *Ascetosporea* (except *Mikrocytos mackini)* but there have been predictions about the relationships within *Rhizaria* (Fig. 1). Over time and with advancements in genome sequencing, it has become clearer that these parasitic microorganisms are more diverse than initially thought (Sierra *et al.* 2016, Bass *et al.* 2019). In many cases, the TSAR group is referred to as the SAR group due to Telonemia not being included in many analyses as it is a relatively new addition to the supergroup (Strassert *et al.* 2019).

These parasites are difficult to sequence as they cannot be cultured in the lab, so the DNA has to be extracted from samples of organisms living in natural habitats (water bodies in this case) (Kerr *et al.* 2018). As the samples are from infected host tissue, there is a lot of other genetic material from other organisms in the sample (the host, other organisms that the host consumes and the microbiome that lives inside). The Burki lab has collected samples from various host tissue (crabs, mussels, oysters, etc.) and has sequenced them with various sequencing techniques (See Appendix A). In the case of most marine invertebrates like mussels and oysters, they are filter-feeders so they consume everything and then filter everything they ingest (Lattos *et al.* 2021, Hamann & Blanke 2022) This makes it challenging to collect samples of *Ascetosporea* that have enough genetic material for sequencing and clean the data (remove contaminants) until a fully assembled genome is produced. Due to this, there is a dearth of genomes and transcriptomes for *Ascetosporeans*.

However, a few genomes have been assembled for these organisms (*M. mackini*, *Marteilia pararefringens* and *Paramarteilia* sp.) in the past (Ward *et al.* 2016, Kerr *et al.* 2018). A result of the challenge of collecting samples for *Ascetosporea* is that the sequenced genomes tend to have low read depths (Ward *et al.* 2016, Kerr *et al.* 2018).

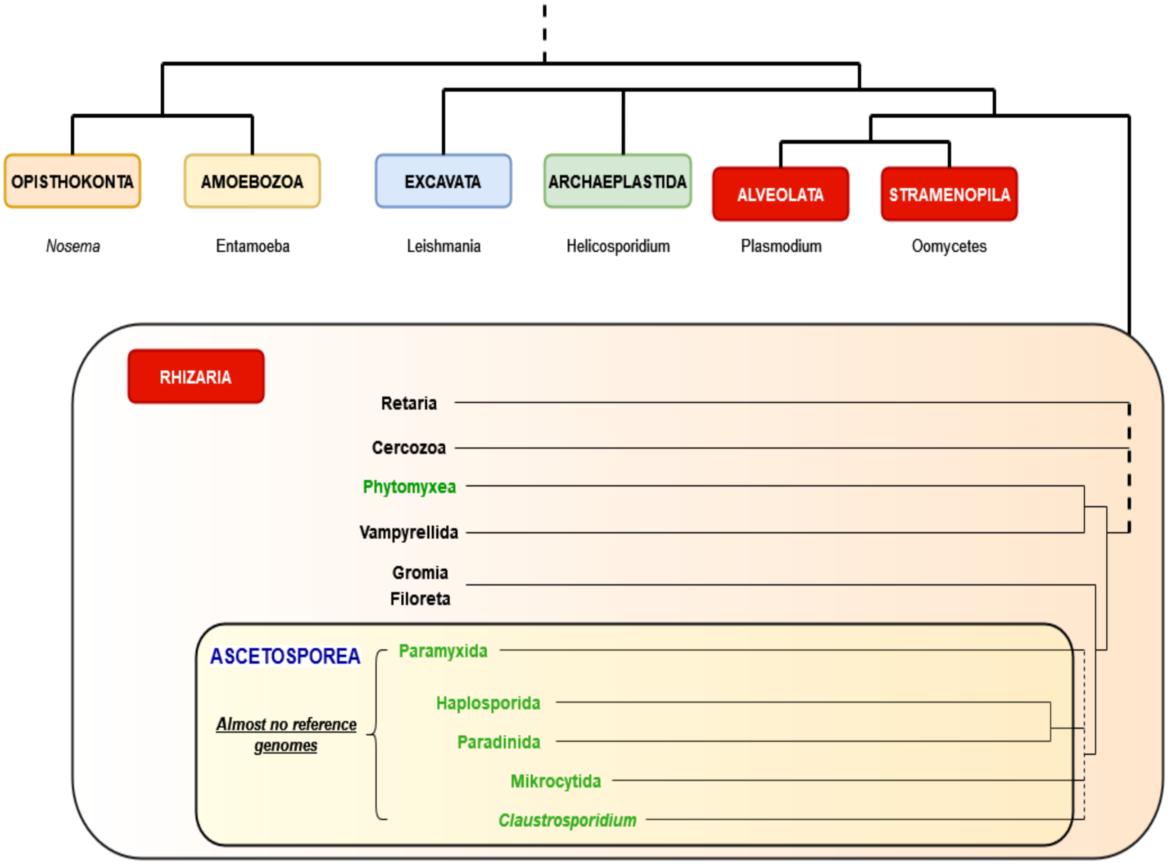


Figure 1: Predicted Phylogeny of the SAR group (coloured in red) and close relatives modified with permission from (Bass *et al.* 2019). Some examples of parasites from other groups can be seen under the group names. Most *Ascetosporean* species do not yet have a genome or a transcriptome. Parasitic groups within *Rhizaria* are marked in the colour "green." Created in DrawIO (Graph 2017) and Inkscape (Inkscape Project 2020).

Within *Rhizaria*, there are a few subgroups present which are parasitic: *Ascetosporea* and *Phytomyxea* (Fig. 1); *Phytomyxea* contains parasites that affect diatoms, other algae and plants (Sierra *et al.* 2016, Bass *et al.* 2019).

The word *'Ascetosporea'* means well-crafted spore bearers and the reasoning behind this is that the group was created to host species and subgroups that had spore ornamentation and formation that were considered to be elaborate. These two spore-forming groups were *Haplosporida* and *Paramyxida*. The sporulation for *Haplosporida* was seen with the walled, uninucleate spores while *Paramyxida* was more intricate, with spores forming inside 'mother cells', which are stem cells (Bass *et al.* 2019, Zakrzewski *et al.* 2019). As of today, there are more subgroups within *Ascetosporea* that aren't necessarily spore forming but were grouped together due to lifestyles and some phylogenetic analyses (Bass *et al.* 2019).

Mikrocytos mackini is one such example of an *Ascetosporean*. It is part of the subgroup *Mikrocytida* (Fig 1), present within *Ascetosporea*. *M. mackini* like many of the other *Ascetoporeans*, is a parasite that affects oysters and other marine aquaculture. It was seen that this protist parasite evolved at a very high rate and did not seem to have any close relatives (Burki *et al.* 2013). It has since been proven to have close relatives (Hartikainen *et al.* 2014).

In 2011, there was a microcell infection (a cell with a nucleus and protoplasm enclosed within a membrane (Abbott & Meyer 2014)) detected in *Cancer pagurus* (edible crab) in the United Kingdom. Sample extraction from *C. pagurus* and sequencing revealed that this infection was being caused by a relative of *M. mackini*, close enough to be called a sister group (Hartikainen *et al.* 2014). This species was described as *Paramikrocytos canceri*. After the establishment of this new *Ascetoporean* parasitic species, there was a PCR assay for screening of over 500 freshwater and marine samples made to target *P. canceri* (Hartikainen *et al.* 2014). It was seen that *P.canceri* was found in many different invertebrates, and not just *C. pagurus* (Hartikainen *et al.* 2014). Due to this research as well as increasing infections caused by these parasites, there is a need for the genomes of these organisms to be sequenced and to verify their evolutionary rates and find their place on the tree of life.

A potential issue of attempting to place the sequenced *Ascetosporeans* on the tree of life are paralogs. Paralogs are results of gene duplication events (Jensen 2001). These “paralogous” genes may not have the same function but are still considered to be closely related. Orthologs are considered to be more functionally similar to each other than paralogs. Unlike paralogs, orthologs are results of speciation events; and they are said to have similar, if not the same function between different species (Gao & Miller 2020). Previously, paralogs were seen as something to remove from phylogenetic analyses, but with advancements in phylogenetic predictions, keeping these paralogs can be advantageous to inferring meaningful phylogenetic signals (Hellmuth *et al.* 2015). These paralogs can be beneficial for phylogenetics, given that the orthologs and paralogs can be differentiated against as paralogs can develop new functions while orthologs normally have similar functions (Dufayard *et al.* 2005, Hellmuth *et al.* 2015, Gao & Miller 2020).

Placing these sequenced *Ascetosporean* microorganisms on the tree of life would contribute to a better understanding of their lifestyle and evolution by observing how they group together on the eukaryotic tree of life. As marine aquaculture populations are important ecologically and act as a sustainable protein source (Stentiford *et al.* 2012, Bondad-Reantaso *et al.* 2012); preserving them could limit drastic effects on food security in poorer parts of the world (Hine *et al.* 2001, Burki *et al.* 2013, Hartikainen *et al.* 2014).

AIMS:

The main aims of this project are:

1. to establish a complete and curated dataset of ~320 gene families with eukaryotes, known paralogs, newly sequenced genomes, as well as prokaryotes
2. to attempt to use this dataset to answer phylogenetic and evolutionary questions about *Ascetosporea* with respect to their evolution and placement on the eukaryotic tree of life.

2 Data

The data was a mixture of prior datasets as well as newly sequenced samples. The gene family dataset (Schön *et al.* 2021) was derived from a previous member of the Burki-lab.

2.1 Sequencing and pre-processing of data

Samples were collected, isolated and sequenced and the data then had to be pre-processed before using it for analysis.

2.1.1 **Sequenced genomes**

There were twelve sequenced *Ascetosporean* genomes present in the data. They had various sequencing techniques used and were sampled in distinct locations by different researchers (See Appendix A). These samples were collected by members of the Burki-lab at Uppsala University as well as David Bass from CEFAS, The Natural History Museum and Oxford University. The data provided was in the form of translated and annotated proteins from the species' sequences (Table 1).

2.1.2 **Pre-processing of data**

Pre-processing of data included sample cleaning, assembly, and annotation. Members of the Burki-lab performed these analyses and cleaning steps. The following workflow for *M. mackini* was used for almost all species with some slight alterations:

1. Adapter trimming: Cleaning and removing adapter sequences from the reads using trimmomatic (Bolger *et al.* 2014).
2. Filtering reads: Reads were filtered by mapping them the genome from a healthy host (not an infected host), and then removing the reads that mapped. This was done using bwa (Li & Durbin 2009). E.g., Samples of *P. canceri* were mapped to the genome of the edible crab (*C. pagurus*).
3. Assembly: SPAdes (Prjibelski *et al.* 2020) was used to create genome assemblies.
4. Blast cleaning: Blast(n) was used against NCBI nt (Altschul *et al.* 1990) to clean the genome based on the hits and e-values (1e-25).
5. Re-assembly: Re-assembly of reads that map to the contigs remaining after the first round of cleaning.
6. Blast cleaning: Using blast(x) against NCBI nr to blast the new contigs and removing hits (e-value 1e-25).
7. Re-assembly: Similar to step 5. Re-assembly of reads to contigs that remain.
8. Steps 6 and 7 were then repeated with blast(x) against NCBI nr.
9. Annotation: Genes were annotated using predictor training, structural annotation, and functional annotation with funannotate (Palmer & Stajich 2019).

The initial assembly for *M. mackini* was 42Mb long, and after performing the steps described above, it was reduced down to 15Mb. This process resulted in cleaned, assembled, and annotated genomes that could then be used to establish phylogenies for the *Ascetosporean* species.

2.2 Gene Family Dataset

The dataset contains 320 gene families which have been seen across almost all eukaryotes and function as ‘housekeeping genes’ i.e., genes required for basic cellular functions (Strassert *et al.* 2021). This dataset was initially created in 2007 by Fabien Burki and the version used was from Schön *et al*. 2021 and contains around 500-700 different eukaryotic species from various groups per gene family.

This was used as a starting point for the project to create single-gene trees for all genes and then ideally move on to a phylogenomic analysis. The known paralogs and prokaryotes were added manually using a dataset from Strassert *et al.* 2021.

2.2.1 **Prokaryotes**

Bacteria and *Archaea* are prokaryotes, and these were chosen to add to the dataset. These were also added to the dataset as it was a quick way to spot sample contamination and eliminate samples which could affect further analysis. These prokaryotes were added from the dataset obtained from Strassert *et al.* 2021. This dataset was chosen as it had the same gene models as the main gene family dataset, but it also contained prokaryotes in the form of *Archaea* and *Bacteria*. The *Bacteria* and *Archaea* present in the dataset were mostly extremophiles (halophilic, thermophilic, and acidophilic, etc.) and were also observed to have similar housekeeping genes as observed in most of the eukaryotes present in the dataset. These prokaryotes could also be observed with other protists from fossil records of the Proterozoic era (Summons & Walter 1990).

In most cases, prokaryotes were used as an outgroup for the SGTs as they do not belong to the class of *Eukaryota*. Any species grouping together with bacteria or archaea could ideally be treated as a prokaryotic contaminant and removed from the analysis.

2.2.2 **Paralogs**

The addition of paralogs was a necessary step in the process as it could tell us which sequences from the lab samples were paralogous in nature if they grouped together on the tree. This was important as it allowed us to separate the sequences in the dataset that may code for some other protein or divergent gene copies.

This was done using the same dataset used for extracting prokaryote sequences (Strassert *et al.* 2021), which also contained paralogs. These paralogs were all tagged manually by Fabien Burki by analysing various phylogenetic trees, while building the main 320 gene family dataset.

Species name	**Host**	**Annotation type**	**No. of protein-coding genes predicted**
M6MM	N/A (Free living heterotrophic amoeba)	Structural and functional annotation	9694
Paramarteilia sp. (now called *Paramarteilia canceri*)	Necor puber (Velvet crab)	N/A	4902
Bonamia ostreae	Ostrea edulis (European flat oyster)	N/A	5253
Marteilia pararefringens (M18)	*Mytilus edulis* (Blue mussel)	Structural	5657
Marteilia pararefringens (DB3)	*Mytilus* sp. (Mussel)	Structural and functional	4737
Marteilia pararefringens (DB4)	*Mytilus* sp. (Mussel)	Structural and functional	5298
Marteilia pararefringens (S151)	*Mytilus* sp. (Mussel)	Structural and functional	4943
Marteilia cochillia (DB6)	*Cerastoderma edule* (Common cockle)	Structural and functional	4559
Marteilia octospora (DB5a)	*Solen* sp. (Bivalves)	N/A	900 (Low coverage or fragmented genome)
Paramikrocytos canceri (2014)	*Cancer pagurus* (Edible crab)	Structural and functional	2409
Paramikrocytos canceri (2018)	*Cancer pagurus* (Edible crab)	Structural and functional	2340
Mikrocytos mackini	*Crassostrea gigas* (Pacific oysters)	Structural and functional	~5000

Table 1: Sequenced genomes, hosts, annotation type and number of genes predicted

3 Methods

All methods necessary for data preparation and analysis except for tree viewing were done using bash and perl in CentOS Linux release 7.9.2009 on UPPMAX (Uppsala Multidisciplinary Center for Advanced Computational Science), Uppsala University, Sweden.

3.1 Dataset preparation

Before any analysis could be performed, the dataset from (Schön *et al.* 2021) needed to be prepared accordingly in order for the software to be run successfully and to meet project goals. For this, the various different datasets had to be merged into a single dataset matching with the gene families.

3.1.1 **Gene and protein identification with blast**

To add the newly sequenced genomes to the dataset, we first needed to know which gene families were present in the sequenced genomes as well as the eukaryote dataset. The *Rhizarians* present in the dataset were extracted which were then used with blobtools v1.1.1 (Laetsch & Blaxter 2017, Laetsch *et al.* 2017) with the 'seqfilter' option to extract the sequences per gene family.

The extracted sequences were used with Blast v2.12.0+ (Altschul *et al.* 1990) as queries with the blast database set to the newly sequenced genomes as mentioned above. Blastp was used with the following options "-max_target_seqs 10 -outfmt 6". These results were then further filtered according to e-value < 1e-20. Using blast output format 6, the genes of interest could be seen. This was used to create a list of only the sequence IDs from the gene models and created a list of filtered hits.

These filtered hits were then used in blobtools v1.1.1 with the 'seqfilter' option to extract those sequences to a file. The file for each set of hits was stored with the name of the gene it was in (based on annotation), and the sequence header was modified to add the genome it was found in (*M. mackini, M. pararefringens* etc.). The blast and filtering methods gave us lists of proteins identified in our sequences and these could then be added to the dataset.

3.1.2 **Prokaryote extraction**

For prokaryotes, a simple 'grep' with the '-i' option to ignore case was used with the search term 'prokaryota' for the genes in the new dataset. This resulted in a list of prokaryotes per gene family in separate files. These were then used as an input to blobtools with the 'seqfilter' option to extract the sequences based on the gene family name.

3.1.3 **Paralog extraction**

In the same dataset as used for the prokaryotes, the paralogs are saved in two types: untagged and tagged paralogs. The tagged paralogs were easy to filter out as they were tagged with "-paralog" at the end of the FASTA header. The untagged paralogs were more challenging as the name of the paralogs changed based on the gene family they were found in, e.g., paralogs in mcm2 were named with mcm3, mcm4 etc. at the end of the FASTA header. A fellow lab member assisted me with the script for extracting paralogs, and it was successful in extracting both tagged and untagged paralogs, as well as some random hits based on the expression pattern used in the script. These erroneous hits needed to be removed from the data before any analysis could be performed.

3.1.4 **Merging the data**

Before proceeding with alignment and tree-making, the various types of data from 3.1.1 (Gene and protein identification with blast) – 3.1.3 (Paralog extraction) had to be merged per gene family. The original dataset, proteins identified in *Ascetosporea*, paralogs, and prokaryotes were merged using a script and stored. A further check was done manually to verify that all files were merged together successfully, which involved counting lines per file and comparing against the merged file. After checking the line counts, there was a discrepancy between the line count of the merged file and the addition of line counts of the separate files. This was theorised to be caused by duplicates being created when merging as well as paralog extraction.

To remedy this, seqkit v0.15.0 (Shen *et al.* 2016) was used with the ‘rmdup’ option to remove all the possible duplicates and store the duplicate names for further studies and reproducibility. After the duplicate removal step, a manual check was performed on a number of files to assess whether removing the duplicates had worked, and the line counts matched the expected numbers. After performing all the necessary merging processes and checks to ensure merging was successful, the dataset was ready for alignment.

3.2 Alignment of data

MAFFT v7.407 (Katoh & Standley 2013) was used for alignment of the genes. For a test run, mafft l-ins-i was used, which is a slow iterative method. This was soon discarded after seeing that it was very time-consuming, and the dataset was too large for this type of algorithm. Instead, mafft-auto was seen as the best choice as it decided the algorithm automatically based on number of sequences and sites. Based on these parameters, mafft-auto chose FFT-NS-2 which is a fast but rough aligning method. The options used with mafft-auto were ‘adjustdirectionaccurately‘, ‘thread 2‘ and ‘reorder.’ The aligning process with FFT-NS-2 did not take very long, and the alignment and tree-making was performed using a single script.

The output of MAFFT was passed through TrimAl v1.4.1 (Capella-Gutierrez *et al.* 2009). When aligning, MAFFT creates gaps which can be removed using trimAl. The ‘gappyout’ option for TrimAl was chosen.

3.3 Phylogenetic analysis

The TrimAl result was then passed into IQ-TREE v2.0-rc2 (Kalyaanamoorthy *et al.* 2017, Minh *et al.* 2020). The single gene trees were made using IQ-TREE and the Model-Finder algorithm (-m TEST) in IQ-TREE. The options used were ‘bb 1000’ for ultrafast bootstrap, ‘alrt 1000’ for assessing branch-wise support and ‘st aa’ for amino acid sequences. All the SGTs for the gene families were created with the LG + G4 model, which is a simple model.

This could be due to the size of the data and the computational power required for more complex models.

IQ-TREE creates trees in Newick format (Olsen 1990, Minh *et al.* 2020), so those trees had to converted to the Nexus format (Maddison *et al.* 1997) before the various taxa and groups could be coloured. For this, Figtree v1.4.4 (Rambaut 2010) was used to convert the trees and for tree-viewing purposes. The trees for all 320 genes were coloured using a script and then saved so that they could be viewed in Figtree.

The trees were either rooted with "*Prokaryota*" or "*Excavata-Discoba*" as suggested by (Schön *et al.* 2021) depending on whether any prokaryotes were present in that specific gene family. Based on this and observation of the trees, the decision was made whether the samples of interest were contaminated, paralogs or correct in phylogenetic placement using the support score created by IQ-TREE (bootstrap and sh-aLRT) and a guide to the phylogeny of the SAR group from Bass *et al.* 2019 (Fig. 1).

The support scores can be seen on the trees in the form of two numbers separated by a "/". The second number shows the UF-Boot support from IQ-TREE eg: 95% support from UFBoot means that there is a 95% chance that that branch is correct. The first number corresponds to the sh-aLRT test (Guindon *et al.* 2010). This is also a support measure for branch strength and normally sh-aLRT >=80% and UFBoot >=95% is considered to be good support for a branch.

3.4 Subset of data

There were 320 separate SGTs that needed to be analysed and checked. As they needed to be checked manually, this was very time intensive; and a subset of genes was created in order to meet time constraints. A total of 25 out of 320 genes were selected as a subset based on a few factors: single copy genes as well as prior research conducted in the lab group where there was a rough phylogeny of these genes. These single copy genes reduced/removed the risk of having paralogs and unnecessary contamination in the dataset due to them only having a copy number of one. These 25 genes were established as single copy orthologs by other Burki-lab members using Orthofinder (Emms & Kelly 2015, Emms & Kelly 2019), which creates orthogroups and gene families and gives a list of genes that may be present in single copy numbers. These were then sued with blast against the main dataset from (Schön *et al.* 2021) and the 25 genes were identified and extracted.

3.4.1 **Gene identification with blast and merging**

Based on the results from 3.3 (Phylogenetic Analysis), to reduce the risk of contamination and paralogs, as well as make it easier with respect to time constraints, it was decided to take a subset of the data and re-run the whole workflow.

For this, the steps from 3.1.1 (Gene and protein identification with blast) were replicated, with the extraction of *Rhizarians* from the dataset. The main difference between the workflow for the whole dataset and the subset was the blast and alignment filtering (Table 2). The filtering was more stringent as it reduced the number of short hits resulting from blast. The new parameters during blast used were q-cov-hsp (qcov_hsp_perc in blast), length and bitscore. Q-cov-hsp (query coverage of high scoring pairs) is a parameter used to quantify the most accurate blast hits that do not contain gaps. A high scoring pair (HSP) is an ungapped local alignment with the highest alignment score. The higher the q-cov-hsp value is, the higher is the coverage of the HSP to the query, which in turn can mean a higher chance of the hit being a true ortholog. During this step, a q-cov-hsp value of 90 was chosen, which would mean that the HSP's would cover at least 90% of the query sequence. These results were then further filtered according to length and bitscore to prevent short hits from interfering with alignment and tree-making. For this, a filter for match length and bitscore was used. These filter values were decided based on the average size of most of the genes and manually inspecting the bitscore for the hits. Using the 'sseqid' from the blast output, the genes of interest could be seen. These were then extracted and sorted. This gave a list of only the sequence IDs from the gene model and created a list of filtered hits. The different blast parameters used can be seen in Table 2.

Step	Filtering parameters
3.1.1 Gene and protein identification with blast	Blast options: -max_target_seqs 10 -evalue 1e-20
3.4.1 Blast and Merge	Blast options: -max_target_seqs 10 -evalue 1e-20 -qcov_hsp_perc 90 External filtering: alignment length >200, bitscore >93 and manual alignment filters

Table 2: Table showing the parameters used with blast for the whole dataset versus the subset of data

The workflow from 3.1.2 (Prokaryote extraction) - 3.1.4 (Merging the data) was repeated, including the merging and removal of duplicates for alignment and tree-making.

3.4.2 Alignment and Phylogenetic Analysis

MAFFT was used, with the same options as 3.2 Alignment of data. TrimAl was used for gap removal of the gaps created by MAFFT with the 'gappyout' option to remove excessive gaps. The alignments for the 25 gene subset were then viewed in AliView v1.28 (Larsson 2014) and checked for alignment lengths and coverages, etc. The samples from the sequenced genomes which had excessive gaps were removed from the alignment.

As seen in 3.3 Phylogenetic Analysis, IQ-TREE was used as the software of choice to create the single gene trees from the processed and edited alignment files. IQ-TREE was run with 4 CPU cores for this step. Most trees were created with the LG+ F + I + G4 model automatically chosen by IQ-TREE. The trees were then coloured and analysed using the same outgroups as in 3.3 Phylogenetic Analysis.

3.5 Possible Lateral Gene Transfer

Lateral Gene Transfer was analysed and tested using the *Microsporidia* that grouped with the sequenced *Ascetosporean* genomes.

3.5.1 Addition of *Microsporidia* with blast

The sequences for *Edha aedis* and *Nosema bombycis* were used as blast queries with the blastp webserver on NCBI (Madden 2003). The ten best *Microsporidian* hits that were not *E. aedis* and *N. bombycis* were chosen and added to the dataset. These were selected based on the bitscore (>80), alignment length (>60% coverage of the reference sequence) and e-value (1e-20).

3.5.2 Alignment and Phylogenetic Analysis

Repeating previous processes, the blast hits were merged with the original single genes, and this was aligned and trimmed using MAFFT and TrimAl. After trimming, the trimmed alignment was used to create the single gene trees for the six genes that showed this grouping with *Microsporidia*. IQ-TREE was used to create the trees with the same options in 3.3 (Phylogenetic Analysis).

3.6 Concatenated trees

3.6.1 Full dataset

To create a supermatrix, a script was used from GitHub (Nylander 2010), which concatenated alignment files together. This supermatrix could be used in two ways to create a phylogenomic analysis: unpartitioned or partitioned. In an unpartitioned analysis, the supermatrix is used as is to create a phylogeny; the software runs it as it runs any other normal aligned fasta file (Kainer & Lanfear 2015).

For a partitioned analysis, a partition file needs to be created with the coordinates of the genes that are present in the supermatrix. The script used to create the supermatrix displays a list of the coordinates for the genes in the output, so this could be used to create the partition file. IQ-TREE was then used to create a partitioned analysis with the partition file and the LG+G4 model was chosen due to it being the model used by IQ-TREE for the single gene trees and time constraints.

This step repeatedly ran out of system memory on UPPMAX and was terminated multiple times by the system administrators.

3.6.2 TSAR analysis

To create a smaller and easily visualised tree, the TSAR clade was chosen with *Telonemia* being an outgroup. This should give us enough data to investigate the evolutionary questions for *Ascetosporea*. Using the supermatrix created in 3.6.1 Full dataset, the various groups from SAR were extracted, *Telonemia* was extracted, and the *Ascetosporea* were also extracted. These were then merged, and duplicates were removed with SeqKit. This merged file did not need to be aligned as the supermatrix in 3.6.1 Full dataset was created using the trimmed alignment of the 25 gene subset. The merged file for TSAR was executed with IQ-TREE to create a phylogenomic tree with the -AUTO model-finder and with -auto for the number of cores being used.

In spite of being smaller than the whole supermatrix, this analysis was also terminated due to memory restrictions.

4 Results

4.1 Full dataset

For the *Rhizarians*, a gene called *eif6* has been highlighted as it was one of the few genes that were present in all analyses (including the subset of data and LGT analysis). Only a few species of *Rhizaria* were found in the dataset, which could have an effect on the blast results.

With the paralogs, the script used sometimes flagged apparent non-paralogs as paralogs i.e., some of the proteins it identified as paralogs were not actually paralogs, so it was imperative to screen for duplicates before alignment and tree-making.

4.1.1 **Alignment**

On looking at the alignment for the *eif6* gene family, it was seen that the species of interest had an extensive number of gaps, which could've been decreasing alignment scores and could influence further analysis, so gaps were removed.

When the same alignment for *eif6* after trimming was observed, we could see that there were a lot of gaps in the alignment of the sequences in *Ascetosporea* even after gap removal. The gappy nature of the species even after gap removal could mean that the gene models were inaccurate, presence of paralogs/contamination and incomplete contigs formed during genome assembly.

4.1.2 **Phylogenetic Analysis**

On closer inspection of the *Ascetosporea* (in purple in Fig. 2 and 3), two areas of the tree contained multiple (>=2) species of *Ascetosporea*. As we could see, it grouped close to *Cryptist*-nucleomorphs (in green in Fig. 2) and *Amoebozoa* (yellow in Fig. 2). *Cryptist*-nucelomorphs are fast-evolving organisms and are results of secondary endosymbiosis (Irwin & Keeling 2019). These *Cryptists* are not part of the TSAR supergroup, so this was an unexpected result. All these branches mostly had good support, as seen by the numbers on the branches. The first set of numbers shows the branch strength with >80 being considered good, and the second set of numbers shows the bootstrap (>95 for ultra-fast bootstrap is considered good support) (Fig. 2 and 3).

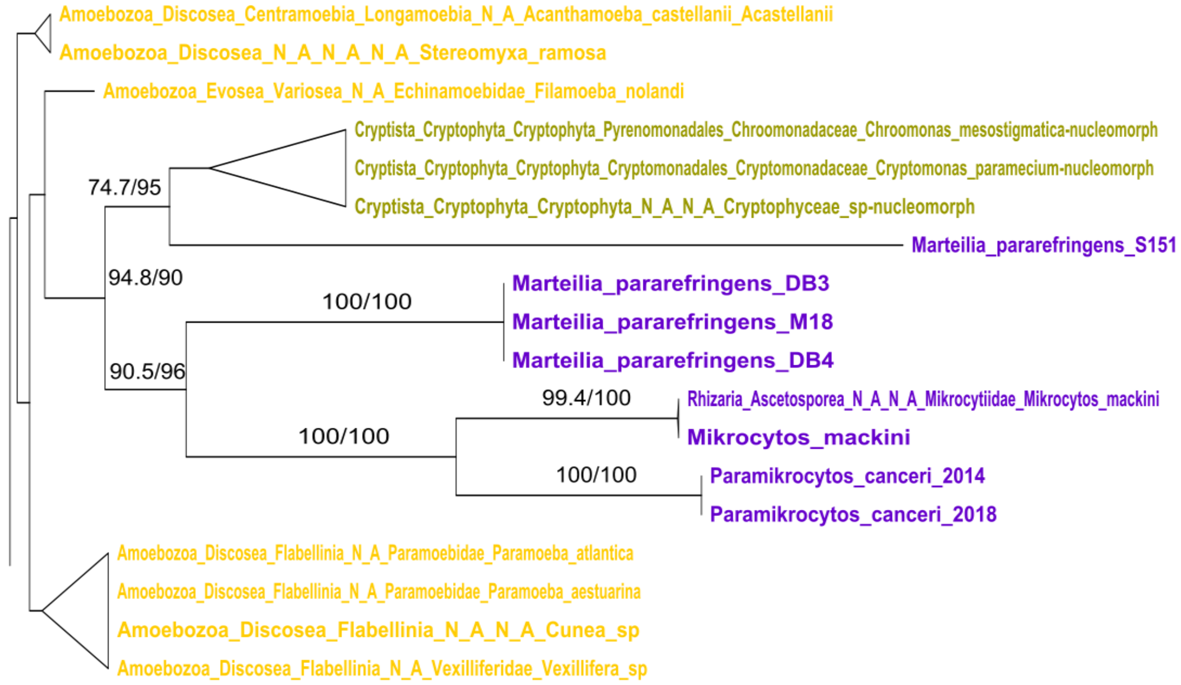

Amoebozoa_Discosea_Centramoebia_Longamoebia_N_A_Acanthamoeba_castellanii_Acastellanii
Amoebozoa_Discosea_N_A_N_A_N_A_Stereomyxa_ramosa
Amoebozoa_Evosea_Variosea_N_A_Echinamoebidae_Filamoeba_nolandi
Cryptista_Cryptophyta_Cryptophyta_Pyrenomonadales_Chroomonadaceae_Chroomonas_mesostigmatica-nucleomorph
Cryptista_Cryptophyta_Cryptophyta_Cryptomonadales_Cryptomonadaceae_Cryptomonas_paramecium-nucleomorph
Cryptista_Cryptophyta_Cryptophyta_N_A_N_A_Cryptophyceae_sp-nucleomorph
74.7/95
Marteilia_pararefringens_S151
94.8/90
100/100
Marteilia_pararefringens_DB3
Marteilia_pararefringens_M18
Marteilia_pararefringens_DB4
90.5/96
99.4/100
Rhizaria_Ascetosporea_N_A_N_A_Mikrocytiidae_Mikrocytos_mackini
100/100
Mikrocytos_mackini
100/100
Paramikrocytos_canceri_2014
Paramikrocytos_canceri_2018
Amoebozoa_Discosea_Flabellinia_N_A_Paramoebidae_Paramoeba_atlantica
Amoebozoa_Discosea_Flabellinia_N_A_Paramoebidae_Paramoeba_aestuarina
Amoebozoa_Discosea_Flabellinia_N_A_N_A_Cunea_sp
Amoebozoa_Discosea_Flabellinia_N_A_Vexilliferidae_Vexillifera_sp

Figure 2: ***eif6*** **single gene tree(one area of interest with *Ascetosporea*). The species of interest can be seen in purple, *Cryptist*-nucleomorphs are in green and *Amoebozoa* is in yellow. The numbers at the branches show the support scores with the number after the '/' representing bootstrap support. Some branch support values have been removed for greater legibility.**

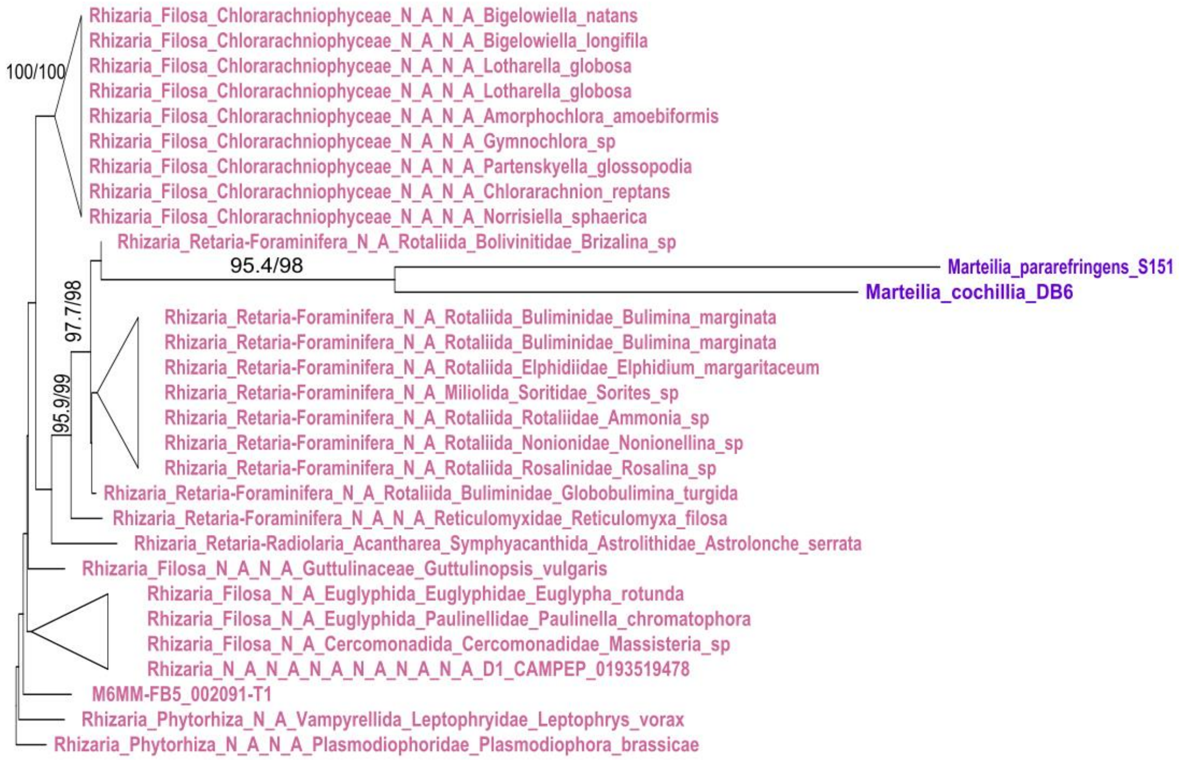

Rhizaria_Filosa_Chlorarachniophyceae_N_A_N_A_Bigelowiella_natans
Rhizaria_Filosa_Chlorarachniophyceae_N_A_N_A_Bigelowiella_longifila
100/100
Rhizaria_Filosa_Chlorarachniophyceae_N_A_N_A_Lotharella_globosa
Rhizaria_Filosa_Chlorarachniophyceae_N_A_N_A_Lotharella_globosa
Rhizaria_Filosa_Chlorarachniophyceae_N_A_N_A_Amorphochlora_amoebiformis
Rhizaria_Filosa_Chlorarachniophyceae_N_A_N_A_Gymnochlora_sp
Rhizaria_Filosa_Chlorarachniophyceae_N_A_N_A_Partenskyella_glossopodia
Rhizaria_Filosa_Chlorarachniophyceae_N_A_N_A_Chlorarachnion_reptans
Rhizaria_Filosa_Chlorarachniophyceae_N_A_N_A_Norrisiella_sphaerica
Rhizaria_Retaria-Foraminifera_N_A_Rotaliida_Bolivinitidae_Brizalina_sp
95.4/98
Marteilia_pararefringens_S151
Marteilia_cochillia_DB6
97.7/98
Rhizaria_Retaria-Foraminifera_N_A_Rotaliida_Buliminidae_Bulimina_marginata
Rhizaria_Retaria-Foraminifera_N_A_Rotaliida_Buliminidae_Bulimina_marginata
Rhizaria_Retaria-Foraminifera_N_A_Rotaliida_Elphidiidae_Elphidium_margaritaceum
95.9/99
Rhizaria_Retaria-Foraminifera_N_A_Miliolida_Soritidae_Sorites_sp
Rhizaria_Retaria-Foraminifera_N_A_Rotaliida_Rotaliidae_Ammonia_sp
Rhizaria_Retaria-Foraminifera_N_A_Rotaliida_Nonionidae_Nonionellina_sp
Rhizaria_Retaria-Foraminifera_N_A_Rotaliida_Rosalinidae_Rosalina_sp
Rhizaria_Retaria-Foraminifera_N_A_Rotaliida_Buliminidae_Globobulimina_turgida
Rhizaria_Retaria-Foraminifera_N_A_N_A_Reticulomyxidae_Reticulomyxa_filosa
Rhizaria_Retaria-Radiolaria_Acantharea_Symphyacanthida_Astrolithidae_Astrolonche_serrata
Rhizaria_Filosa_N_A_N_A_Guttulinaceae_Guttulinopsis_vulgaris
Rhizaria_Filosa_N_A_Euglyphida_Euglyphidae_Euglypha_rotunda
Rhizaria_Filosa_N_A_Euglyphida_Paulinellidae_Paulinella_chromatophora
Rhizaria_Filosa_N_A_Cercomonadida_Cercomonadidae_Massisteria_sp
Rhizaria_N_A_N_A_N_A_N_A_N_A_D1_CAMPEP_0193519478
M6MM-FB5_002091-T1
Rhizaria_Phytorhiza_N_A_Vampyrellida_Leptophryidae_Leptophrys_vorax
Rhizaria_Phytorhiza_N_A_N_A_Plasmodiophoridae_Plasmodiophora_brassicae

Figure 3: ***eif6*** **single gene tree (another area of the tree with *Ascetosporea*). The *Ascetosporea* can be seen in purple, while the other Rhizarian species and M6MM can be seen in pink. The numbers at the branches show the support scores with the number after the '/' representing bootstrap support. Ideally bootstrap >95 for UFBoot. Some branch support values have been removed for greater legibility.**

The other section of the tree with *Ascetosporea* was a clade of *M. pararefringens* (purple in Fig. 3) and *M. cochillia* (purple) in *Rhizaria*, and it placed within *Rhizaria* (pink) as initially predicted to. We can see the species of interest with long branches compared to the *Rhizaria* and only a couple place with the *Rhizaria* (Fig. 3). Two taxa being placed within *Rhizaria* was enough to warrant making evolutionary inferences about these *Ascetosporeans*. Looking at the other results across gene families led me to believe that there were anomalies in the data. These anomalies were creating unexpected results in the placement and clustering of taxa. This could have been due to the fragmented nature of the genes in the alignments, that the placements for most the species of interest were not as predicted.

The other taxa from the species of interest grouped together with separate groups in *Archaeplastida-Rhodophyta*, *Opisthokonta-Metazoa*, and *Excavata-Discoba*. These were treated as sample contaminants that made their way into the samples as only one taxon each seemed to be clustering together with those groups.

For other genes, there were multiple clades of *Ascetosporea* spread all over the tree, which does not give us much data to answer evolutionary questions. In those cases, the genomes were studied further, and it was seen that a few of them had poor read coverage and could be fragmented on a gene level. This resulted in multiple hits of genes from one species to pass through the blast filters and group together in the trees (See Appendix B).

Due to these issues and time constraints, a subset of the data was taken based on gene copy numbers using prior research done in the lab with ortholog clustering, and more stringent filtering was done on the sequences before and after alignment.

All the trees created, and scripts used are available to view and analyse. (See Appendix C)

4.2 Subset of data

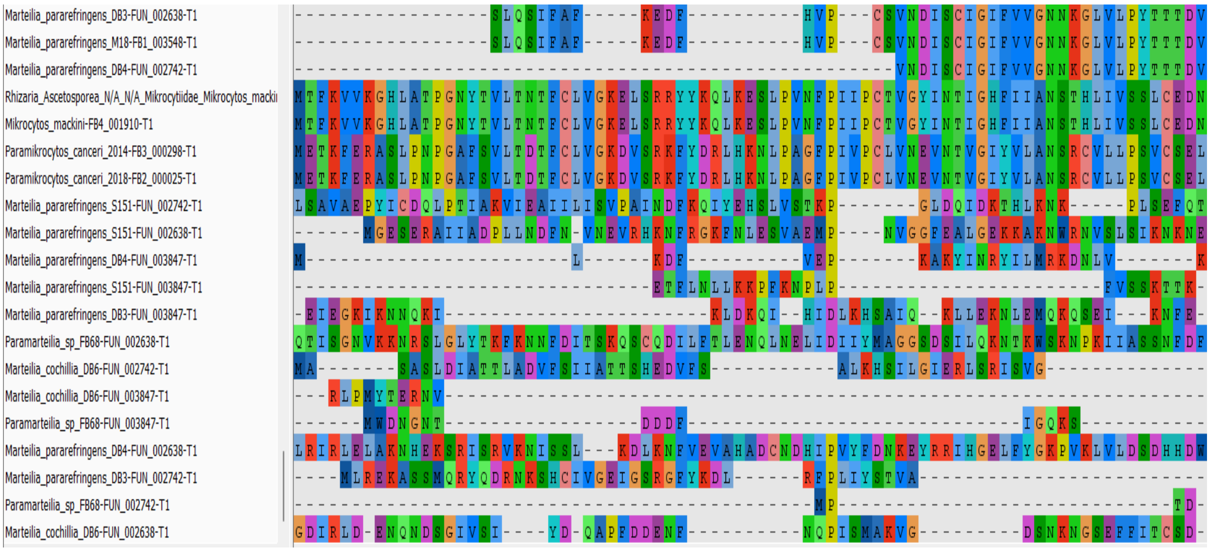
For the subset of data, the steps listed in 3.4 Subset of data were followed with more stringent blast and alignment filtering due to the challenges faced while computing the data in the main dataset.

4.2.1 **Blast**

For blast, the filters used were different, with there being thresholds for coverage of high scoring pairs (q-cov HSPs) and e-value. With just the extra filtering steps, the amount of *Ascetosporean* hits was reduced for all genes, as previously the blast hits seemed to be very short in length. These results were then further filtered using the length of found matches and bitscore. This filtering removed a lot of possible noise from the data and could enable us to establish possible homologies between species. After performing these filtering procedures, only a few blast hits remained, some of which were duplicates.

4.2.2 Alignment

There were two types of alignments in the smaller dataset, default mafft-auto and a TrimAl gappy-out alignment. The alignments were done using the combination of new filtered blast searches, paralogs, prokaryotes, and the original gene family dataset from Schön *et al*. 2021.



```
Marteilia_pararefringens_DB3-FUN_002638-T1 -----------------SLQSIFAF-----KEDF----------HVP---CSVNDISCIGIFVVGNNKGLVLPYTTTDV
Marteilia_pararefringens_M18-FB1_003548-T1 -----------------SLQSIFAF-----KEDF----------HVP---CSVNDISCIGIFVVGNNKGLVLPYTTTDV
Marteilia_pararefringens_DB4-FUN_002742-T1 ----------------------------------------------------VNDISCIGIFVVGNNKGLVLPYTTTDV
Rhizaria_Ascetosporea_N/A_N/A_Mikrocytiidae_Mikrocytos_macki MTFKVVKGHLATPGNYTVLTNTFCLVGKELSRRYYKQLKESLPVNFPIIPCTVGYINTIGHFIIANSTHLIVSSLCEDN
Mikrocytos_mackini-FB4_001910-T1 MTFKVVKGHLATPGNYTVLTNTFCLVGKELSRRYYKQLKESLPVNFPIIPCTVGYINTIGHFIIANSTHLIVSSLCEDN
Paramikrocytos_canceri_2014-FB3_000298-T1 METKFERASLPNPGAFSVLTDTFCLVGKDVSRKFYDRLHKNLPAGFPIVPCLVNEVNTVGIYVLANSRCVLLPSVCSEL
Paramikrocytos_canceri_2018-FB2_000025-T1 METKFERASLPNPGAFSVLTDTFCLVGKDVSRKFYDRLHKNLPAGFPIVPCLVNEVNTVGIYVLANSRCVLLPSVCSEL
Marteilia_pararefringens_S151-FUN_002742-T1 LSAVAEPYICDQLPTIAKVIEAIILISVPAINDFKQIYEHSLVSTKP-------GLDQIDKTHLKNK-----PLSEFQT
Marteilia_pararefringens_S151-FUN_002638-T1 ------MGESERAIIADPLLNDFN-VNEVRHKNFRGKFNLESVAEMP----NVGGFEALGEKKAKNWRNVSLSIKNKNE
Marteilia_pararefringens_DB4-FUN_003847-T1 M-----------------------L------KDF----------VEP-------KAKYINRYILMRKDNLV-------K
Marteilia_pararefringens_S151-FUN_003847-T1 -------------------------------ETFLNLLKKPFKNPLP-----------------------FVSSKTTK-
Marteilia_pararefringens_DB3-FUN_003847-T1 -EIEGKIKNNQKI-----------------------KLDKQI--HIDLKHSAIQ--KLLEKNLEMQKQSEI---KNFE-
Paramarteilia_sp_FB68-FUN_002638-T1 QTISGNVKKNRSLGLYTKFKNNFDITSKQSCQDILFTLENQLNELIDIIYMAGGSDSILQKNTKWSKNPKIIASSNFDF
Marteilia_cochillia_DB6-FUN_002742-T1 MA-------SASLDIATTLADVFSIIATTSHEDVFS-----------ALKHSILGIERLSRISVG--------------
Marteilia_cochillia_DB6-FUN_003847-T1 ---RLPMYTERNV------------------------------------------------------------------
Paramarteilia_sp_FB68-FUN_003847-T1 ------MWDNGNT-----------------DDDF-----------------------------IGQKS-----------
Marteilia_pararefringens_DB4-FUN_002638-T1 LRIRLELAKNHEKSRISRVKNISSL---KDLKNFVEVAHADCNDHIPVYFDNKEYRRIHGELFYGKPVKLVLDSDHHDW
Marteilia_pararefringens_DB3-FUN_002742-T1 ----MLREKASSMQRYQDRNKSHCIVGEIGSRGFYKDL------RFPLIYSTVA-------------------------
Paramarteilia_sp_FB68-FUN_002742-T1 ---------------------------------------------MP-----------------------------TD-
Marteilia_cochillia_DB6-FUN_002638-T1 GDIRLD-ENQNDSGIVSI----YD-QAPFDDENF----------NQPISMAKVG---------DSNKNGSEFFITCSD-
```

Figure 4: TrimAl Alignment of eif6 zoomed in on species of interest. The sequences at the bottom show the species of interest: *M. mackini, P. canceri, M. pararefringens, M. cochillia* and *Paramarteilia* sp.

In the case of *eif6*, the alignment did not need any further filtering with respect to the species of interest as it wasn't excessively 'gappy'.

E.g., in the gene *rplp0*, with the editing of the alignment, six sequences of *M. cochillia, M. pararefringens and Bonamia ostreae* needed to be removed due excessive gaps (~50%).

4.2.3 Phylogenetic analysis

On viewing the edited and filtered Single Gene Tree (SGT) for *eif6*, it was observed that the *Ascetosporeans* (purple in Fig. 5) were all grouping together, but this time they were seen to be close relatives of *Opisthokonta-Fungi-Microsporidia* (brown in Fig. 5), which are also parasites of invertebrates, mainly insects. This clade also had very good support shown by the second set of numbers on the branches (bootstrap) (Fig. 5).

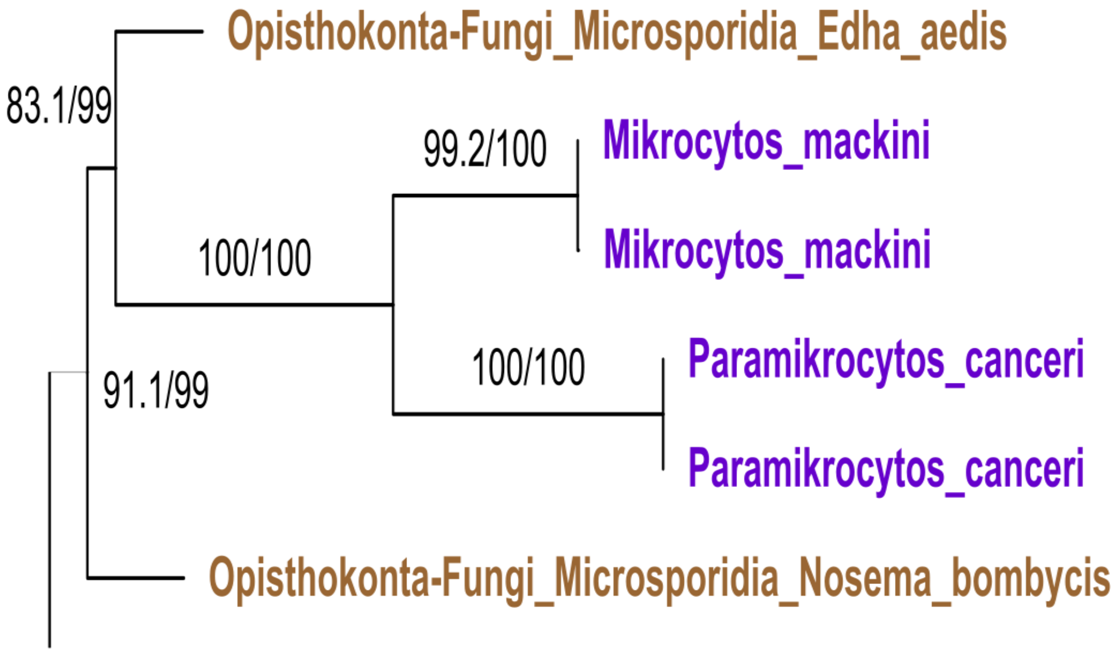


Figure 5: Species of interest on the filtered *eif6* single gene tree (seen in purple) grouping with *Microsporidia* (brown). The numbers show the branch strength test and bootstrap support with a value of 80/95 being considered good.

For the filtered tree of *eif6* (Fig. 5), most of the species of interest grouped differently than they did before (Fig. 2 and 3). As compared to Fig. 2 and 3, most *Ascetosporean* sequences were removed during the filtering process and there was one only clade left which could be useful in answering phylogenetic questions. This tree (Fig. 5) did not have all the species of interest that were included in the analysis due to stricter filtering.

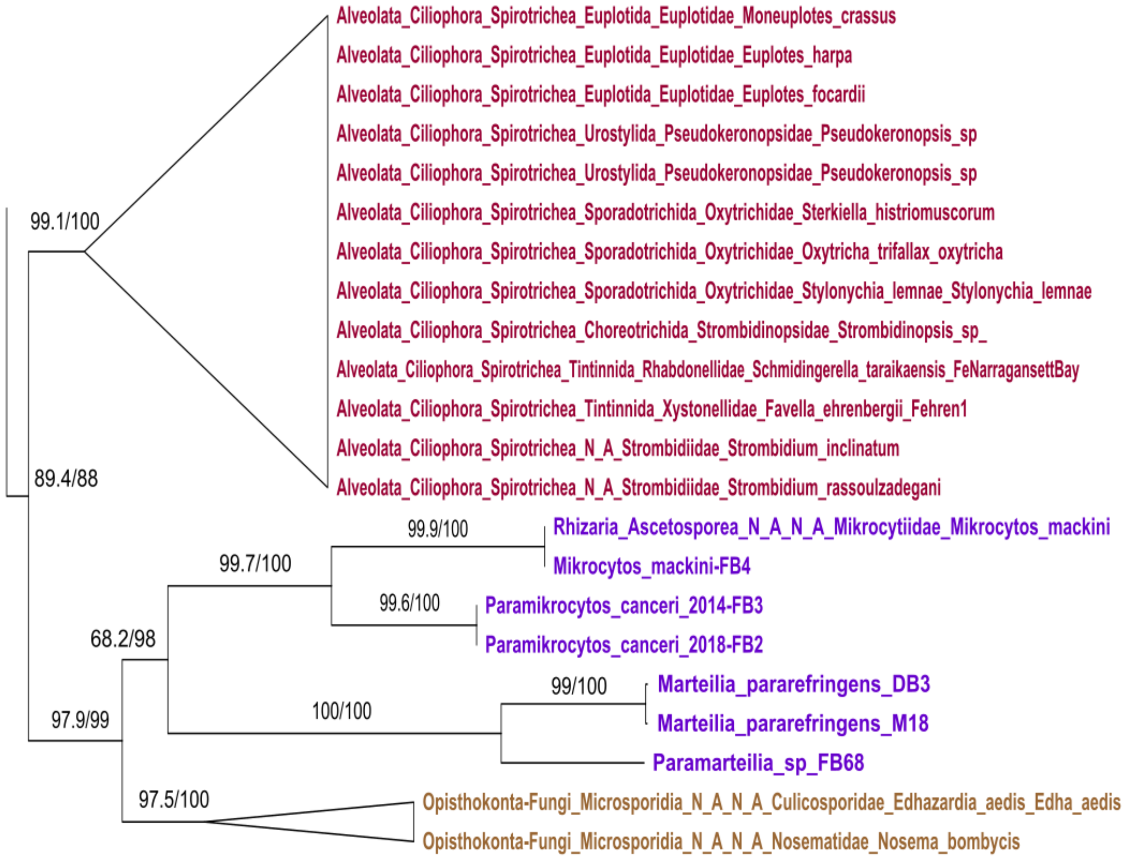

Alveolata_Ciliophora_Spirotrichea_Euplotida_Euplotidae_Moneuplotes_crassus
Alveolata_Ciliophora_Spirotrichea_Euplotida_Euplotidae_Euplotes_harpa
Alveolata_Ciliophora_Spirotrichea_Euplotida_Euplotidae_Euplotes_focardii
Alveolata_Ciliophora_Spirotrichea_Urostylida_Pseudokeronopsidae_Pseudokeronopsis_sp
Alveolata_Ciliophora_Spirotrichea_Urostylida_Pseudokeronopsidae_Pseudokeronopsis_sp
99.1/100
Alveolata_Ciliophora_Spirotrichea_Sporadotrichida_Oxytrichidae_Sterkiella_histriomuscorum
Alveolata_Ciliophora_Spirotrichea_Sporadotrichida_Oxytrichidae_Oxytricha_trifallax_oxytricha
Alveolata_Ciliophora_Spirotrichea_Sporadotrichida_Oxytrichidae_Stylonychia_lemnae_Stylonychia_lemnae
Alveolata_Ciliophora_Spirotrichea_Choreotrichida_Strombidinopsidae_Strombidinopsis_sp_
Alveolata_Ciliophora_Spirotrichea_Tintinnida_Rhabdonellidae_Schmidingerella_taraikaensis_FeNarragansettBay
Alveolata_Ciliophora_Spirotrichea_Tintinnida_Xystonellidae_Favella_ehrenbergii_Fehren1
Alveolata_Ciliophora_Spirotrichea_N_A_Strombidiidae_Strombidium_inclinatum
89.4/88
Alveolata_Ciliophora_Spirotrichea_N_A_Strombidiidae_Strombidium_rassoulzadegani
99.9/100
Rhizaria_Ascetosporea_N_A_N_A_Mikrocytiidae_Mikrocytos_mackini
99.7/100
Mikrocytos_mackini-FB4
99.6/100
Paramikrocytos_canceri_2014-FB3
68.2/98
Paramikrocytos_canceri_2018-FB2
99/100
Marteilia_pararefringens_DB3
97.9/99
100/100
Marteilia_pararefringens_M18
Paramarteilia_sp_FB68
97.5/100
Opisthokonta-Fungi_Microsporidia_N_A_N_A_Culicosporidae_Edhazardia_aedis_Edha_aedis
Opisthokonta-Fungi_Microsporidia_N_A_N_A_Nosematidae_Nosema_bombycis

Figure 6: Filtered *rplp0* tree zoomed in on the species of interest (purple) with various other groups: *Alveolata* (red) and *Microsporidia* (brown). The numbers show the branch strength test and bootstrap support with a value of 80/95 being considered good.

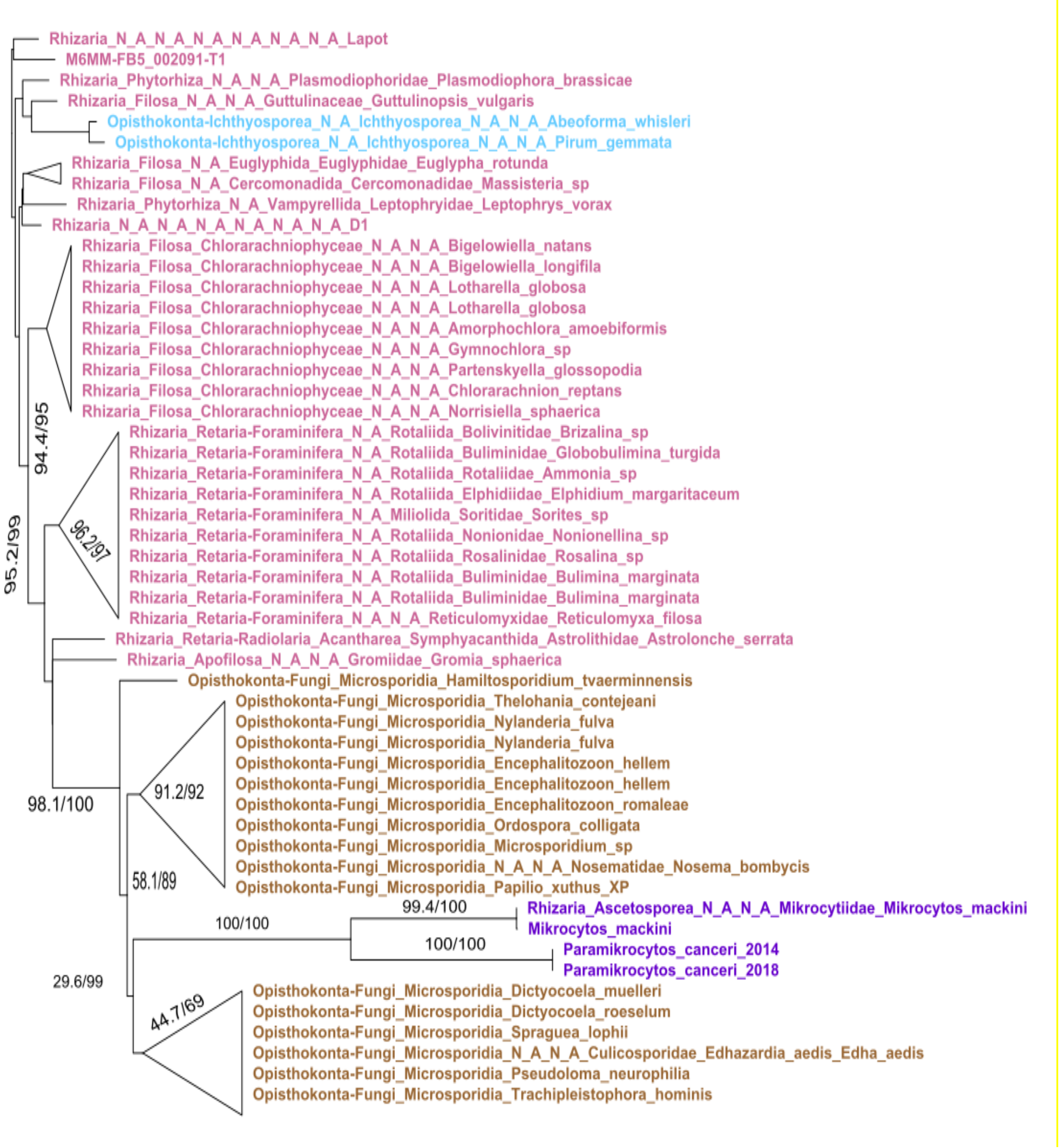
In the tree for gene *rplp0*, the grouping of *Ascetosporea* (purple in Fig. 6) was as a sister clade to *Alveolata* (red in Fig. 6) and close to the predicted position of TSAR (Bass *et al.* 2019). The result, similar to *eif6*, was that the *Ascetosporean* species unexpectedly grouped together with a fungal species (*Microsporidia* [brown in Fig. 6]) with good support. On inspecting the trees further in other genes, the same grouping with *Microsporidia* (as seen in Fig. 5 and 6) occurred in six more genes out of the subset of 25, which could mean that there was lateral gene transfer occurring between the *Ascetosporean* species and the *Microsporidia* seen in the group.

4.3 Possible LGT (Lateral Gene Transfer)

I observed in six SGTs that the *Ascetosporeans* were grouping together with two *Microsporidia*, namely *E. aedis* and *N. bombycis*. These *Microsporidians* infect invertebrates and have similar lifestyles so this could mean that there was lateral gene transfer somewhere in time between these species. The six genes further studied were: *EIF2A* (Eukaryotic translation Initiation Factor 2A), *eif6* (Eukaryotic translation Initiation Factor 6), *pno1* (Partner of NOB1 homolog), *rpl7a* (Ribosomal Protein L7a), *rplp0* (Ribosomal Protein Lateral Stalk Subunit P0), and *rps23* (Ribosomal Protein S23).

The pipeline described in 3.6 (Possible Lateral Gene Transfer) was used in conjunction with earlier steps with MAFFT and IQ-TREE, except this time ~20 sequences of *Microsporidia* were added using blast. Trees were remade using the same data with the blast hits and are available to view (See Appendix C)

On studying the SGTs, the various *Ascetosporean* taxa and their support on ancestral nodes were noted, and it was seen where they were branching. Based on these observations, there were multiple possible inferences made (Table 3).



Rhizaria_N_A_N_A_N_A_N_A_N_A_N_A_Lapot
M6MM-FB5_002091-T1
Rhizaria_Phytorhiza_N_A_N_A_Plasmodiophoridae_Plasmodiophora_brassicae
Rhizaria_Filosa_N_A_N_A_Guttulinaceae_Guttulinopsis_vulgaris
Opisthokonta-Ichthyosporea_N_A_Ichthyosporea_N_A_N_A_Abeoforma_whisleri
Opisthokonta-Ichthyosporea_N_A_Ichthyosporea_N_A_N_A_Pirum_gemmata
Rhizaria_Filosa_N_A_Euglyphida_Euglyphidae_Euglypha_rotunda
Rhizaria_Filosa_N_A_Cercomonadida_Cercomonadidae_Massisteria_sp
Rhizaria_Phytorhiza_N_A_Vampyrellida_Leptophryidae_Leptophrys_vorax
Rhizaria_N_A_N_A_N_A_N_A_N_A_N_A_D1
Rhizaria_Filosa_Chlorarachniophyceae_N_A_N_A_Bigelowiella_natans
Rhizaria_Filosa_Chlorarachniophyceae_N_A_N_A_Bigelowiella_longifila
Rhizaria_Filosa_Chlorarachniophyceae_N_A_N_A_Lotharella_globosa
Rhizaria_Filosa_Chlorarachniophyceae_N_A_N_A_Lotharella_globosa
Rhizaria_Filosa_Chlorarachniophyceae_N_A_N_A_Amorphochlora_amoebiformis
Rhizaria_Filosa_Chlorarachniophyceae_N_A_N_A_Gymnochlora_sp
Rhizaria_Filosa_Chlorarachniophyceae_N_A_N_A_Partenskyella_glossopodia
Rhizaria_Filosa_Chlorarachniophyceae_N_A_N_A_Chlorarachnion_reptans
Rhizaria_Filosa_Chlorarachniophyceae_N_A_N_A_Norrisiella_sphaerica
Rhizaria_Retaria-Foraminifera_N_A_Rotaliida_Bolivinitidae_Brizalina_sp
Rhizaria_Retaria-Foraminifera_N_A_Rotaliida_Buliminidae_Globobulimina_turgida
Rhizaria_Retaria-Foraminifera_N_A_Rotaliida_Rotaliidae_Ammonia_sp
Rhizaria_Retaria-Foraminifera_N_A_Rotaliida_Elphidiidae_Elphidium_margaritaceum
Rhizaria_Retaria-Foraminifera_N_A_Miliolida_Soritidae_Sorites_sp
Rhizaria_Retaria-Foraminifera_N_A_Rotaliida_Nonionidae_Nonionellina_sp
Rhizaria_Retaria-Foraminifera_N_A_Rotaliida_Rosalinidae_Rosalina_sp
Rhizaria_Retaria-Foraminifera_N_A_Rotaliida_Buliminidae_Bulimina_marginata
Rhizaria_Retaria-Foraminifera_N_A_Rotaliida_Buliminidae_Bulimina_marginata
Rhizaria_Retaria-Foraminifera_N_A_N_A_Reticulomyxidae_Reticulomyxa_filosa
Rhizaria_Retaria-Radiolaria_Acantharea_Symphyacanthida_Astrolithidae_Astrolonche_serrata
Rhizaria_Apofilosa_N_A_N_A_Gromiidae_Gromia_sphaerica
Opisthokonta-Fungi_Microsporidia_Hamiltosporidium_tvaerminnensis
Opisthokonta-Fungi_Microsporidia_Thelohania_contejeani
Opisthokonta-Fungi_Microsporidia_Nylanderia_fulva
Opisthokonta-Fungi_Microsporidia_Nylanderia_fulva
Opisthokonta-Fungi_Microsporidia_Encephalitozoon_hellem
Opisthokonta-Fungi_Microsporidia_Encephalitozoon_hellem
Opisthokonta-Fungi_Microsporidia_Encephalitozoon_romaleae
Opisthokonta-Fungi_Microsporidia_Ordospora_colligata
Opisthokonta-Fungi_Microsporidia_Microsporidium_sp
Opisthokonta-Fungi_Microsporidia_N_A_N_A_Nosematidae_Nosema_bombycis
Opisthokonta-Fungi_Microsporidia_Papilio_xuthus_XP
Rhizaria_Ascetosporea_N_A_N_A_Mikrocytiidae_Mikrocytos_mackini
Mikrocytos_mackini
Paramikrocytos_canceri_2014
Paramikrocytos_canceri_2018
Opisthokonta-Fungi_Microsporidia_Dictyocoela_muelleri
Opisthokonta-Fungi_Microsporidia_Dictyocoela_roeselum
Opisthokonta-Fungi_Microsporidia_Spraguea_lophii
Opisthokonta-Fungi_Microsporidia_N_A_N_A_Culicosporidae_Edhazardia_aedis_Edha_aedis
Opisthokonta-Fungi_Microsporidia_Pseudoloma_neurophilia
Opisthokonta-Fungi_Microsporidia_Trachipleistophora_hominis
94.4/95
95.2/99
96.2/97
98.1/100
91.2/92
58.1/89
99.4/100
100/100
100/100
29.6/99
44.7/69

Figure 7: The *eif6* gene shows possible Lateral Gene Transfer flowing from *Rhizaria* (pink) into *Microsporidia* (brown) and the species of interest (purple). Branch strengths and bootstrap support can be seen with the number on the branches. Some branch support values have been removed for greater legibility.

We could see in the trees that the species of interest (in purple) were grouping within the *Microsporidians* (brown) with that clade being sister to the main *Rhizarian* clade (pink). Most branches also had good support scores with UFBoot >97 (Fig. 7).

These newly added *Microsporidian* species grouped together with our species of interest in most of the six genes chosen for further testing. Some of the species of interest were removed in the filtering steps due to their poor match lengths, bitscore and e-value.

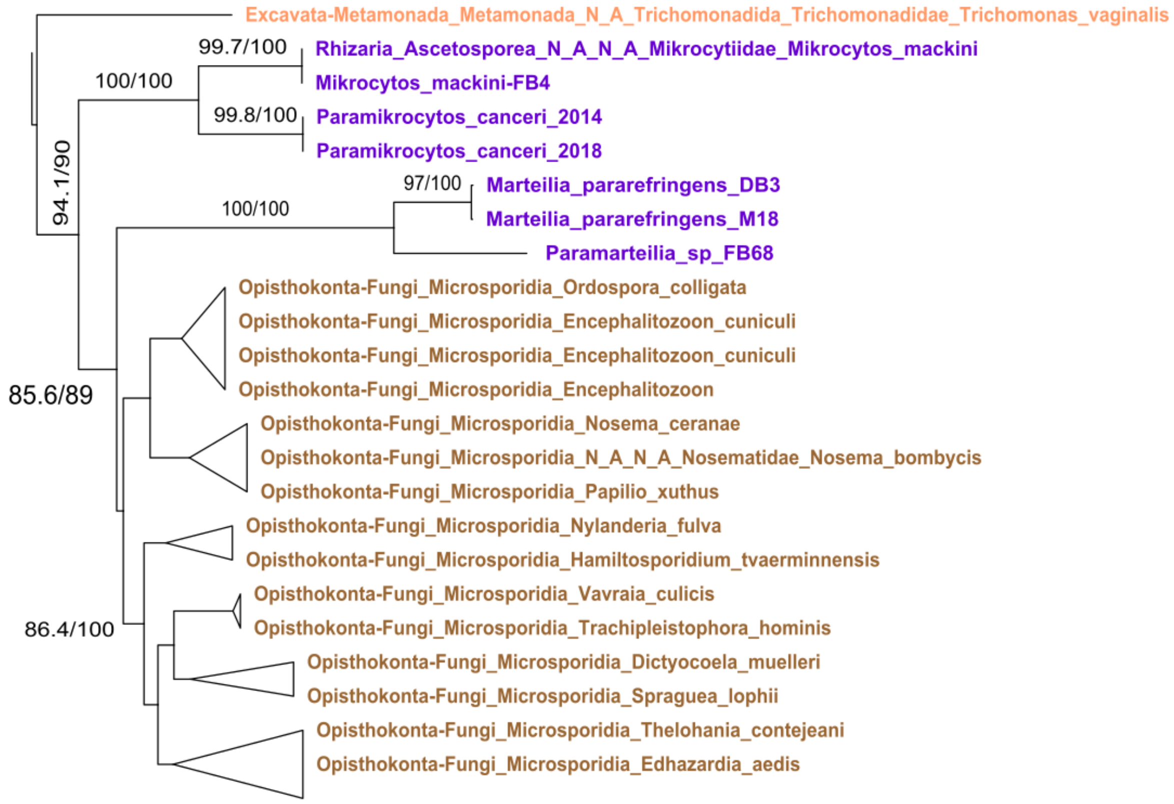

Excavata-Metamonada_Metamonada_N_A_Trichomonadida_Trichomonadidae_Trichomonas_vaginalis
99.7/100
Rhizaria_Ascetosporea_N_A_N_A_Mikrocytiidae_Mikrocytos_mackini
100/100
Mikrocytos_mackini-FB4
99.8/100
Paramikrocytos_canceri_2014
Paramikrocytos_canceri_2018
94.1/90
97/100
Marteilia_pararefringens_DB3
100/100
Marteilia_pararefringens_M18
Paramarteilia_sp_FB68
Opisthokonta-Fungi_Microsporidia_Ordospora_colligata
Opisthokonta-Fungi_Microsporidia_Encephalitozoon_cuniculi
Opisthokonta-Fungi_Microsporidia_Encephalitozoon_cuniculi
85.6/89
Opisthokonta-Fungi_Microsporidia_Encephalitozoon
Opisthokonta-Fungi_Microsporidia_Nosema_ceranae
Opisthokonta-Fungi_Microsporidia_N_A_N_A_Nosematidae_Nosema_bombycis
Opisthokonta-Fungi_Microsporidia_Papilio_xuthus
Opisthokonta-Fungi_Microsporidia_Nylanderia_fulva
Opisthokonta-Fungi_Microsporidia_Hamiltosporidium_tvaerminnensis
Opisthokonta-Fungi_Microsporidia_Vavraia_culicis
86.4/100
Opisthokonta-Fungi_Microsporidia_Trachipleistophora_hominis
Opisthokonta-Fungi_Microsporidia_Dictyocoela_muelleri
Opisthokonta-Fungi_Microsporidia_Spraguea_lophii
Opisthokonta-Fungi_Microsporidia_Thelohania_contejeani
Opisthokonta-Fungi_Microsporidia_Edhazardia_aedis

Figure 8: ***rplp0*** **gene tree showing possible LGT between species of interest (purple) and** ***Microsporidia*** **(brown).). Branch strengths and bootstrap support can be seen with the number on the branches. Some branch support values have been removed for greater legibility.**

The best examples of possible LGT between species were the genes *eif6* and *rplp0*. These trees showed some evidence of gene transfer between *Microsporidia* (brown) and *Ascetosporea (purple)*. The *Excavata-Metamonada* (orange) taxon was treated as an artifact of Long Branch Attraction due to the long branch lengths of the *Microsporidia* and *Ascetosporeans*. This was further reinforced by the *Metamonada* grouping outside the clade of relevance (Kück *et al.* 2017) (Fig. 8).

Gene name	**No. of Ascetosporean taxa present**	**No. of distinct clades formed**	**Possible inferences**
EIF2A	*14*	2	1. Lateral Gene Transfer: Yes 2. Sample contamination: Yes
eif6	*4*	1	Lateral Gene Transfer: Yes (with possible gene flow from *Rhizaria* towards *Ascetosporea* and *Microsporidia.* Good support.) (Fig. 7)
pno1	*1*	1	Lateral Gene Transfer: Possibly. No other *Ascetosporeans* to make evolutionary inferences.
rpl7a	*12*	3	1. Lateral Gene Transfer: Possibly 2. Sample contamination: Yes
rplp0	*13*	2	1. Lateral Gene Transfer: Yes (between *Microsporidia* and *Ascetosporea*. Good support.) (Fig. 8) 2. Sample contamination: Yes
rps23	*1*	1	Lateral Gene Transfer: Possibly. No other *Ascetosporeans* to make evolutionary inferences.

Table 3: Summary of SGTs of genes with possible Lateral Gene Transfer

4.4 Concatenated supermatrix tree

As mentioned in 3.6 (Concatenated trees), the full dataset and the subset of SAR taxa was concatenated separately, and trees were attempted to be made through this analysis. Due to the size of these files, this analysis was automatically terminated multiple times by UPPMAX as it exceeded memory capacities. Due to these technical challenges, it has not been added to this report.

5 Discussion

5.1 Initial Single Gene Trees

On observing the ~320 single gene trees closely, rooted to *Prokaryota* or *Excavata-Discoba*, it was seen there some taxa were placed in their expected positions while some taxa deviated from the expected position (possible contamination). There were instances of *Ascetosporea* grouping with *Cryptist*-nucleomorphs which are nuclei present in eukaryotes which evolve at a very high rate (Irwin & Keeling 2019). These nucleomorphs are fast-evolving and this grouping could be occurring due to Long Branch Attraction (LBA). Long Branch Attraction is a phenomenon that occurs when taxa with long branches are attracted to each other. This can interfere with phylogenetic signals and mask the true position of taxa (Bergsten 2005, Philippe *et al.* 2005). Faster-evolving organisms are seen to have longer branch lengths on phylogenetic trees. As *Ascetosporea* have been seen to be fast evolving, they tend to get attracted to these *Cryptist*-nucleomorphs. This signal on the tree is probably an artifact of LBA, but single gene trees do not provide much phylogenetic information, which is why phylogenomic methods are preferred to create species-trees instead of gene-trees.

Many species of interest had multiple hits per gene when performing blast searches against annotated genomes. These appeared to be very small sequences that matched exactly. This was speculated to be incorrect gene models or short contigs causing multiple hits to appear. Due to the multiple blast hits of varying lengths belonging to the same species, there were seemingly multiple “copies” of the same gene in the same species, sometimes even 15-20 copies of that gene present in the annotation. Some of these were seen to be prokaryote contamination through blast and the rest were theorised to be poor gene models, non-homologs or short contigs. It could have been a case of one gene being split into two during cleaning and assembly; so, it would annotate as two separate genes and in turn two separate blast hits.

5.2 Subset of genes

The gene subset consisted of 25 single copy orthologs found through Orthofinder. As they were single copy genes, the reasoning behind this was that there should be only one copy of a gene present per species. At first glance at the SGTs, a lot of the multiple gene copies/gene fragmentations were removed, due to the strict blast and alignment filtering that was not done previously. There was still prokaryote contamination, but the filtering removed most of this. It was also seen that the *M. mackini* and *P. canceri* were almost always present as sister groups with good support (Fig. 2, 5, 6 and 7) as also shown by (Hartikainen *et al.* 2014) but were never consistent in their position, which was theorized to be within the *Rhizaria* group of the TSAR supergroup. *M. pararefringens* and *Paramarteilia* sp. were the other two species that grouped together in almost all the other trees (Fig. 2, 6 and 8). The support for most of these clades was considered good. (See Appendix C)

5.3 Lateral Gene Transfer

When viewing the SGTs, it was observed that the species of interest were grouping together with the *Microsporidia E. aedis* and *N. bombycis* in some of the trees. On adding genes from the additional *Microsporidian* blast hits, the *Ascetosporean* species still grouped together with the old and newly added Microsporidia. The initial hypothesis was that the samples were contaminated with *Microsporidia*, but this was dismissed as all the samples for *Ascetosporea* were collected in different locations and at separate times. Another reason for this dismissal was that the samples were also sequenced in multiple sequencing runs and at separate times. This variance in sample locations, time of collection and sequencing runs made it highly unlikely that the same *Microsporidians* had contaminated multiple different samples. As the same grouping between *Ascetosporea* and *Microsporidia* was seen multiple times, this was hypothesized to be a case of Lateral Gene Transfer (LGT). The *Microsporidia* have a similar lifestyle as *Ascetosporea*, in that they are both eukaryotes which infect invertebrates and are intracellular (Han & Weiss 2017). Depending on how the grouping occurred, the gene flow could be analysed between the species by tracking the common ancestors and how they seemed to evolve. Another case of LGT between *Ascetosporea* and *Microsporidia* was also seen in a phylogenetic analysis of the transposable elements of both. This research into transposable elements was conducted by my co-supervisor Ioana Onut Brännström of the Burki lab.

5.4 Pre-processing of data

The pre-processing of the sequenced samples involved significant amounts of computation. During these processes, there are many different variables which could have caused the apparent issues in cleaning, genome assembly, annotation, and analysis we encountered. Cleaning normally involves taking the genomes of the hosts the *Ascetosporeans* were found in and trying to map the sequencing reads obtained to those genomes to remove the host from the data. In some cases, there may not be a publicly available host genome; which can increase the complexity of cleaning sequencing reads. In those instances, the final genomes of the *Ascetosporeans* could not be assembled accurately. This greatly reduced the amount of data present for the study as a lot of sequenced data could not be cleaned properly. Overall, due to the complexity of extracting genetic material from host tissues, this resulted in either fragmented or incomplete gene models when the genomes were assembled. As we could see in the SGTs, there were often multiple copies of genes with differing annotations. This could be due to the sequencing techniques used as well as types of studies in which they were involved. As these were extraneous factors, the amount of data for an already little studied group was reduced even further. However, an advantage of sequencing new samples, running analyses and conducting this study is that there is now more data for *Ascetosporea* than before, which could be very useful in future studies of this supergroup.

As the dataset was being compiled and made, the sequences were studied, and they were mostly all the same length for all taxa per gene. This would help in alignment as most alignment algorithms rely on matches and mismatches to score alignments. Gaps in alignments are normally given higher penalty scores as they normally don't help with most evolutionary inferences, but there is debate about whether to use gappy or trimmed alignments for phylogenetic analyses (Portik & Wiens 2021). As per (Saurabh *et al.* 2012), gaps can be introduced in different places depending on the software used. Gaps are also hypothetical insertion/deletion events in a multiple sequence alignment. They do not always contain important phylogenetic or evolutionary information. However, it is also theorised that gaps are phylogenetically informative and can be used for phylogenetics provided that these gaps are filtered and noise is removed (Donath & Stadler 2018). Multiple residue gaps in protein sequences can also be an indicator of monophyly (Lloyd & Calder 1991). Gap removed alignments were used due to personal preference, but future work can change this aspect of the study. Before the SGT's were created, the alignments were looked at after gap removal and it was observed that the species of interest were very gappy. The substantial amounts of gaps can relay different phylogenetic signals and according to (Roure *et al.* 2013), it can be better to include short sequences in phylogenetic analysis. This can be a further subject of study in the case of *Ascetosporea* due to the number of short hits and sequences present in the data.

5.5 Future research

For future prospects with the results of this study, the hypothesized LGT between *Ascetosporea* and *Microsporidia* should be looked at further as we see the same pattern in six genes out of a subset of 25. It stands to reason that this relationship exists not only at a gene level, but also on a genome wide level. This could be done by performing similar single gene tree analyses across the ~320 genes and then creating a large supermatrix tree with the new data. If this LGT holds across a majority of the ~320 genes, we can say that there are some evolutionary links between the *Microsporidia* and *Ascetosporea*. In this case, the LGT for the subset of data is also a significant result as it showed us something that wasn't studied before.

Another future possibility is to create a phylogenomic study of the smaller dataset. This is done as single gene trees generally do not have much phylogenetic signals, but phylogenomic studies do; this was initially executed but due to time constraints was not able to be completed in time. The initial single gene and incomplete supermatrix analysis was conducted using a simple LG+G4 model. The next steps would be to use more complex models like general time reversible (GTR) models, or more complex LG ± F ± Γ models with varying parameters to create a phylogenomic tree with the current dataset and run Bayesian phylogeny methods in order to obtain more accurate results.

The last step would be to thoroughly examine the dataset and attempt to identify potential contaminants in the newly sequenced data. As could be seen throughout the analysis and the SGTs, there were a lot of contaminants and incorrect phylogenetic placements of certain groups. This can be remedied by examining every single gene from the ~320 for possible prokaryotes and other contaminants that may have resulted from the types of samples collected. This will have to be done by blast and other related methods and will be time-consuming. Once the contamination has been removed, it will be easier for future research to be conducted using this data.

Aquaculture is very important for developing countries in ensuring their food security. Rising sea temperatures and other environmental changes are affecting the sustainability of seafood and aquaculture. Studies have shown that a pescatarian diet is more sustainable for the environment as a major protein source. However, rising temperatures have affected aquaculture populations by exposing them to different pathogens like *Ascetosporea* and other antibiotic resistant bacteria (Bass *et al.* 2019, Reverter *et al.* 2020, Fisher *et al.* 2021). Due to these issues, there is a danger to many populations in terms of mass mortalities, and this will in turn affect humanity by rendering this major source of protein inedible. As (Stentiford *et al.* 2012) have shown, nearly 40% of shrimp production is already affected by viral pathogens. This, along with *Ascetosporea* infecting other marine populations can cause a collapse of the global seafood trade and in turn severely affect developing countries. Understanding how *Ascetosporea* affect aquaculture populations could help in preserving these populations.

6 Acknowledgment

I would like to express gratitude to my supervisors **Fabien Burki** and **Ioana Onut Brännström** for their help and support during the thesis. I would not have learned so much and made so much progress without their constant support and input during the project. I would also like to thank **Markus Hiltunen** of the Burki Lab for taking me through the data pre-processing steps, answering my questions and helping me understand the finer details and challenges involved in my project.

I would also like to thank my Subject Reader, **Lisa Klasson** for asking me questions about my project during our meetings and giving me practical solutions to problems I was facing. These were instrumental to me in refocusing my ideas to move forward with the project.

I would be remiss if I do not mention my fellow master's students in the Department of Organismal Biology**: Ines Borges**, **Thomas Huber**, **Eleanor Sheridan** and **Boel Olsson** for their company during lunch, the SysBio seminars and fikas. I am also grateful to the other members of Burki Lab: **Mahwash Jamy**, **Miguel Mendez Sandin** and **Anne Walraven** for their help and guidance for future career prospects.

I would like to thank my fellow master's in bioinformatics students: **Vivianne**, **Melih**, **Greta**, **Milena** and **Patrick** for the random meetups and Dungeons and Dragons campaign; my corridor mates **Leon Libbey**, **Lucas Bruylands** and **Ting-An Lin** for distracting me when I was stressed with spontaneous cooking sessions and movie nights.

I would also like to thank **Anders Alfjorden** of the Burki Lab and National Veterinary Institute, Sweden for his help with understanding the sampling techniques and for his microscopy resulting in the picture on the cover page.

Lastly but most importantly, I would like to thank my mother **Chandana Bhawe**, my father **Kunal Bhawe**, my brother **Aniket**, and my partner **Akanksha**, as I would not have learned and achieved as much as I have without their support, trust, and motivation.

7 References

Abbott C, Meyer G. 2014. Review of Mikrocytos microcell parasites at the dawn of a new age of scientific discovery. Diseases of Aquatic Organisms 110: 25–32.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. Journal of Molecular Biology 215: 403–410.

Bass D, Ward GM, Burki F. 2019. Ascetosporea. Current Biology 29: R7–R8.

Bergsten J. 2005. A review of long-branch attraction. Cladistics: The International Journal of the Willi Hennig Society 21: 163–193.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114–2120.

Bondad-Reantaso MG, Subasinghe RP, Josupeit H, Cai J, Zhou X. 2012. The role of crustacean fisheries and aquaculture in global food security: Past, present and future. Journal of Invertebrate Pathology 110: 158–165.

Burki F, Corradi N, Sierra R, Pawlowski J, Meyer GR, Abbott CL, Keeling PJ. 2013. Phylogenomics of the intracellular parasite Mikrocytos mackini reveals evidence for a mitosome in rhizaria. Current biology: CB 23: 1541–1547.

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25: 1972–1973.

Carnegie RB, Meyer GR, Blackbourn J, Cochennec-Laureau N, Berthe FCJ, Bower SM. 2003. Molecular detection of the oyster parasite Mikrocytos mackini, and a preliminary phylogenetic analysis. Diseases of Aquatic Organisms 54: 219–227.

Donath A, Stadler PF. 2018. Split-inducing indels in phylogenomic analysis. Algorithms for molecular biology: AMB 13: 12.

Dufayard J-F, Duret L, Penel S, Gouy M, Rechenmann F, Perrière G. 2005. Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. Bioinformatics 21: 2596–2603.

Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biology 16: 157.

Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biology 20: 238.

Fisher MC, Moore SK, Jardine SL, Watson JR, Samhouri JF. 2021. Climate shock effects and mediation in fisheries. Proceedings of the National Academy of Sciences of the United States of America 118: e2014379117.

Gao K, Miller J. 2020. Primary orthologs from local sequence context. BMC bioinformatics 21: 48.

Graph J. 2017. drawio. WWW document 2017: https://github.com/jgraph/drawio-desktop/releases/tag/v18.1.3. Accessed 1 June 2022.

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Systematic Biology 59: 307–321.

Hamann L, Blanke A. 2022. Suspension feeders: diversity, principles of particle separation and biomimetic potential. Journal of the Royal Society, Interface 19: 20210741.

Han B, Weiss LM. 2017. Microsporidia: Obligate Intracellular Pathogens Within the Fungal Kingdom. Microbiology Spectrum, doi 10.1128/microbiolspec.FUNK-0018-2016.

Hartikainen H, Stentiford GD, Bateman KS, Berney C, Feist SW, Longshaw M, Okamura B, Stone D, Ward G, Wood C, Bass D. 2014. Mikrocytids are a broadly distributed and divergent radiation of parasites in aquatic invertebrates. Current biology: CB 24: 807–812.

Hellmuth M, Wieseke N, Lechner M, Lenhof H-P, Middendorf M, Stadler PF. 2015. Phylogenomics with paralogs. Proceedings of the National Academy of Sciences of the United States of America 112: 2058–2063.

Hine PM, Bower SM, Meyer GR, Cochennec-Laureau N, Berthe FC. 2001. Ultrastructure of Mikrocytos mackini, the cause of Denman Island disease in oysters Crassostrea spp. and Ostrea spp. in British Columbia, Canada. Diseases of Aquatic Organisms 45: 215–227.

Inkscape Project. 2020. Inkscape Project.

Irwin NAT, Keeling PJ. 2019. Extensive Reduction of the Nuclear Pore Complex in Nucleomorphs. Genome Biology and Evolution 11: 678–687.

Jensen RA. 2001. Orthologs and paralogs - we need to get it right. Genome Biology 2: interactions1002.1.

Kainer D, Lanfear R. 2015. The Effects of Partitioning on Phylogenetic Inference. Molecular Biology and Evolution 32: 1611–1627.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nature Methods 14: 587–589.

Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution 30: 772–780.

Kerr R, Ward GM, Stentiford GD, Alfjorden A, Mortensen S, Bignell JP, Feist SW, Villalba A, Carballal MJ, Cao A, Arzul I, Ryder D, Bass D. 2018. *Marteilia refringens* and *Marteilia pararefringens* sp. nov. are distinct parasites of bivalves and have different European distributions. Parasitology 145: 1483–1492.

Kück P, Wilkinson M, Groß C, Foster PG, Wägele JW. 2017. Can quartet analyses combining maximum likelihood estimation and Hennigian logic overcome long branch attraction in phylogenomic sequence data? PloS One 12: e0183393.

Laetsch DR, Blaxter ML. 2017. BlobTools: Interrogation of genome assemblies. F1000Research 6: 1287.

Laetsch DR, Koutsovoulos G, Booth T, Stajich J, Kumar S. 2017. DRL/blobtools: BlobTools v1.0.1. doi 10.5281/zenodo.845347.

Larousse M, Galiana E. 2017. Microbial Partnerships of Pathogenic Oomycetes. PLoS Pathogens 13: e1006028.

Larsson A. 2014. AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics 30: 3276–3278.

Lattos A, Chaligiannis I, Papadopoulos D, Giantsis IA, Petridou EI, Vafeas G, Staikou A, Michaelidis B. 2021. How Safe to Eat Are Raw Bivalves? Host Pathogenic and Public Health Concern Microbes within Mussels, Oysters, and Clams in Greek Markets. Foods 10: 2793.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England) 25: 1754–1760.

Lloyd DG, Calder VL. 1991. Multi-residue gaps, a class of molecular characters with exceptional reliability for phylogenetic analyses. Journal of Evolutionary Biology 4: 9–21.

Madden T. 2003. The BLAST Sequence Analysis Tool. National Center for Biotechnology Information (US)

Maddison DR, Swofford DL, Maddison WP. 1997. Nexus: An Extensible File Format for Systematic Information. Systematic Biology 46: 590–621.

Milner DA. 2018. Malaria Pathogenesis. Cold Spring Harbor Perspectives in Medicine 8: a025569.

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Molecular Biology and Evolution 37: 1530–1534.

Nylander J. 2010. catfasta2phyml.

Olsen G. 1990. Olsen, Gary. "The" Newick's 8: 45" tree format standard." World-Wide-Web Reference. http://evolution. genetics. washington. edu/phylip/newick doc. html (1990).

Palmer J, Stajich J. 2019. nextgenusfs/funannotate: funannotate v1.5.3. doi 10.5281/zenodo.2604804.

Philippe H, Zhou Y, Brinkmann H, Rodrigue N, Delsuc F. 2005. Heterotachy and long-branch attraction in phylogenetics. BMC evolutionary biology 5: 50.

Portik DM, Wiens JJ. 2021. Do Alignment and Trimming Methods Matter for Phylogenomic (UCE) Analyses? Systematic Biology 70: 440–462.

Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. 2020. Using SPAdes De Novo Assembler. Current Protocols in Bioinformatics 70: e102.

Rambaut A. 2010. FigTree.

Reverter M, Sarter S, Caruso D, Avarre J-C, Combe M, Pepey E, Pouyaud L, Vega-Heredía S, de Verdal H, Gozlan RE. 2020. Aquaculture at the crossroads of global warming and antimicrobial resistance. Nature Communications 11: 1870.

Roure B, Baurain D, Philippe H. 2013. Impact of Missing Data on Phylogenies Inferred from Empirical Phylogenomic Data Sets. Molecular Biology and Evolution 30: 197–214.

Saurabh K, Holland BR, Gibb GC, Penny D. 2012. Gaps: An Elusive Source of Phylogenetic Information. Systematic Biology 61: 1075–1082.

Schön ME, Zlatogursky VV, Singh RP, Poirier C, Wilken S, Mathur V, Strassert JFH, Pinhassi J, Worden AZ, Keeling PJ, Ettema TJG, Wideman JG, Burki F. 2021. Single cell genomics reveals plastid-lacking Picozoa are close relatives of red algae. Nature Communications 12: 6651.

Shen W, Le S, Li Y, Hu F. 2016. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. PLOS ONE 11: e0163962.

Sierra R, Cañas-Duarte SJ, Burki F, Schwelm A, Fogelqvist J, Dixelius C, González-García LN, Gile GH, Slamovits CH, Klopp C, Restrepo S, Arzul I, Pawlowski J. 2016. Evolutionary Origins of Rhizarian Parasites. Molecular Biology and Evolution 33: 980–983.

Stentiford GD, Neil DM, Peeler EJ, Shields JD, Small HJ, Flegel TW, Vlak JM, Jones B, Morado F, Moss S, Lotz J, Bartholomay L, Behringer DC, Hauton C, Lightner DV. 2012. Disease will limit future food supply from the global crustacean fishery and aquaculture sectors. Journal of Invertebrate Pathology 110: 141–157.

Strassert JFH, Irisarri I, Williams TA, Burki F. 2021. A molecular timescale for eukaryote evolution with implications for the origin of red algal-derived plastids. Nature Communications 12: 1879.

Strassert JFH, Jamy M, Mylnikov AP, Tikhonenkov DV, Burki F. 2019. New Phylogenomic Analysis of the Enigmatic Phylum Telonemia Further Resolves the Eukaryote Tree of Life. Molecular Biology and Evolution 36: 757–765.

Summons RE, Walter MR. 1990. Molecular Fossils and Microfossils of Prokaryotes and Protists from Proterozoic Sediments. American Journal of Science 290A: 212–244.

Ward GM, Bennett M, Bateman K, Stentiford GD, Kerr R, Feist SW, Williams ST, Berney C, Bass D. 2016. A new phylogeny and environmental DNA insight into paramyxids: an increasingly important but enigmatic clade of protistan parasites of marine invertebrates. International Journal for Parasitology 46: 605–619.

Zakrzewski W, Dobrzyński M, Szymonowicz M, Rybak Z. 2019. Stem cells: past, present, and future. Stem Cell Research & Therapy 10: 68.

8 Appendix A

Sequencing data

Organism	Host	DNA Isolation	DNA Sequencing	RNA Sequencing	Genome Annotation	Genes predicted
M6MM	Host: free living heterotrophic amoeba	Phenol chloroform	Paired-end 150bp read-length, NovaSeq 6000	Paired-end 150bo read-length, MiSeq	Structural and functional	9694
Paramarteilia sp. (now called *Paramarteilia canceri*)	Host: Necor puber (Velvet crab)	Qiagen DNA blood and tissue	Paired-end 250bp read-length, NovaSeq 6000	Sample too low quality to be sequenced	N/A	4902
Bonamia ostrae	Host: Ostrea edulis (European flat oyster)	Qiagen DNA blood and tissue	Paired-end 250bp read-length, NovaSeq 6000	-	N/A	5253
Marteilia pararefringens (M18)	Host: Mytilus edulis (Blue mussel)	N/A	Paired-end 150bp read length, HiSeqX	-	Structural annotation	5657
Marteilia pararefringens (DB3)	Host: Mytilus sp. Infected host tissue.	N/A	MiSeq	-	Structural and functional	4737
Marteilia pararefringens (DB4)	Host: Mytilus sp. Infected host tissue.	N/A	MiSeq	-	Structural and functional	5298
Marteilia pararefringens (S151)	Host: Mytilus sp.	N/A	NovaSeq	-	Structural and functional	4943
Marteilia cochillia (DB6)	Host: Cerastoderma edule (Common cockle)	N/A	MiSeq	-	Structural and functional	4559
Marteilia octospora (DB5a)	Host: Solen sp. (Bivalves). Infected host tissue.	N/A	MiSeq	-	N/A	900
Paramikrocytos canceri (2014)	Host: Cancer pagurus (Edible crab)	N/A	Illumina MiSeq	-	Structural and functional	2409
Paramikrocytos canceri (2018)	Host: Cancer pagurus (Edible crab)	N/A	Illumina HiSeq 2000	Illumina Miseq PE300	Structural and functional	2430
Mikrocytos mackini	Host: Crassostrea gigas (Pacific oysters)	N/A	Illumina Miseq 600. Paired-end 200bp and 300bp lengths.	Illumina HiSeq 2000	Structural and functional	~5000

Table 4: Sequenced genomes and data related to sequencing techniques

9 Appendix B

atad1 SGT

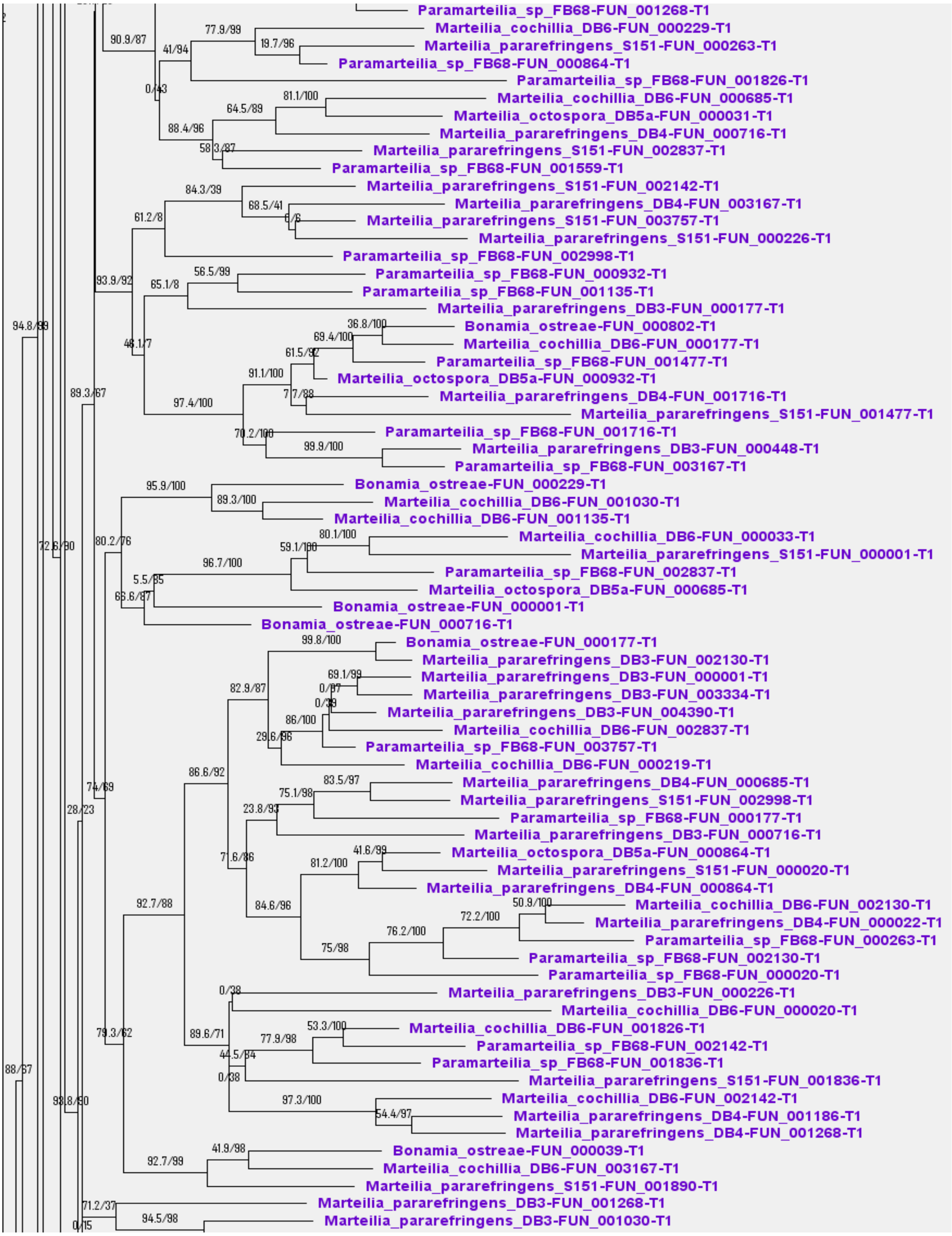

Paramarteilia_sp_FB68-FUN_001268-T1
77.9/99
Marteilia_cochillia_DB6-FUN_000229-T1
90.9/87
19.7/96
41/94
Marteilia_pararefringens_S151-FUN_000263-T1
Paramarteilia_sp_FB68-FUN_000864-T1
Paramarteilia_sp_FB68-FUN_001826-T1
0/43
81.1/100
Marteilia_cochillia_DB6-FUN_000685-T1
64.5/89
Marteilia_octospora_DB5a-FUN_000031-T1
88.4/96
Marteilia_pararefringens_DB4-FUN_000716-T1
58.3/87
Marteilia_pararefringens_S151-FUN_002837-T1
Paramarteilia_sp_FB68-FUN_001559-T1
84.3/39
Marteilia_pararefringens_S151-FUN_002142-T1
68.5/41
Marteilia_pararefringens_DB4-FUN_003167-T1
61.2/8
6/8
Marteilia_pararefringens_S151-FUN_003757-T1
Marteilia_pararefringens_S151-FUN_000226-T1
Paramarteilia_sp_FB68-FUN_002998-T1
56.5/99
93.9/92
Paramarteilia_sp_FB68-FUN_000932-T1
65.1/8
Paramarteilia_sp_FB68-FUN_001135-T1
Marteilia_pararefringens_DB3-FUN_000177-T1
36.8/100
94.8/99
Bonamia_ostreae-FUN_000802-T1
69.4/100
48.1/7
Marteilia_cochillia_DB6-FUN_000177-T1
61.5/92
Paramarteilia_sp_FB68-FUN_001477-T1
91.1/100
Marteilia_octospora_DB5a-FUN_000932-T1
89.3/67
7.7/88
97.4/100
Marteilia_pararefringens_DB4-FUN_001716-T1
Marteilia_pararefringens_S151-FUN_001477-T1
70.2/100
Paramarteilia_sp_FB68-FUN_001716-T1
99.9/100
Marteilia_pararefringens_DB3-FUN_000448-T1
Paramarteilia_sp_FB68-FUN_003167-T1
95.9/100
Bonamia_ostreae-FUN_000229-T1
89.3/100
Marteilia_cochillia_DB6-FUN_001030-T1
Marteilia_cochillia_DB6-FUN_001135-T1
80.1/100
72.6/90
80.2/76
Marteilia_cochillia_DB6-FUN_000033-T1
59.1/100
Marteilia_pararefringens_S151-FUN_000001-T1
96.7/100
Paramarteilia_sp_FB68-FUN_002837-T1
5.5/85
66.6/87
Marteilia_octospora_DB5a-FUN_000685-T1
Bonamia_ostreae-FUN_000001-T1
Bonamia_ostreae-FUN_000716-T1
99.8/100
Bonamia_ostreae-FUN_000177-T1
Marteilia_pararefringens_DB3-FUN_002130-T1
69.1/99
82.9/87
Marteilia_pararefringens_DB3-FUN_000001-T1
0/87
Marteilia_pararefringens_DB3-FUN_003334-T1
0/39
Marteilia_pararefringens_DB3-FUN_004390-T1
86/100
Marteilia_cochillia_DB6-FUN_002837-T1
29.6/96
Paramarteilia_sp_FB68-FUN_003757-T1
Marteilia_cochillia_DB6-FUN_000219-T1
86.6/92
83.5/97
Marteilia_pararefringens_DB4-FUN_000685-T1
74/69
75.1/98
28/23
Marteilia_pararefringens_S151-FUN_002998-T1
23.8/93
Paramarteilia_sp_FB68-FUN_000177-T1
Marteilia_pararefringens_DB3-FUN_000716-T1
41.6/99
71.6/86
Marteilia_octospora_DB5a-FUN_000864-T1
81.2/100
Marteilia_pararefringens_S151-FUN_000020-T1
Marteilia_pararefringens_DB4-FUN_000864-T1
92.7/88
84.6/96
50.9/100
Marteilia_cochillia_DB6-FUN_002130-T1
72.2/100
Marteilia_pararefringens_DB4-FUN_000022-T1
76.2/100
Paramarteilia_sp_FB68-FUN_000263-T1
75/98
Paramarteilia_sp_FB68-FUN_002130-T1
Paramarteilia_sp_FB68-FUN_000020-T1
0/38
Marteilia_pararefringens_DB3-FUN_000226-T1
Marteilia_cochillia_DB6-FUN_000020-T1
53.3/100
79.3/62
89.6/71
Marteilia_cochillia_DB6-FUN_001826-T1
77.9/98
Paramarteilia_sp_FB68-FUN_002142-T1
44.5/84
88/67
Paramarteilia_sp_FB68-FUN_001836-T1
0/38
Marteilia_pararefringens_S151-FUN_001836-T1
97.3/100
93.8/90
Marteilia_cochillia_DB6-FUN_002142-T1
54.4/97
Marteilia_pararefringens_DB4-FUN_001186-T1
Marteilia_pararefringens_DB4-FUN_001268-T1
41.9/98
Bonamia_ostreae-FUN_000039-T1
92.7/99
Marteilia_cochillia_DB6-FUN_003167-T1
Marteilia_pararefringens_S151-FUN_001890-T1
71.2/37
Marteilia_pararefringens_DB3-FUN_001268-T1
94.5/98
0/15
Marteilia_pararefringens_DB3-FUN_001030-T1

Figure 9: SGT for the gene *atad1*. Gene hits and multiple copies for the gene from the main dataset (purple). To prevent instances like this, the stricter filtering methods were chosen for the subset of data.

10 Appendix C

Data

Github for Scripts used: https://github.com/burki-lab/Phylogenomics_Data

The following Figshare links can be used to access the Single Gene Trees created:

Main dataset: https://figshare.com/s/658a912db7229a726ded

Subset of data: https://figshare.com/s/0cc3ee7ae98c816861f2

Lateral Gene Transfer trees: https://figshare.com/s/29f7ab5265ad5b044c6a

Legend for all the above trees: https://figshare.com/s/ccde4aabed5951302b4b

The genes used in the subset of the data are as follows:

- ATSAR2 (Arabidopsis Thaliana Secretion-Associated Ras Super family 2)
- EIF2A (Eukaryotic translation Initiation Factor 2A)
- eif6 (Eukaryotic translation Initiation Factor 6)
- fbl (Fibrillarin)
- FCF1 (*FCF1* RRNA-Processing Protein)
- gnb2L1 (Guanine Nucleotide-Binding protein subunit beta-2-Like 1)
- osgep (O-Sialoglycoprotein Endopeptidase)
- pno1 (Partner of NOB1 homolog)
- rpl13 (Ribosomal Protein L13)
- rpl17 (Ribosomal Protein L17)
- rpl18a (Ribosomal Protein L18a)
- rpl26 (Ribosomal Protein L26)
- RPL34 (Ribosomal Protein L34)
- rpl5 (Ribosomal Protein L5)
- rpl6 (Ribosomal Protein L6)
- rpl7a (Ribosomal Protein L7a)
- rplp0 (Ribosomal Protein Lateral Stalk Subunit P0)
- rps16 (Ribosomal Protein S16)
- rps18 (Ribosomal Protein S18)
- RPS19 (Ribosomal Protein S19)
- rps23 (Ribosomal Protein S23)
- rps26 (Ribosomal Protein S26)
- rps3a (Ribosomal Protein S3a)
- rps4y1(Ribosomal Protein S4 Y-Linked 1)
- srp54 (Signal Recognition Particle 54)