THE CELL BIOLOGY AND ECOLOGY OF HETEROTROPHIC EUKARYOTES IN A TAILINGS RECLAMATION SITE IN NORTHERN ALBERTA

By

Elisabeth Helen Richardson

A thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

In

Ecology

Department of Biological Sciences

University of Alberta

ABSTRACT

The wastewater products from the bitumen extraction process, called tailings, are contained in pits which cover over 77km2 of the Northern Albertan landscape. Reclaiming these tailings is an essential part of the life cycle of oil sands mines. One method under development converts tailings ponds into end-pit lakes (EPLs). EPLs contain a layer of tailings capped with freshwater, which, over time, are projected to resemble natural lakes in the Northern Alberta boreal forest region. The first full-scale EPL in the Athabasca Oil Sands is Base Mine Lake (BML) on the Syncrude mine site. I carried out an initial assessment of the eukaryotic microbial population of BML in 2015, two years after BML was water capped. Using eDNA sequencing, I was able to identify a substantial heterotrophic flagellate population in BML. I continued this eDNA study over four years (2015-2018) and observed changes the community over time. After addition of alum to BML in early 2016, photosynthetic eukaryotes returned over the subsequent summers. There was also a substantial increase in species richness in the heterotrophic eukaryotes, though there were no clear trends associated with any variables other than time.

One group that exhibited an increase in species richness over time was the ciliates. This phylum has been noted as a potential source of bioindicators, but little is known about the diversity of their cell biology. I ran a comparative genomic analysis on species across the diversity of the ciliates and identified substantial variation in the membrane trafficking components across each class. The results of this analysis suggest that the traditional model organisms for ciliates, all of which are found in the class Oligohymenophorea, may not be representative of ciliate diversity and additional model organisms or analysis may be necessary to determine resistance or sensitivity of specific ciliate classes or species to anthropogenic disturbance.

PREFACE

Chapter 1 is based upon the review by Richardson, E. & Dacks, J. B. Microbial Eukaryotes in Oil Sands Environments: Heterotrophs in the Spotlight. Microorganisms 7, 178 (2019). Though little of the original text is included, many of the themes found in this introductory chapter were first published here. This review outline was jointly written and developed by ER and JBD, the text was written by ER, and edited by ER and JBD.

Chapter 2 (excluding the preface and afterword) has been previously published as Richardson, E. et al. Phylogenetic Estimation of Community Composition and Novel Eukaryotic Lineages in Base Mine Lake: An Oil Sands Tailings Reclamation Site in Northern Alberta. Journal of Eukaryotic Microbiology 67, 86–99 (2020). Minor edits have been made to the original text for consistency and clarity. In this study, the studies were designed by JBD, ER, and DB. Sample collection was carried out by Syncrude, and eDNA extraction and sequencing was performed by PD and AS. The OTU clustering pipeline was designed by LP, who also contributed to data interpretation. ER carried out the eDNA analysis and ordination. ER also carried out the phylogenetic analysis in close collaboration with DB. ER wrote the manuscript, which was edited by all authors. Employees at Syncrude also provided input on the manuscript.

Chapter 3 (excluding the preface and afterword) is currently in preparation for publication. This manuscript is a collaboration with Dr. Peter Dunfield, Dr. Angela Smirnova, and Dr. David Bass, and Dr. Rolf Vinebrooke. In this study, the studies were designed by PD, JBD, RV and ER. Sample collection was carried out by Syncrude, and eDNA extraction and sequencing was performed by PD and AS. All analyses in the study were completed by ER with input from DB, JBD, and RV. Employees at Syncrude also provided input on the manuscript.

Chapter 4 (excluding the preface and afterword) contains data and figures from a previously published manuscript and data and figures which is currently in preparation for publication. Figure 4.8 was previously published as Figure 6 and Figure S4A of Sparvoli, D. et al. Remodeling the Specificity of an Endosomal CORVET Tether Underlies Formation of Regulated Secretory Vesicles in the Ciliate Tetrahymena thermophila. Current Biology 28, 697-710.e13 (2018). My

contribution to this paper was in collaboration with Dr. Sparvoli and Dr. Turkewitz, who carried out the molecular biology wet lab work associated with the project as indicated in Chapter 4. The remainder of the chapter is a manuscript currently in preparation with Dr. Joel B Dacks. In these sections, JBD and ER designed the experiments, ER performed the experiments. JBD assisted with data interpretation.

And she quoth, 'Now indeed I know thee, for in sooth art thou Parzival!
Didst thou see the mournful monarch? Didst thou see the wondrous Grail?
Ah! tell me the joyful tidings, may his woe at last be stilled?
Well is thee that the blessèd journey thou hast ta'en, now shall earth be filled,
As far as the winds of heaven may blow, with thy fair renown;
Naught on earth but shall do thee service, fulfilment each wish shall crown!'

Then he quoth, 'Nay, I asked no question!' 'Alas I' cried the mournful maid,
'That ever mine eyes have seen thee, who to question wast sore afraid!
Such marvels they there have shown thee, yet no word might they win from thee,
When thou sawest the Grail, and those maidens who serve It, from falsehood free,
Fair Garschiloie, and yet fairer Repanse de Schoie the queen.
Thou hast seen the knives of silver, thou the bleeding spear hast seen—
Alas! wherefore hast thou sought me? Dishonoured, accurst art thou
Who bearest wolf's fang empoisoned! And deep in thine heart I trow
Is it rooted, the plant of falsehood, and afresh doth it ever spring!
Thou shouldst have had pity on him, Anfortas, their host and king,
And have asked of his bitter sorrow, on whom God hath a wonder sped,
Now thou livest, and yet I tell thee to bliss art thou henceforth dead!'

"Parzival", Wolfram von Eschenbach

Translated by Jessie L. Weston

ACKNOWLEDGEMENTS

I began my PhD by moving to Canada from the UK, and I am extremely grateful both to the people I left behind and the new ones I met along the way during the course of this journey. I would first like to extend my thanks to everyone who welcomed me to Canada and helped me settle here - I've been fortunate to be so kindly received for nearly six wonderful years. I have also received funding from the provincial and federal governments of Alberta and Canada in the forms of an Alberta Innovates - Technology Futures and a Vanier scholarship. I hope I can repay this investment in my potential.

I would have never started my PhD, let alone at the University of Alberta, without Richard Dorrell. I know I am not the first person who has thanked Richard in a thesis acknowledgement, and I am equally certain I will not be the last. He is one of the most dedicated and inspirational scientists I have ever had the pleasure of knowing, and his passion for inspiring and supporting his students is something I hope I carry forward throughout the rest of my life. He was also who introduced me to my PhD supervisor, Joel Dacks. Joel has been an incredible supervisor and mentor since I booked my first flight to Edmonton, and has continually supported me, my scientific career, and my various extracurriculars throughout my PhD. He helped me design and re-design my PhD project through setbacks and even discipline crossovers, without ever losing his enthusiasm for my progress or my science. I would also like to thank my supervisory committee, Rolf Vinebrooke, Lisa Stein, and Camilla Nesbo for their thoughtful and insightful commentary on my project throughout the years, and, in particular, Rolf for taking on a co-supervisory role when it became clear that the ecology of Base Mine Lake would make a substantial part of my doctoral research. This transition would not have been possible without the support of the Biological Sciences faculty, staff, and students, and I'd like to thank everyone in the department for making me feel welcome and helping me complete my requirements in the programme. Lastly, I would not have become the scientist I am today without the Dacks Lab 2014-2020, which is an incredible group of scientists and even more incredible group of friends, ones who I am proud to call my colleagues. I'm sure if I listed everyone individually, I would forget someone, so I hope it suffices to say that I love you all.

Personally, I would like to thank everyone who supported me in getting to this point and sticking it out to the end, even when it hasn't been easy. My whole family, including my wonderful mum and dad, who have been supporting my academic pursuits for decades even when they cause me to move five thousand miles away from home, and my brother Sam, the original Dr. Richardson. I'd also like to thank my grandmother Margaery Jellicoe in particular for hosting me whenever I worked in London.

When you're doing a PhD so far from home, it's even more important to have good friends around you, and I'm incredibly lucky to have some amazing ones. First of all, I'd like to thank my sisters in science Emily Herman, Laura Lee and Shweta Pipaliya for being with me through the trials and tribulations of grad school, and with whom I now have an emoji code for every possible academic situation. I'd like to thank everyone in Let's Talk Science who I worked with over my years as a co-ordinator, particularly my fellow co-ordinators; a group I joined to stave off some of the homesickness when I first came to Edmonton ended up being a huge part of my grad school life and the place where I met some of my best friends, including Anna Henderson, Vanessa Carias, Natasha Donahue, and Alison Muller. I'd also like to thank Sasha van der Klein both for her amazing friendship over the years and encouraging me to run for Vice-President Labour of the Graduate Students' Association at the University of Alberta, one of the most rewarding (and eye-opening!) experiences of my graduate student career, and Jennifer Tollmann for international support and being my European correspondent.

Finally, I would like to thank my partner Gregory Altrogge, who is the best man I have ever met.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND GLOSSARY OF TERMS

18S rRNA: the rRNA gene associated with the 18S (small) subunit of the eukaryotic ribosome.

Amplicons: Fragments of eDNA amplified by PCR using gene-specific primers.

AP: Adaptor Protein

ASV: Autosomal Sequence Variant

BCR: Beaver Creek Reservoir

BLAST: Basic Local Alignment Search Tool

BML: Base Mine Lake

CIPRES: CyberInfrastructure for Phylogenetic RESearch

COG: Conserved Oligomeric Golgi complex

COPI: COatomer Protein I

COPII: COatomer Protein II

CORVET: class C CORe Vacuole / Endosome Tethering

DNA: Deoxyribose Nucleic Acid

eDNA: environmental DNA

eRNA: environmental RNA

EPL: End-Pit Lake

EukBank / EukRef / EukMap: Eukaryotic (sequence) Bank, Eukaryotic (sequence) Reference, Eukaryotic (sequence) Map (Reference Database)

FFT: Fluid Fine Tailings

FISH: Fluorescent In-Situ Hybridisation

GenBank: Gene Bank (Reference Database)

GRASP: Golgi Reassembly And Stacking Protein

HF: Heterotrophic Flagellate

HMM: Hidden Markov Model

HOPS: HOmotypic fusion and vacuole Protein Sorting

HTAC: HeteroTetrameric Adaptor Complex

HTS: High-Throughput Sequencing

MAFFT: Multiple Alignment using Fast Fourier Transform

MAC / MIC : MACronucleus / MICronucleus

Metagenomics: the entire sequence complement from an eDNA sample extraction.

Metatranscriptomics: the entire sequence complement from an eRNA sample extraction.

MFT: Mature Fine Tailings

MIC: Maximal Information Coefficient

MLSB: Mildred Lake Settling Basin

MMETSP: Marine Microbial Eukaryote Transcriptome Sequencing Project

mothur: HTS analysis platform.

MTC: Multisubunit Tethering Complex

MTS: Membrane Trafficking System

MUSCLE: Multiple Sequence Comparison by Log Expectation

NMDS: Non-Multimetric Dimensional Scaling

OSPW: Oil Sands Process Water

OTU: Operational Taxonomic Unit

PAH: PolyAromatic Hydrocarbon

PCR: Polymerase Chain Reaction

PERMANOVA: PERmutational Multivariate ANalysis Of VAriance

PFLA: Phospholipid Fatty Acid

Porewater: the water that is released from FFT as it settles into MFT (equivalent to OSPW)

PR2: Protist Ribosomal Reference database (Reference Database)

QIIME2: Quantative Insights Into Microbial Ecology 2 (HTS analysis platform)

Reclamation: physical reconstruction of a disturbed landscape to equivalent land use capacity
Remediation: removal or containment of contaminants from an environment

SAG-D: Steam-Assisted Gravity Drainage

SILVA: This is not an acronym, from the Latin for forest. (Reference Database)

Swarm: OTU clustering algorithm

Tailings: the liquid waste from the bitumen extraction process.

TIME: Temporal Insights into Microbial Ecology

TRAPPI / II: TRAnsport Protein Particle complex

TSAR: Telonemids, Stramenopiles, Alveolates and Rhizaria

TSET: This is not an acronym, one of the heterotetrameric adaptin complexes

TWI: Tailings / Water Interface

USEARCH / VSEARCH: OTU clustering algorithm

V1-V9 regions: Variable regions of the 18S rRNA

WIP: West In-Pit

Chapter 1
PROTISTS, PONDS, AND PETROLEUM

1.1 Protist taxonomy and diversity: DNA as a classification tool

Efforts to catalogue the diversity of life on Earth have been ongoing for thousands of years. One of the earliest examples of grouping organisms by their traits, the precursor to modern taxonomy and systematics, is found in Aristotle's *History of Animals*, written circa 350 BCE[1]. However, it was not until the 18th century that formal attempts to classify organisms in a systematic manner began with the Linnaean classification system, a modified version of which is still in use today[2]. Our collective encyclopaedia of the organisms with which we share the planet has expanded exponentially since the early days of naming plants and animals, and this includes the vast reservoirs of microbiological life invisible to the human eye.

Carolus Linnaeus's taxonomy was developed as an observational system, and classification of organisms based on visible features remains a key aspect of taxonomy and systematics[3]. However, appearances can be deceiving, and this is never truer than when applying morphological classifications to microbes. Single-celled organisms, particularly eukaryotes, exhibit dizzying morphological variation (i.e. polymorphism). Further, morphological and ultrastructural analyses via microscopy have led to incredibly sophisticated organismal descriptions and identifications of diverse cellular structures[3]. However, there are limitations on the traits that can be distinguished via a microscope. For many single-celled heterotrophs, with their small size, limited morphological diversity, and similar life histories, there is simply not enough information to determine the true extent of their divergence[4].

With the discovery of the structure of DNA and the advent of DNA sequencing and analysis in the 20th century, however, microbiologists have developed a powerful new tool with which to complete their analyses[5]. The promise of such techniques were recognised soon after the structure of DNA was established in the early 1960s[6]. Though using DNA as a method of establishing taxonomy and species boundaries were first elucidated with respect to animals (particularly concerning human and primate evolution[6,7]), the same ideas were readily applicable to the microbial world. One of the earliest examples of using molecular phylogenetic techniques to

establish a molecular phylogeny across all domains of life was carried out by Schwartz and Dayhoff in 1979[8]. In this paper, the authors used ribosomal RNA, ferredoxins, and cytochrome c sequences to create composite trees describing the potential evolutionary origins of nuclear DNA, mitochondria, and chloroplasts. Even at this early stage, it was evident that to untangle taxonomic relationships deep within the tree of life such as the bacterial/archeal/eukaryotic divergence(s), it would be necessary to determine which candidate genes could be used most effectively across all known organisms.

Ribosomes and the associated ribosomal machinery are used to translate genetic information into protein. Often referred to as the 'central dogma' of biology, this process is one of the few constants that unites all living organisms—all the way back to the hypothesised origins of life[9]. Ribosomes are complex, multimolecular structures that contain both RNA and proteins[10]. The DNA involved in the construction of the RNA portions of the ribosome (described as the rRNA small and large subunit[11]) is a natural target for classification. They are conserved in some form in all eukaryotes and bacteria, they are not translated into protein and therefore are not subject to the selection pressure to maintain a coding sequence, and they consist of highly conserved and hypervariable regions[12]. In eukaryotes, the most commonly used rRNA gene for classification is the 18S small subunit[13]. Initially, the promise of using ribosomal DNA sequences to establish taxonomy (later developed into the technique of DNA barcoding) seemed limitless, and some suggested that it would come to replace morphologically-based taxonomies entirely[14,12,15–17]. Vast databases of 18S rRNA genes from known organisms have been compiled to act as reference databases for taxonomic reference and phylogenetic analysis[18–21], and there has been extensive research into whether 18S taxonomies can accurately reconstruct the diversity of samples[13,22–25]. Most of these studies have concluded that the taxonomic classification based on 18S assessments is only as good as the reference database; accordingly, much of the effort in taxonomic research is currently going toward curation of these databases[21]. From a phylogenetic perspective, there is also strong evidence that 18S more accurately resolves some taxa than others; for example, animal phylogenies are much better recovered using 18S rRNA than plant phylogenies[26].

To be clear, our understanding of DNA has by no means ended debates over taxonomy. Since molecular biology methods were established in the 1960s, there has been heated debate over

whether molecular techniques could—or should—supersede the established natural history of these species[27], and the tree of eukaryotes, particularly with respect to single-celled eukaryotes, is still under constant revision, with expanded sampling and new methods of classification often shaking its branches[28]. There has also been a substantial backlash against the concept of DNA-only taxonomy for microbial eukaryotes. Papers advocating a combined morphological and barcoding approach to assigning taxonomy often use extremely emotive language to defend morphological classification, citing the "perils" of DNA barcoding[29] or describing morphological taxonomy as "sadly considered old fashioned"[4]. As a recent advance in molecular barcoding technology, the presence of field sequencers that allow swift generation of DNA sequence *in situ* for organismal identification[30] was also met with resistance from other researchers, citing a lack of investment in museums and physical specimen collections[31]. Despite this, 18S phylogenies have provided many notable results, including the recent phylogenies of the heterotrophic nanoflagellate groups Glissomonada[32] and Microsporidia[33], two previously intractable clades known for their small size and difficult resolution. In fact, the DNA-based revised trees of the Microsporidia indicated that this clade, previously thought to be relatively small and entirely parasitic, has a much larger and more diverse membership including species which were previously classified within the sister group Rozellids[33]. The increased popularity of DNA-only classifications and phylogeny shows no sign of slowing down[34–37]. The current known diversity of eukaryotes, based on a combination of morphological and phylogenetic evidence, is indicated in Figure 1.1. Briefly, the majority of eukaryotic diversity is divided into two clades: Amorphea and Diaphoretickes. Within these groups are the more commonly recognised multicellular classifications: the Amorphea contains the Animalia and Fungi, while the Diaphoretickes contain the Archeaplastida. However, the span of eukaryotic diversity also includes numerous other single-celled protist and algal phyla, some of which are not contained within these two domains or do not yet have a definitive classification[28].

1.2 Microbial ecology and the age of high-throughput sequencing

Microbes, including bacteria, have been observed since the development of microscopy in the 17th century. However, early microscopes did not resolve bacterial structures well, and the vast

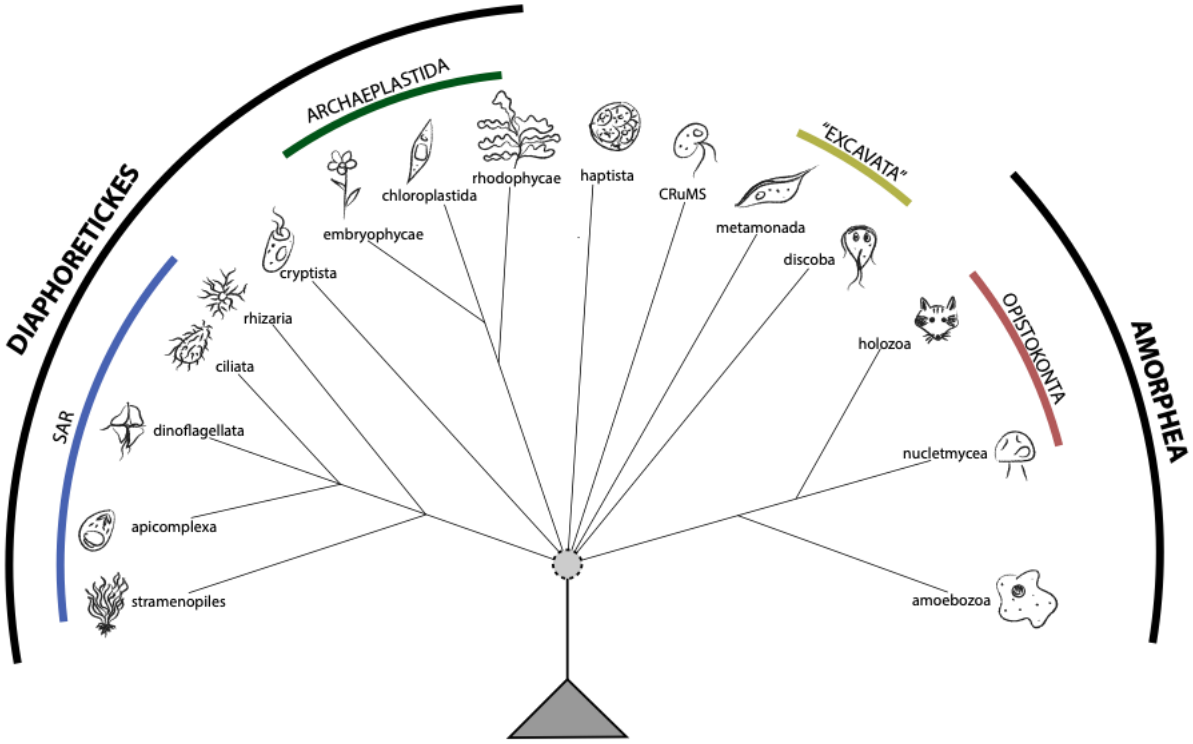


Figure 1.1 Diversity of eukaryotes
The current known diversity of known eukaryotes and the taxonomic relationships between them. Based on Burki and Keeling (2019) and Adl et al. (2019).

majority of what was known about bacterial function was based upon observing the effects of bacterial processes on the environment. One of the earliest examples of this was Pasteur's experiments on fermentation; by manipulating the bacteria present in the environment, he was able to observe the effects of bacteria on wine[38]. Much of the development of microbiology over the next century centered around disease prevention and culturing of microorganisms—most famously through the rival microbial laboratories of Pasteur and Koch[39].

However, it was the work of Sergei Winogradsky around the turn of the 20th century that established the effects of microbial processes on the environment. He was the first to observe biogeochemical processes from a microbial perspective and was able to establish experimentally that bacteria were essential for global nitrogen cycling[40]. Efforts to observe biogeochemical processes still use indirect methods of assessing microbiological contributions. Increasingly, however, specific processes have become associated with specific bacterial clades, and measuring the community composition of the bacteria has become a more relevant metric for understanding the nutrient cycling. For example, phospholipid fatty acid analysis (PFLA) is used to determine the distribution of idiosyncratic lipids found in various groups of bacteria and used to make inferences about the sample as a whole. These biodiversity assessments are an essential part of understanding global ecology and have become ever more time sensitive in the context of a globally changing climate[41].

In the era of high-throughput sequencing, DNA analysis can be used for ecology as well as taxonomic classification. Extraction of DNA directly from environmental samples for sequencing, known as eDNA sequencing, is an extremely popular technique for assessing biodiversity, and there are numerous studies dedicated to exploring the best way to achieve maximal DNA recovery from various environments[42–45]. With the combination of eDNA and high-throughput sequencing, the opportunity to understand of the diversity and abundance of eukaryotes has exploded[28]. Though the development of these new techniques has brought technological challenges[46], important information has been obtained from all manner of global environments that has forced microbiologists to re-examine some widely held preconceptions about microbial eukaryotic diversity and abundance.

Many of the techniques developed for assessing microbial communities and abundances were, and still are, developed with bacteria in mind. It is therefore important to note that studies of eukaryote-specific community composition face challenges which do not impact studies of the prokaryotic microbial communities of the same environments[13]. Eukaryotes can also have multiple protective membranes and a nucleus before the genome can be accessed, and may potentially possess shells, scales, or other cell protections that hinder cell lysis[28]. In the majority of environments bacteria outnumber eukaryotes, and because bacterial DNA is more abundant metagenomic samples usually contain only a small percentage of eukaryotic DNA[47]. Bacteria also have smaller genomes that are more likely to be fully covered by a metagenomic sample, which can also bias interpretation of species abundance and community structure[47].

Because of the additional difficulties associated with carrying out high-throughput environmental DNA studies on eukaryotes, information on eukaryotic microbial communities has lagged behind that of bacteria. Due to increased sequencing effort and extensive curation of eukaryote-specific gene databases, it is now feasible to analyse bacterial and eukaryotic communities within a sample simultaneously and retrieve broadly comparable results[48,49]. However, it is not accurate to assume that the environmental factors driving bacteria and protists are the same. An analysis of protists in soil by Bates et al. (2013)[50] sampled a global selection of soil and biome types. One of the most notable discoveries was that, unlike bacteria, the most important explanatory variable predicting community composition for protists was soil moisture content. Soils from the Dry Valleys in the Antarctic had similar protist communities to the Mojave Desert, while the humid Peruvian jungle resembled the Caribbean island of Puerto Rico[50].

However, one cannot assume abundance of taxa nor their roles in the microbial community based on the known properties of the environment. The same paper also identified a high percentage of Apicomplexa in the soil, a result mirrored in a study of soil transcriptomes by Geisen et al. (2015)[51]. This work focused on uncovering the metabolically active portion of the protist community, and discovered a massive and overlooked community of parasites, which appeared to be active and likely associated with hosts, not dormant in cyst form within the soil as was previously assumed[51]. Further studies have found that organisms that have long been known to science, but previously thought to be rare and relatively unimportant, can actually be highly

abundant and have simply been undetected. The Tara Oceans project[52] showed that diplonemids, a group of organisms within the Excavata that were considered to be rare, actually account for the majority of marine protist biomass[53]. They likely have a vital role in nutrient cycling in the open oceans[53].

18S rRNA are now a standard part of the eukaryotic microbial ecologist's toolkit with a relatively low barrier of entry compared to earlier bioinformatic methods. Platforms such as QIIME[54] and mothur[55] allow nonbioinformaticians to produce informative research on eDNA samples, and this has massively expanded the potential for evaluating the microbial ecology of understudied environments using an eDNA approach[34]. Figure 1.2 illustrates an example of an eDNA workflow and shows the stages necessary for completing an 18S study from raw input material to sequencing output. Many environments or geographic areas now have some environmental survey data available via GenBank or EukRef, and it is possible to search both of these databases for any environmental sequences that correspond to one's own sequences of interest[18,21]. This EukRef database has been successfully used for global biodiversity assessments and taxonomic classification of several heterotroph groups, including ciliates[56], kinetoplastids[57], and diplonemids[58]. The best practices for surveying the eukaryotic microbial community of an environment via its eDNA are still disputed[51,59]. As previously noted, metagenomic studies (where eDNA is sequenced directly and detection of species is limited by the quantity of that species' DNA in the sample) often retrieves very little eukaryotic DNA[47]. A common solution for this issue is production of amplicons, short DNA sequences amplified by 18S-specific primers[35,60]. Amplicon-driven studies using the 18S rRNA subunit are now effectively mapping protist diversity, abundance, and microbial ecology on a global scale[21,35]. Use of amplicons adds several stages to the diversity assessment pipeline, all of which have the potential to introduce biases into the analysis. There are several hypervariable regions in the 18S rRNA gene, and primers are available to target multiple regions effectively[23]. The most commonly used, the V4 region, is approximately 400nt long in most organisms and is by far the best represented in database curation efforts[21]. There is some dispute as to which hypervariable region most accurately represents the diversity of the studied community; while Tanabe et al. (2015) and Hu et al. (2015) found that the V4 region provided as much explanatory power as the entire 18S subunit[61,62], Wylezich et al. (2018) found that V4 sequencing did not provide accurate classifications of communities in

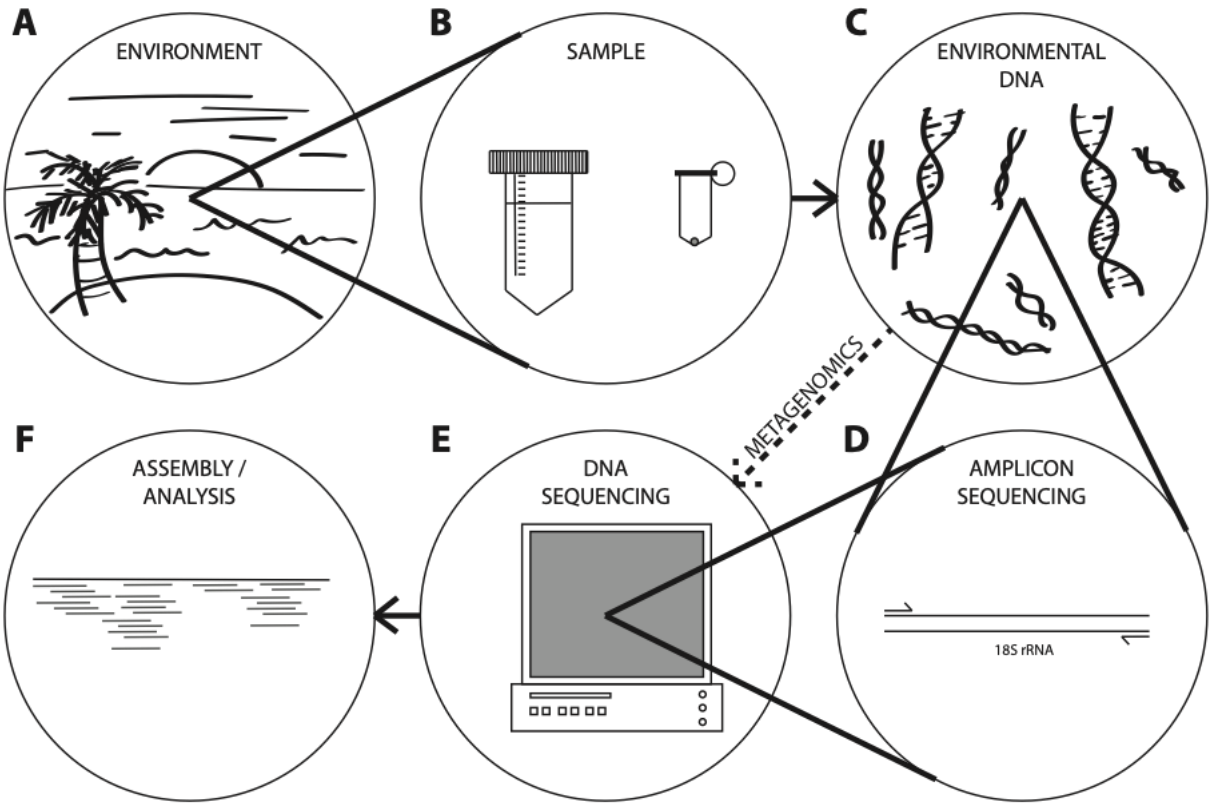


Figure 1.2: Overview of eDNA extraction and analysis protocol
A: A soil or water sample is collected from the environment of interest. **B:** The DNA is extracted from the sample, usually via commercially available kits. **C:** Environmental DNA can then be stored or immediately used for analysis. **D:** For amplification of the eukaryote-specific signal, amplicons can be created by using PCR on a gene of interest. **E:** DNA is sequenced. **F:** Downstream alignment and analysis of the sequences, including assembly of sequenced reads created in E.

oxygen-deficient environments[63]. Jamy et al. (2019) advocated using the entire 18S and 28S rRNA genes and found that using long-read sequencing substantially improves phylogenetic resolution[64]. This technique involves using an electrical current to assess the base pair sequence of a single strand of DNA or RNA passing through a pore and can be used to create single sequencing reads thousands of base pairs in length[65].

Following the selection of the amplified gene region(s), one of the most intense areas of disagreement is the clustering threshold for operational taxonomic units (OTUs). These species-level proxies for diversity, first proposed by Sneath and Sokal in the 1960s[66] and adopted early in the development of DNA barcoding methods for classification[17], involve clustering DNA sequences based on their overall similarity to each other compared to other sequences in the dataset. There are many available OTU clustering algorithms, though the two most commonly used are VSEARCH[41] and swarm2[67]. Importantly, the OTUs produced from a round of clustering are comparable only to each other as they have been defined in reference to each other; it is inaccurate to compare OTUs between two separately clustered datasets[41,67]. For bacteria, which generally have more mutable species boundaries due to their extensive horizontal gene transfer, a 97% clustering threshold (where reads with 3 base pair differences or fewer every 100 base pairs are clustered as a single out) is often used[68], though recent studies have also suggested this may be resulting in inaccurate classifications[69,70]. For eukaryotes, however, 97% threshold is far too broad to accurately depict species-level diversity, and an exact clustering threshold appropriate to reflect species-level diversity in protists is likely nonexistent; the corresponding percentage difference in what constitutes 'species-level diversity' varies between eukaryotic kingdoms[71]. There are two potential solutions to this problem: use a high clustering threshold (usually around 99%) and accept the constraint of uneven clustering or do not cluster at all[35]. The latter goal can be accomplished either by using a 'zero-radius OTU' (or 'ZOTU') or by using autonomic sequence variants (or 'ASVs'[72]). ASVs are created using a process of denoising rather than clustering, where potential sequencing errors or single-nucleotide polymorphisms are removed from a sequence without removing taxonomic resolution by clustering based on a percentage similarity. Because the sequences are not clustered, it is possible to analyse sequences independently of the overall dataset and therefore compare taxonomic classifications and diversity estimates obtained in different studies by different researchers— a notable advantage over OTU clustering methods[73].

Despite some differences of opinion, there is an increasing push towards standardisation of methodology for analysing protist communities; this has widespread support from researchers. Geisen et al. (2019)'s paper contains guidelines for these studies which outline the most common techniques currently in use in the field[35], and user-friendly pipelines and protocols are available more generally for analysing eDNA data[54,55]. For a detailed discussion of the methodological approach used to determine the protocols used for eDNA surveys in this thesis, please refer to Appendix I.

1.3 Studies on the ecological effect of the petroleum industry on heterotrophic protists.

One of the goals of global biodiversity assessment is to set a baseline to assess change in microbial populations over time. While the overarching effects of global warming and climate change are key areas of inquiry, humans are also impacting environments more directly and acutely through a variety of interventions, whether intentional or accidental.

The petroleum industry is worth trillions of dollars, and, despite global efforts to wean societies from dependence on fossil fuels, it still comprises an enormous part of global economy and trade[74]. Extraction of hydrocarbons from underground is traditionally understood as drilling and siphoning of liquid oil deposits – however, as conventional oil fields run low, efforts are increasingly focused on nonconventional sources such as shale oil, fracking, and oil sands[74,75]. These unconventional oil sources generally result in much more environmental disturbance than oil drilling, and much less is understood about the way they affect the local ecosystem[76]. This issue is exacerbated with respect to the microbial ecosystem, which is often poorly understood compared to the local macrofauna and flora[77]. Plant and animal movements have often been tracked for decades or centuries due to their importance as crop, prey animals, or in recreation in the local region; for example, the movements of caribou herds in Jasper National Park have been tracked by colonists since 1811[78]. Though historical trends in microbial ecology can be extrapolated from archaeological evidence such as ice cores and preserved stomach contents[79], there is much less historical literature on the subject, particularly with regard to abundance and diversity assessments. As such, much of the information we have around how eukaryotic communities respond to oil exposure involves accidental releases of hydrocarbons in transit, and involves a sudden, acute, and temporary hydrocarbon exposure; for example, scenarios may include oil spills from tankers or

pipelines into the ocean. The largest oil spill study followed the 2010 accidental release of billions of barrels of oil into the Gulf of Mexico from the Deepwater Horizon well off the coast of Louisiana—a research effort that is still ongoing[80]. The Gulf of Mexico Research Initiative provided substantial experimental data on algal responses to oil exposure. The results varied; some show differing responses down to the species level with no overall trend toward prokaryotes or eukaryotes as more sensitive or resistant[81], whereas other studies show that within eukaryotes, size of cell and evolutionary derivation may affect hydrocarbon resistance[82].

Most studies of how nonphotosynthetic protists (or, as they are often referred to in ecology, heterotrophic flagellates) are involved in chronic hydrocarbon contamination entail 'before and after' studies comparing communities before and after disturbances occur. For example, protist community changes have been evaluated in this manner for soils associated with chronic hydrocarbon pollution in Brazil[83] and marine environments surrounding oil rigs in Norway[84]. In both cases, substantial changes in community structures were found, with lower community diversity in the disturbed compared to undisturbed environments. In the case of the Brazilian soils, only ciliates were evaluated, and the group Colpodea spp. was identified as being particularly resistant to pollution[83]. In the case of the Norwegian oil rigs, several indicator taxa for disturbances were identified, including multiple ciliate species[84]. As eDNA studies become more prevalent, particularly in the context of evaluating protist responses to hydrocarbons, the list of known hydrocarbon-resistant, hydrocarbon-susceptible, and hydrocarbon-degrading heterotrophic flagellates is certain to grow.

Though the before and after studies provide valuable information into how protists, particularly algae, are affected by exposure to oil and additional chemicals such as the dispersant Corexit[81,85–87], the environmental pressures associated with long term hydrocarbon exposure are very different and may well lead to lasting effects on cellular adaptations, community composition, and related ecosystem function and its services to human society. The Athabasca Oil Sands is an example of an environment with both natural and anthropogenically induced chronic hydrocarbon exposure and provides an opportunity to study the long-term effects of both states on the local environment.

1.4 The Athabasca Oil Sands

The geological deposit of bituminous sands covering Northern Alberta and extending into neighbouring Saskatchewan currently known as the Athabasca Oil Sands has been known to the local Indigenous peoples since time immemorial and was first described by colonial explorers in 1778[88]. The terrain overlaying the Athabasca Oil Sands deposit is part of the Canadian boreal forest[89], which stretches across Western Canada from the Yukon and Northwest Territories all the way into portions of Northern Ontario and Quebec. This boreal region is characterised by not only its vast forested area, but also a high abundance of lakes, streams and rivers. It is also notable for its extreme temperature variations, which can range from +30^{o}C in the summer to -50^{o}C in the winter[89]. The Athabasca Oil Sands, like many northern areas of Canada, was left largely unexplored before Canadian confederation. The first experimental attempts to separate refineable oil from oil sands were not described until the 1880s when Dr. Robert Bell, a representative of the Canadian geological survey, attempted to extract bitumen from oil sands using hot water[88]. However, over the subsequent century, the processes described by Bell were scaled up into a truly industrial operation. Most of the operational mines in the Athabasca region today were established in the 1970s and 1980s[88] when exploitation of nonconventional oil sources became industrially feasible and profitable, and the current bitumen output from the Athabasca oil sands is estimated at about 2.8 million barrels a day[90].

There are two major methods of extraction in the Albertan oil sands region. When industrial bitumen extraction from the Athabasca Oil Sands commenced, open-pit mining was the only commercially viable method of extraction[88]. Open-pit mining (which accounted for the majority of extraction in the 20th century and currently accounts for 500km^{2} of impacted land in the oil sands area[91]) involves the removal of topsoil from a bitumen deposit, uncovering the bitumen-coated sands beneath. The sand is then removed from the site and the bitumen is separated from the sands via a water-intensive washing process; hot water and surfactants are used to separate the hydrophobic bitumen from the extracted rock and silt. The bitumen is separated from the mixture of water and detergents for processing and eventual refinement as crude oil, and the liquid waste, known as tailings, is stored on the mine site[91]. The 500km^{2} figure associated with mine pits does not account for additional disruption associated with maintaining the mine site; for example, linear

disruptions such as roads leading to oil sands mines are known to impact the grazing of local large herbivores and predator-prey relationships with other animals in the region[76].

The second, less disruptive method of oil extraction is Steam-Assisted Ground Drainage (SAG-D), and this method accounts for 80% of the oil extraction currently occurring in the Albertan oil sands region[91]. This protocol involves extraction of oil in situ rather than removal of sands to a different refinery. Steam is pumped into the bitumen deposits in areas 75m deep or greater, causing displacement of the bitumen matrix surrounding the sands. The bitumen is then siphoned directly out of the ground. Though this process is both less water-intensive and has a smaller direct ecological footprint than open-pit mining, the disturbance is more dispersed, with additional linear and edge disruption that can be particularly harmful to local wildlife[92].

SAG-D uses less water to recover bitumen (0.5 barrels of water per barrel of bitumen as opposed to the 2-4 barrels of water per barrel of bitumen necessary for open-pit mining). It also does not produce tailings waste as no sands are removed from the ground[91]. However, it is more economical to extract bitumen from oil sands via open-pit mining when the deposits are close to the surface, and this technique therefore still accounts for 20% of oil sands production[88]. The liquid waste that is used in this process, known as tailings, is then diverted for storage as it cannot be returned to the environment due to its exposure to industrial chemicals[93]. This mixture of water from the extraction process, residual bitumen, silt, surfactants, naphthenic acids, heavy metals, and salts, is currently stored in structures known as tailings ponds (Figure 1.3). They generally consist of the semisettled, partially solid tailings and an overwater cap produced from the water displaced from the settling tailings[94]. This overwater is periodically siphoned off and reused for further bitumen extraction. Fresh tailings from the extraction process (fluid fine tailings, or FFT), gradually settles over time into thicker, mature tailings (MFT) and a water cap[93].

Tailings ponds can be 40-50m deep and cover multiple square kilometres; there are an estimated billion m^3 of tailings currently stored in northern Alberta's mining sites[93]. These tailings ponds covered 88km^2 of the landscape as of 2013[90], and current oil sands mining operations produce an additional 1 million m^3/day of fresh tailings[95]. Since this waste cannot be released into the local watershed until it is deemed to be reclaimed—and therefore no longer potentially harmful to the local flora and fauna—the footprint of these tailings ponds is currently increasing[93]. The

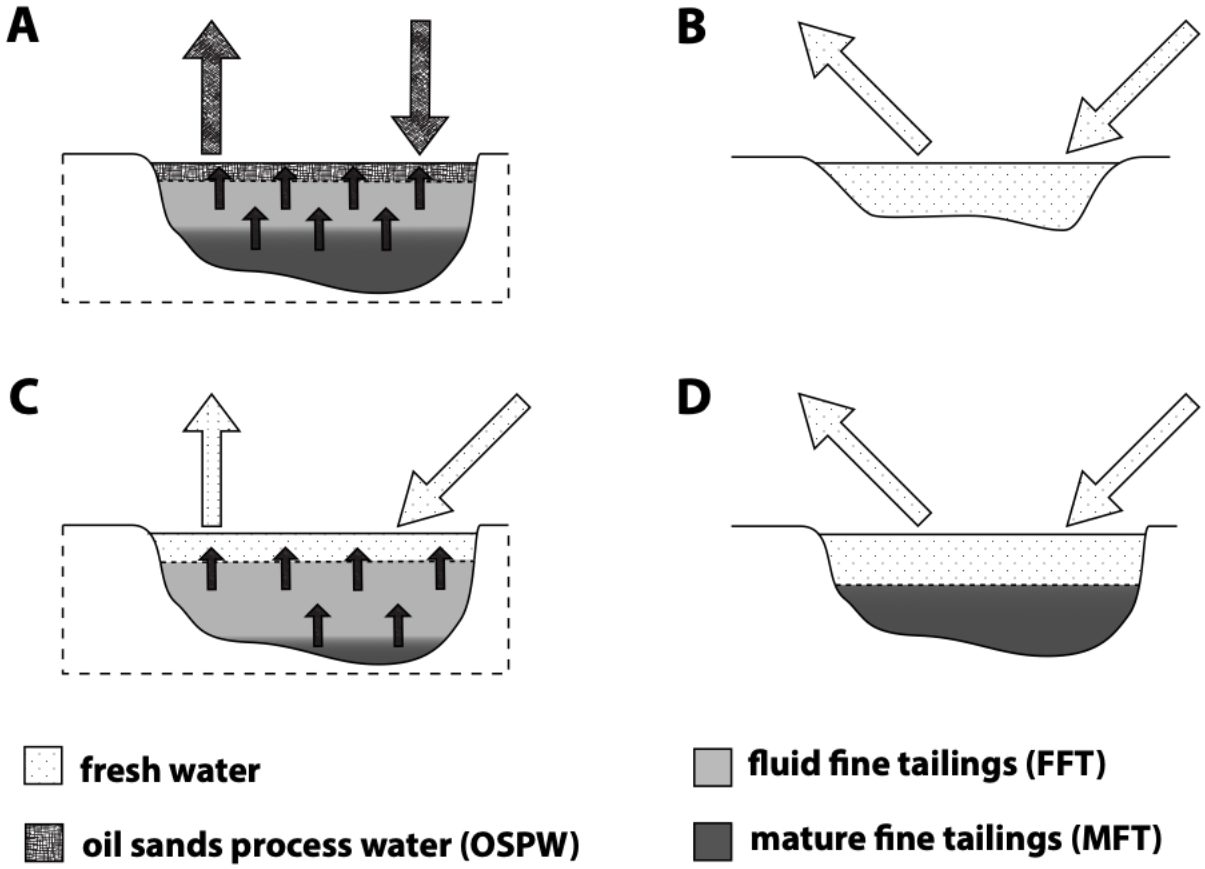


Figure 1.3: Cross-sections of water bodies from the oil sands region
A: An active tailings pond. Tailings are in the process of settling, with porewater from the tailings entering the thin water cap. This OSPW is reused in the extraction process. **B:** A freshwater lake from the Alberta oil sands region, containing freshwater. **C:** An end-pit lake early in the reclamation process. Water enters from freshwater inputs like lakes and is extracted for reuse into the oil sands process. Tailings are in the process of settling, with porewater from the tailings entering the water cap. **D:** Hypothetical end point of reclamation. The tailings are fully settled and separated from the freshwater cap, which is fully integrated into the local watershed.

components of fluid fine tailings (FFT) that are of greatest concern, as well as most extensively researched, are residual hydrocarbons, naphthenic acids, heavy metals, and salts[96]. Despite these ecological challenges, expansion of the Albertan Oil Sands is continuous. The region contains approximately 165 billion barrels of oil harvestable with current technology, and may contain an additional 250 billion barrels of oil that could be extracted if that technology improves[91].

The Canadian Association of Petroleum Producers forecasts a production rate of 4.25 million barrels a day by 2035, up from the current 2.8 million[75]. As reclamation of liquid tailings waste is not a proven technology and is projected to take several decades, it is also necessary to establish methods of identifying early success or promising indicators of reclamation progress[97]. Microbes have long been used as ecosystem indicators, and the ease of eDNA assessments of different environments makes them an even more promising candidate for future use in this area [46].

1.5 Heterotrophic protists in the oil sands region

The Athabasca Oil Sands bitumen deposit ranges from hundreds of metres to only centimetres below the earth's surface, and heavy hydrocarbons are naturally occurring compounds in the local soils and watersheds. Microbial species have presumably adapted to the presence of bitumen in freshwater and soils[98]. Accordingly, one would expect a certain amount of resilience from the microbial communities surrounding tailings ponds to the presence of bitumen and associated hydrocarbons. However, the extent to which this might confer a physiological and community resistance to survival in mine tailings remains unknown[93,95,97]. Ecological studies of the interactions between bitumen and microbial communities in northern Alberta are limited, and most have been performed in the context of establishing a baseline for reclamation of mining-influenced environments. Yergeau et al. (2012) used 454 sequencing to survey the bacterial communities found in natural streams in the Albertan region compared to those with bitumen-containing soils in the riverbed[99]; they found significant differences between bacterial communities of bitumen-associated and nonbitumen-associated streams. The authors suggested a profound difference in the ecology that likely persisted through multiple trophic levels[99]. Reid et al. (2018) also studied metatranscriptomics of hydrocarbon-rich freshwater sediments in the Athabasca Oil Sands region to determine a baseline of microbial processes before the influence of oil sands mining[100]. Though they did not distinguish between protists and bacteria, they focused on chemical processes most

associated with bacterial communities such as nitrogen fixation—and, as noted previously, metatranscriptomic analyses tend to be dominated by bacterial sequences. They found genes associated with the metabolism of nitrogen, sulphur, and methane, as well as hydrocarbon degradation, suggesting that the presence of hydrocarbons was integrated into the sediments and the community was adapted to bitumen presence even outside of the influence of oil sands extraction[100]. In a later paper, Reid et al. (2019) also suggested that from a biogeochemical perspective, artificial reservoirs associated with oil sands mine sites may be a good proxy endpoint for oil sands reclamation as they exhibit many similar microbial metabolic processes to natural lakes, including chemotaxis from hydrocarbon contaminants and high photosynthetic activity[101]. Similarly, Wong et al. (2015) surveyed the bacterial communities of bitumen outcrops in Northern Albertan streams using 16S/18S rRNA sequencing[102]. These outcrops, where heavy bitumen is exposed to the stream directly, reach temperatures of up to 55-60^{o}C when exposed to sunlight. The Wong et al. (2015) study identified a community of hydrocarbon-degrading microbes that included a substantial thermophilic fungal component. This shows that protists are an integral part of the hydrocarbon ecosystem in northern Alberta; fungi in particular are excellent candidates for bioremediation as they contain the majority of highly thermophilic (capable of surviving in 50+^{o}C temperatures) taxa and also have been identified as capable of degrading hydrocarbons in vitro and in situ[95,96,102].

Studies of the bitumen deposits themselves have not yielded any information on microbial eukaryotes; to date there has been no published successful recovery of 18S amplicons from within these environments[94]. However, the bacterial communities within the bitumen have been analysed from cores taken from bitumen deposits using 16S rRNA amplicon sequencing and metagenomics[103]. The most surprising result from these studies has been a considerable contribution to nutrient cycling from an abundant aerobic bacterial population; aerobic taxa such as *Pseudomonas* and *Acinetobacter* were dominant in 16S samples, and genes associated with oxygen-dependent metabolic processes were common in the metagenomes[104,105]. These environments were initially thought to be too challenging for eukaryotic communities; however, amplicon studies of microbial eukaryotes have found thriving communities in extremely challenging environments, including deep in Lake Baikal, in the Dry Valleys of Antarctica, and even in water droplets precipitated into clouds[106–108]. Since initial assessments of the

environmental conditions of bitumen deposits and the composition of the prokaryotic community appear to be incorrect, revisiting the issue of eukaryotic microbes in bitumen deposits would likely reveal some kind of eukaryotic component.

1.6 Tailings ponds and heterotrophic protists

The ecology of tailings ponds, and how reclaimed tailings pond sites can be integrated into local watersheds, is an important question for reclamation of the oil sands region[97]. Initial studies of tailings pond ecology through high-throughput sequencing focused on bacteria as bacterial methanogenesis is responsible for the release of enormous quantities of greenhouse gases – at its peak, the tailings pond Mildred Lake Settling Basin was releasing 40 million litres a day of methane into the atmosphere due to its bacterial community. These studies, summarised in Foght et al. (2017) and Siddique et al. (2018) *inter alia*, showed that there was a diverse bacterial community and extensive nutrient cycling involving anaerobic and aerobic processes in these tailings ponds[94,95]. The first indication that eukaryotes may be present in tailings ponds was reported in Saidi-Mehrabad et al. (2013), a paper that focused on methanogenic and methanotrophic bacteria present in enrichment cultures created from tailings pond samples[109]. They noted the presence of an amoeboid organism that comprised up to 40% of the DNA abundance in metagenomes of these enrichment cultures after 48 hours, apparently feeding on the bacteria. However, though they identified the amoeba as most closely related to the known species *Protacanthamoeba bohemica*, they did not determine its origin or definitively show that it came from the tailings pond sample rather than postcollection contamination[109]. *P. bohemica* was originally identified as a parasite of freshwater fish and has since been isolated in drinking water and in environmental screens of freshwater; switching between parasitic and free-living heterotrophy has been noted in other amoebae such as the human pathogen *Naegleria fowleri*[110]. This suggests that *P. bohemica* would theoretically be capable of surviving in low light and potentially nutrient poor tailings as well as freshwater, but it is not conclusive. Amoebae have also been noted for their ability to degrade hydrocarbons in mesocosm experiments[111].

The most comprehensive study of eukaryotes in tailings ponds comes from Aguilar et al. (2016)[112]. In this study, the eukaryotic community of the soft, anoxic tailings sediments of West-in Pit and the water cap of Mildred Lake Settling Basin (MLSB), two tailings ponds in northern Alberta,

were characterized before the decommission of these sites in 2012. This study used 18S-optimised amplicon sequencing to ensure the maximum possible signal from the eukaryotic cells found within these tailings, but sequences identifiable as eukaryotes were sparse—to ensure the most accurate possible classification of this small eukaryotic sample, the authors used phylogenetics to determine how closely the extracted sequences were related to known species[112]. Most of the extracted sequences were heterotrophic, from the groups Rhizaria, Amoebozoa and Fungi; this is consistent with the low light and low oxygen environment. However, there was a notable presence of sequences from majority photosynthetic clades such as Euglena and Chlorophyta[112]. While the number of sequences extracted was too small for identification of any trophic webs or ecological trends associated with eukaryotes, the consistent detection of a eukaryotic signature in the three samples of tailings sediments argued against the eukaryotic signal merely being an artefact and instead strongly shows a eukaryotic community in this environment.

The authors also mined metagenomes produced from the MLSB sediments and the overwater, taken from the thin oxic layer at the top of the water cap, for eukaryotic sequences. A relatively low abundance of eukaryotes detectable using metagenomic methods combined with the low eukaryotic presence meant even fewer sequences were identified; those that were present included heterotrophs, phototrophs, and also some sequences classified as Metazoa, suggesting the presence of zooplankton or insects[112]. However, these sequences may not have been derived from living organisms and therefore cannot be used to make any ecological conclusions. However, it is notable that a substantial percentage of the initial protist diversity identified in studies of hydrocarbon-associated water bodies are heterotrophic; heterotrophic flagellates have historically been understudied compared to their photosynthetic algal counterparts[32,33].

1.7 Base Mine Lake, a pilot tailings reclamation project in the oil sands area

Though land previously associated with oil sands mining has, in some cases, been certified as reclaimed[113], it is a testament to the difficulty of reclaiming fluid fine tailings (FFT) that there is no scientifically or legally recognised strategy of reintegrating the liquid in tailings ponds back into the local watershed[97]. Studies into how oil sands process water (OSPW) affect protists in situ have been underway since the 1980s, when oil exploration in northern Alberta expanded as an industry[114]. These mesocosms, which have been established at various sites in Alberta, have tested

how indigenous microorganisms react to being exposed to the OSPW and FFT. Most of the eukaryotes observed in these studies are phototrophs as primary production is essential for nutrient cycling[115]. However, some phototrophs indigenous to OSPW may also have a role in direct bioremediation. Ruffell et al. (2016) incubated 21 algal strains extracted from sites within or near oil sands mining operations with OSPW and used mass spectroscopy to determine the fraction of oil degraded by the algae in culture[116]. Though multiple eukaryotic algae were used in the experiment, the only species identified as a candidate for industrial bioremediation was a cyanobacterium. Leung et al. (2003)[114] used cell counts to observe the eukaryotic phototroph communities in multiple water bodies in the Fort McMurray region, some of which were affected by OSPW. They noted that Chlorophyta were particularly dominant in their samples, particularly in those highly affected by OSPW, and suggested that indigenous phytoplankton communities may be somewhat resistant to the presence of hydrocarbons and associated contaminants[114].

In 2012, the largest reclamation site for FFT was established at the Syncrude mine site near Fort McMurray (Figure 1.4). This site, called Base Mine Lake (BML), is an example of a proposed FFT reclamation technique called 'wet-capping'[117]. FFT from a mine site (in this case, the decommissioned tailings pond Mildred Lake Settling Basin) are covered with a layer of freshwater that is continually replenished until the tailings are settled into a thick layer and the freshwater has diluted all potential contaminants to ecologically tolerable conditions. Geophysical studies of BML have shown positive change over the last five years. Oxygen penetration into the water is comparable to natural lakes in the northern Alberta region, and heavy metal and hydrocarbon contaminations levels have returned to levels which are considered acceptable for undisturbed lakes in the local area[118]. However, light penetration and salt concentrations remain an issue; in particular, the salt and pH levels are projected to remain elevated for decades to come[118]. Studies of the bacteria and geochemistry of BML have detected an abundance of methanotrophic and nitrifying bacteria, with nitrification becoming a particularly prevalent process since 2016. This is potentially problematic from a reclamation perspective as both of these processes consume oxygen; while the top layer of the water cap, created from a combination of OSPW and freshwater from a local reservoir, has comparable oxygen concentrations to a natural Albertan lake, the water cap close to the tailings interface has persistently low oxygen concentrations[119]. Modelling of the current rate of oxygen consumption by bacteria in the tailings/water interface suggests that,

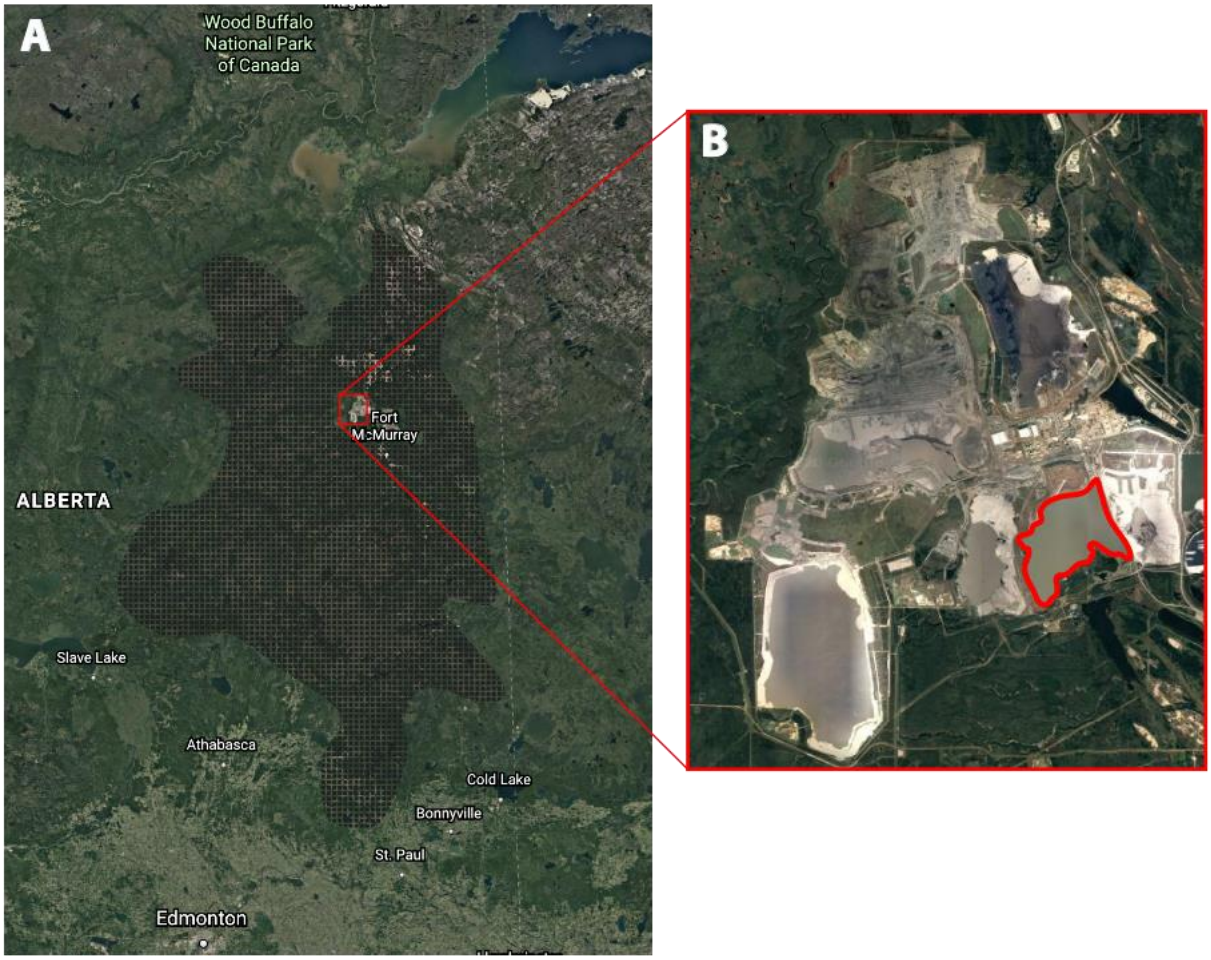


Figure 1.4: Map of the Athabasca Oil Sands Region
Images taken from Google Earth. **A:** The overall span of the Athabasca Oil Sands deposit, shaded in brown. Note that there are additional deposits by Cold Lake and Peace River, not indicated, and further oil sands across the provincial border in Saskatchewan. **B:** A close-up of the Syncrude Mildred Lake mine, with the position of Base Mine Lake bordered in red.

without intervention, these processes could cause transient or complete anoxia of the lower 3.5m of the 10m-depth water column[120]. Bacteria in BML also showed substantial depth stratification both functionally and taxonomically[121]. In studying the biogeochemical processes associated with the water capping technology under study in Base Mine Lake, novel organisms and processes have been identified. A novel lineage of mixotrophic bacteria associated with nitrogen cycling, MBAE14, and a novel bacteria-infecting phage have both been identified as metagenome-assembled genomes found in samples from BML. Also, these metagenomic studies identified a novel form of enzyme capable of oxidizing short-chain alkanes, ammonia, and methane from known lineages of *Betaproteobacteria*[122].

OSPW from BML have also been used for numerous microcosm and in vitro studies. As BML has been the only full-scale end-pit lake in the oil sands region for several years, water column studies and geophysical modelling have been extensively used to predict its responses to both environmental changes and industrial intervention. For example, Kong et al. (2019) used samples and field data from BML to validate their model of methane emissions from end-pit lakes, extending their model of methane emissions from tailings ponds more generally[123], and Poon et al. (2018) used BML samples to experiment on methods of improving clarity and reducing turbidity by altering pH in end-pit lake water caps[124]. Cap water from BML, alongside OSPW, has also been used to test the effects of water cap technology on organisms which would be expected in the end-pit lake environment at the end stages of reclamation. As one might expect for a lake, much of this research in multicellular organisms has focused on fish. Embryos and juvenile fish from Japanese medaka (*Oryzias latipes)*[125], rainbow trout (*Oncorhynchus mykiss*)[126], walleye (*Sander vitreus*)[127], and fathead minnow (*Pimephales promelas*)[128] have been experimentally exposed to OSPW, with noted toxic and mutagenic effects at high concentrations. Efforts have been made to understand the mechanisms of this, particularly in the model organism *O. latipes*; these studies suggest that toxic effects are likely induced via oxidative stress pathways and inhibition of ABC transporters in the early stages of development[125,129]. However, it is also worth noting that these effects are only observed at concentrations of OSPW much higher than those found in the BML water cap, such as those found on the caps of active tailings ponds[125]. Microbial inocula isolated from BML cap water have also been evaluated for their biodegradation potential of bitumen and napthenic acids. Yu et al. (2018) found that BML microbes were able to

successfully degrade residual bitumen, the addition of which caused a shift in the microbial community toward a higher abundance of known hydrocarbon-degrading taxa[130]. When presented with model napthenic acids, short-chain hydrocarbons commonly used in the bitumen extraction process, the model algae *Chlorella kessleri* and *Botryococcus braunii* were able to degrade these chemicals much more effectively when cocultured with bacteria indigenous to BML[131].

Every mine site in the Athabasca Oil Sands region produces substantial quantities of tailings, and all have some form of pit lake infrastructure that will have to be reclaimed after production ceases. BML is the only full scale reclamation site in the region, but other demonstration lakes have since been established, such as Lake Miwasin at the Suncor mine site[132]. Due to the quantity of tailings that will need to be incorporated into the environment and the individual approaches each mine site is taking to reclaim their tailings ponds, each individual end-pit lake system will be an extremely large and complex water body with its own unique biology and history[93]. Because eDNA approaches allow us to evaluate microbial communities on a high-throughput scale, there is potential for the integration of holistic bioindicator taxa that can be empirically shown as influenced by the anthropogenic disturbance specific to the industrial site under reclamation. An advantage of eDNA approaches is all taxa—whether extremely large and distinct or tiny and relatively homogenous in appearance—are evaluated using a molecular approach; therefore, ease of visual identification is no longer a relevant factor for establishing bioindicator taxa, leaving a gap that has the potential to be filled by heterotrophic protists or heterotrophic flagellates.

1.8 Hydrocarbons and heterotrophic protist cell biology

Beyond understanding the effects of hydrocarbons on microbial communities, it is equally important to understand the effects of hydrocarbons on the cells of organisms. Tailings waste has been applied in vitro to multicellular organisms, including HeLa cells and fish embryos[125,127,133]. These experiments have shown profound local toxicity to these multicellular tissues and organisms, implying that macrofauna such as fish and plants would struggle to survive in the environmentally challenging conditions of tailings, fresh process water, and early stage reclamation environments, though the toxicity of end-pit lake waters would certainly decrease over time due to dilution and reclamation processes[118]. This is another aspect in which eukaryotic microbiological studies provide potential advantages to our understanding of these environments.

By studying these organisms, which can be grown quickly and inexpensively in culture and have a similar fundamental cell structure to most macroflora and fauna, they can be used to establish potential cellular processes affected by hydrocarbons that can be expanded into the health of other organisms or into ecosystem processes[134].

Rogerson and Berger carried out one of the earliest studies of the effects of hydrocarbons on protists in vitro in the late 1980s[135]. They made use of electron microscopy to examine the effects of crude oil on the ciliate *Colpoda colpodium*, a common freshwater protist that has (since this study) been shown to exhibit remarkable resistance to hydrocarbons. Rogerson and Berger also noted that exposure to crude oil was not fatal to the organism even at relatively high concentration, but did result in pronounced visible morphological changes including misshapen organelles and an increase in intracellular vesicles[135]. This showed that the membranes of the organism were affected heavily by the introduction of hydrocarbons, which is unsurprising since these cell structures are lipid-based. This is consistent with eukaryotes being more heavily affected by hydrocarbons than bacteria as their cell biology is more heavily reliant on membranes and compartmentalisation of cellular processes by lipid-bound structures[28].

However, other studies have suggested that, rather than being incapacitated by hydrocarbons, protists may have an active role in bioremediation. The exact process for protists' contributions to hydrocarbon degradation is unknown; some in vitro studies have shown that inhibition of protist grazing inhibits hydrocarbon degradation[136,137], suggesting that protists are enhancing degradation by preying on bacteria, while other studies suggest that protists are degrading hydrocarbons directly. Also, Gilbert et al. (2014)[138] showed that heterotrophic grazers such as ciliates may act as biological dispersants of hydrocarbons, contributing to the bioremediation of hydrocarbon contamination by breaking down the lipid droplets that can cut off oxygen to the water column and also prove lethal to macrofauna and flora[138]. These studies demonstrate some of the benefits of including heterotrophic protists as reclamation indicators; they represent a trophic level intermediate between bacteria, which are profoundly affected by geochemistry, and macrofauna, which may be too affected by local toxicity to establish populations. Many heterotrophic protists are highly resilient and can survive low light, periods of partial or complete anoxia, and presence of toxic compounds[139]; even primarily phototrophic algae employ heterotrophic strategies in

periods of nutrient limitation or low light[140]. Industrial evaluation of protist communities has become common practice in wastewater treatment plants ever since it was established in the 1970s that 'ciliated protozoa' are essential for water treatment to be successful[141]. The limitations and successes of heterotrophic protists in wastewater treatment are reviewed in Foissner et al. (2016), and his most relevant conclusion in the context of oil sands reclamation is that for heterotrophic protists to be useful as bioindicators, one must understand the taxonomy and ecology of these enigmatic organisms[141]. eDNA and other molecular approaches appear to be filling this niche even with the undisputed decline of microscopy as a relevant technique for community ecology[13]. Phylogenetic approaches have led to the expansion of our understanding of two of the most morphologically homogenous yet genetically diverse of these heterotrophic clades; the fungal Microsporidia[33] and the cercozoan Glissmonada[32,142]. The explosion of identified species belonging to these clades, as well as an increased appreciation of their ecological abundance and relevance, indicates the growing potential for heterotrophic protists as underappreciated bioindicators, both of industrial disturbance and reclamation.

1.10 Thesis scope

This thesis examines the heterotrophic nanoflagellate communities of Base Mine Lake, an end-pit lake in northern Alberta over four years in its early reclamation process (2015-2018) via a series of amplicon-based eDNA community assessments. I also investigate the cell biology of membrane trafficking pathways associated with resistance to hydrocarbons in the ciliates, a phylum noted for its use as a bioindicator in anthropogenically influenced environments.

What is the heterotrophic community composition in the early reclamation period of Base Mine Lake? In the second chapter, I evaluate the heterotrophic protist diversity in detail for Base Mine Lake in the summer of 2015, two years after water-capping. I use a combination of diversity analyses and phylogenetic taxonomic analysis to establish month-to-month trends across the lake and also identify potentially novel diversity in five mostly heterotrophic groups (amoebozoa, cercozoa, Ciliophora, excavata, and fungi). This chapter represents the first evaluation of protist diversity in a tailings pond reclamation environment.

What is the heterotrophic community composition of Base Mine Lake over time? In the third chapter, I continue to use the combined diversity analysis and phylogenetic taxonomic analysis to determine the heterotrophic protist community over four summers, beginning with the 2015 data. I determine whether, at this early stage of reclamation, any trends can be observed seasonally in any of the heterotrophic groups and find that most of the variation is explained by the year of sampling, suggesting that the heterotrophic microbiome is still being established in this early stage of reclamation. I also use cooccurrence analysis to determine whether there are any specific sequences that appear to correlate strongly with reclamation time in linear or nonlinear fashions to identify any other potential organisms of interest that may act as bioindicators for early reclamation.

What is the baseline genetic diversity of the ecologically relevant heterotrophic phylum Ciliophora, and how can this inform our understanding of hydrocarbon resistance? In the fourth chapter, I combine eDNA, transcriptome, and genome information from the ciliate phylum to determine a baseline level of diversity across the ciliate membrane trafficking system by using comparative genomics and phylogenetics. The membrane trafficking system has been strongly implicated in resistance to hydrocarbons and has been extensively studied in model Oligohymenophorean ciliates. I use comparative genomics to extend these findings across ciliate diversity and determine how widespread ciliate-specific membrane trafficking innovations are within more basal ciliate clades.

How does the collected data inform reclamation in the Athabasca Oil Sands region? In the fifth chapter, I comment upon the usefulness and relevance of heterotrophic flagellates, and particularly ciliates, as reclamation indicators in the Albertan Oil Sands region. Ultimately, most of the heterotrophic groups at the community level seem to still be in flux at this stage of tailings pond reclamation. However, there are multiple heterotroph species, including ciliates, which appear to show a much stronger response to tailings reclamation and may be useful as future bioindicators of reclamation success. There are also groups of heterotrophs which seem to thrive at different stages of the reclamation process, and some evidence for replacement of heterotroph niches in the reclaimed community.

Chapter 2

PHYLOGENETIC ESTIMATION OF COMMUNITY COMPOSITION AND NOVEL EUKARYOTIC LINEAGES IN BASE MINE LAKE, AN OIL SANDS TAILINGS RECLAMATION SITE IN NORTHERN ALBERTA

2.1 Preface

One of the most substantial challenges for reclamation in the Athabasca Oil Sands is the unprecedented scale of the operation. Entire landscapes displaced by mining activity have to be reformed, including wetlands[143]. While some formerly aquatic sites have been confirmed via conversion to grassland or forest (for example, via the filling of a mining pit that used to be a lake before mining), reclamation of liquid tailings into a lake environment that is integrated into the local watershed has not yet successfully been completed[143]. Base Mine Lake is the first such pilot project and has been developed from mesocosm and smaller scale test ponds on the Syncrude site dating back to the 1980s[144].

The problems with assessing reclamation success in Base Mine Lake (BML) are therefore twofold: since the environment under development (an end-pit lake) has no true analogue in the natural environment, it is necessary to establish both what a successful endpoint for reclamation will be as well as the protocols necessary to reach this point[143]. This chapter aims to address these issues by surveying the microbial heterotroph environment in BML in the ice-free summer period of 2015. BML's initial water-capping process had been completed in 2013, and dilution with freshwater from Beaver Creek Reservoir (BCR) meant that most of the industrial contaminants like heavy metals in the overwater were at a negligible concentration; however, the environment in BML was still far from that of a natural Albertan lake[145]. Most notably, the turbidity of the lake was extremely high, and the salinity was also elevated compared to the local watersheds. In the winter of 2016, due to industrial intervention and the Horse River Fire (an enormous wildfire that consumed much of the nearby city of Fort McMurray and the surrounding boreal forest), there were considerable changes in the geochemistry and abiotic variables of BML and the local environment[146]. The summer of 2016 saw a significant reduction in turbidity, which led to a proliferation of photosynthetic organisms and a return to the phototrophic blooms common in other

lakes in the region (discussed in more detail in Chapter 3). This meant that 2015 was an interesting sample point both from an environmental (as an example of an end-pit lake in very early stages of reclamation) and a microbiological perspective (as an example of the microbial community before phototroph blooms returned).

There were two major aims in the study described in this chapter. The first was to establish the baseline diversity of BML in its early stages of reclamation, as a mechanism for tracking the changes in this microbial community during active reclamation. We determined that BML in 2015 was an environment dominated by heterotrophs as phototrophs had not yet become established. The second aim of this study was to establish BML as a site for bioprospecting, whether for novel diversity within the microbial heterotrophs or for novel adaptations within known groups. Bioprospecting studies in understudied environments have proven fruitful for uncovering both new lineages, such as the recently identified *Hemimastix kukwesjijik* that solidified Hemimastigophora as a new kingdom level of diversity[147], and for uncovering previously unidentified life histories from known groups, such as the free-living Rhodelphia as a sister clade to the photosynthetic red algae[148]. We used phylogenetics to manually annotate the taxonomic affiliations of each of the OTUs identified in this study and determined a considerable proportion were more related to each other than any other known publicly available sequences; this indicated that BML was an excellent site at which to identify novel eukaryotic diversity. This chapter provides the first assessment of community level eukaryotic microbial diversity in a reclamation environment in northern Alberta after Aguilar et al. (2016)[149] demonstrated the presence of a diverse eukaryotic community in the tailings ponds of the Athabasca oil sands.

For the work in this chapter and the subsequent chapter, I would like to thank employees at Syncrude and on the Base Mine Lake project for providing their samples and expertise. I would also like to thank Peter Dunfield and everyone at the Dunfield lab for their insights and assistance on the project, and in particular Angela Smirnova for her technical expertise and skill in extracting and sequencing the eDNA from the Base Mine Lake samples. I'd like to thank Maria Aguilar for her help getting started on eDNA, and for her support as I started out as a graduate student. David Bass also provided invaluable advice and support on this project, and I would like to thank him for hosting me at the Natural History Museum in June 2017 for a crash course in eukaryotic

phylogenetics and taxonomy. Lucas Paoli's pipeline for OTU clustering, described in Chapter 2, was also my introduction to programming in R. While I am definitely not thanking him for that, his ecological expertise on the other hand was an excellent addition to the project and one for which I am grateful. Giselle Walker and Micah Dunthorn also provided the pan-eukaryotic and ciliate backbones, respectively, used in the phylogenetic analyses.

2.2. Introduction

2.2.1 Protists in hydrocarbon-influenced environments

Eukaryotic microbes are key players in ecological communities. Photosynthetic eukaryotes such as diatoms and dinoflagellates are responsible for most of the carbon fixation in the open ocean and contribute massively toward global nutrient cycling[150]. Soils contain substantial communities of microbial eukaryotes with diverse ecological niches. Detrivores such as fungi degrade dead and decaying matter, heterotrophic eukaryotes consume bacteria and other microbial eukaryotes, and parasitic eukaryotes can encyst in soils or parasitize other eukaryotes[151,152]. As our understanding of the eukaryotic microbial world increases, the role of environmental stressors becomes more evident. Eukaryotic communities can be influenced by anthropogenic change (such as hydrocarbon exposure) in their local environments. Exposure can occur through industrial extraction and processing, spills and leaks during transport, or in disposal of waste products. In the aftermath of the Deepwater Horizon disaster, where 4.9 million barrels of oil were released into the Gulf of Mexico, intensive monitoring of this environment has been performed[153] The effects of crude oil, oil derivatives, and hydrocarbons on phytoplankton and zooplankton have been well studied, showing that while the phytoplankton were sensitive to the presence of hydrocarbons, this is highly dependent on species and environmental conditions[154–158]. Microbial eukaryotes also have their own roles to play in reclamation of hydrocarbon-affected environments. Many species of fungi are able to directly degrade hydrocarbons[159,160], and there is evidence that protist grazing increases the efficiency of hydrocarbon degradation by bacterial communities [161]. Biophysical studies show that grazing organisms, particularly ciliates, may act as natural hydrocarbon dispersants as they move through the water column looking for food[162]. Ciliates in particular have been suggested as bioindicators for urban pollution due to their sensitivity to changes in contaminants in waste water[163,164]. Hydrocarbon exposure, especially in extraction sites, is also accompanied by exposure to other environmental stressors such as salts and heavy metals. Adaptations of eukaryotic

microorganisms to high-salt environments and environments contaminated with heavy metals have been extensively studied in a few model species, such as the halophile *Dunaliella salina* and the metal-tolerant *Euglena gracilis*. In these species, biochemical pathways that attenuate the cellular damage caused by acid and metal ions have been determined, including high levels of beta-carotene and glycerol in *D. salina* and glutathione in *E. gracilis*[165,166]. Both organisms also show adaptations in their membrane trafficking system to adapt to elevated export of organic acids[167]. Although it is well known that eukaryotic microbes, both as community assemblages and on a cellular level, are affected by hydrocarbon exposure, few studies have focused on their remediation potential.

2.2.2 Tailings, tailings ponds and end-pit lakes

Extraction of marketable oil from oil sands requires treatment with hot caustic water along with naphtha and/or paraffin solvents[168]. This extraction process produces fluid fine tailings (FFT) comprised of water, silt, clay, residual bitumen organics, and unrecovered solvents. FFT porewater contains several compounds of concern, including naphthenic acids. FFT is contained in basins or tailings ponds to dewater. The tailings solids are reclaimed, and the water is a source of recycled water for bitumen extraction. While construction of reclaimed landscapes from some oil sands mining sites has been successful[169], reclamation strategies for the tailings ponds have not yet been fully developed and tested[143]. To remedy this, there are numerous projects in northern Alberta aimed at investigating tailings pond reclamation techniques. One of these is Base Mine Lake, a former mine pit under reclamation using the technique of water-capping, where a water cap (in the case of BML, at an average depth of 8 m) is placed over tailings waste (in the case of BML, at an average depth of 45 m)[145,170]. This forms an end-pit lake (EPL) which is hypothesised to physically sequester the tailings beneath the water cap and demonstrate improved water quality over time. In the long term, it is hoped that it will become a functional component of the closure landscape and support locally common flora and fauna[171]. BML was filled with tailings over the course of nearly 2 decades, then capped with water in 2013. The progress of this site has been rigorously monitored and researched ever since[170]. While metal ion concentrations have been well within water quality guidelines since 2016 due to dilution with freshwater, the salinity (due primarily to Na^{+}, Cl^{-}, SO_4^{2-} and HCO_3^{-} levels) is still elevated as compared to Athabasca River water (median conductivity=2700 S/cm; median total dissolved solids=1700 mg/L). BML acts as a typical

temperate dimictic lake, with spring and fall mixis events. In summer, the water cap is thermally stratified with a corresponding vertical gradient of dissolved oxygen (DO) from oxic in the epilimnion to close to anoxic in the hypolimnion[172]. In 2015, BML exhibited reduced light penetration and high turbidity due to residual suspended solids in the water cap[173]. Turbidity was highest in the spring after ice melt and gradually reduced over the summer until the lake iced over again in late October. Also, turbidity was transiently increased during the summer by storms[174]. Further, White and Liber[145] reported a wild population of the 'water flea' *Daphnia pulex* in samples collected from Base Mine Lake in 2016.

2.2.3 Protists in the oil sands

The in vitro effect of the contaminants present in tailings on organisms from bacteria to vegetation to fish is well documented[175–177]. It has already been noted that abundant and diverse bacterial communities are associated with tailings pond water[168]. Oil Sand Process Water taken from Base Mine Lake in the early stages of reclamation (between 2012 and 2016) has shown mutagenic and genotoxic effects on cell cultures and *Oryzias latipes* (medaka fish) embryos[178,179]. In environmental metagenomic studies, eukaryotic microbial signatures are often overwhelmed by the sheer quantity of prokaryotes, making detection of rare species impossible. Aguilar et al. (2016)[149] used metagenomic techniques to analyse the presence of eukaryotes in West In-Pit, a tailings storage facility, but these techniques proved to be limited as it was difficult to recover eukaryote DNA. To overcome these technical limitations, the authors used a PCR amplification approach of the V4 region of the 18S rRNA genes and showed that eukaryotic microbes were indeed present in tailings ponds and reclamation sites[149]. However, as this paper was based on an extremely small number of samples and recovery of eukaryotic sequences was poor, it was impossible to determine the exact diversities of these communities. Another notable result from Aguilar et al. (2016)[149] was that the majority of the detected sequences were not present in reference databases and were often highly distinct from their closest matches in GenBank, which points to the existence of substantial novel diversity in these sites.

2.2.4 Scope of study

High-throughput 18S rRNA gene amplicon sequencing was used to determine the microbial eukaryote community present in the reclamation site Base Mine Lake over the summer of 2015.

The results show a diverse and novel eukaryotic microbial community dominated by heterotrophs. These results have the potential to inform reclamation efforts in the region and indicate that in its current state, Base Mine Lake is a good site for bioprospecting for hydrocarbon-tolerant eukaryotes.

2.3 Methods

2.3.1 Sampling and water chemistry

Samples were taken from three platforms located at the centre of the lake (Platform 1), the north eastern section (Platform 2), and the south western section (Platform 3) of the lake (Figure S2.1) at intervals from below the surface to the tailings/water interface. Sampling took place from March to September 2015. Metadata from each sample can be found in Table S2. 1L polycarbonate containers for water samples were rinsed three times with water before being filled to the top and capped. These were transported to the University of Calgary on ice in coolers at about 5$_{o}$C and processed within 7 days of sampling. Water was slightly saline (2700 µS/cm) and alkaline (pH 7.9 – 8.5); additional water chemistry information was not available. Biological materials from 1L of water samples were harvested at 8,000 × *g* for 10 min at 4ºC using the Avanti J-E high-performance centrifuge (Beckman Coulter Life Sciences, Indianapolis, USA) and stored at -20ºC until processed for DNA extraction.

2.3.2 DNA extraction, PCR amplification, and sequencing

DNA was extracted using FastDNA Spin Kit for Soil (MP Biomedicals, Solon, OH, USA). Short fragments (380 bp) of the V4 region of the eukaryotic 18S rRNA gene were amplified using universal primers for eukaryotes (bolded) [180] with Illumina adapters attached to the 5'-ends (not bolded) : Ill_18S_F 5' -TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCAGCA(G/C)C(C/T)GCGGTAATTC C-3'; Ill_18S_R 5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGACTTTCGTTCTTGAT(C/T)(A/G)A -3'. Each 30 µl PCR mixture contained 1 µl of DNA template, 1.56 µl of primer solution (10 µM), 3 µl of 10x PCR buffer (Life Technologies Inc. Burlington, ON, Canada), 2 µl of deoxynucleoside triphosphate (dNTPs) (25 mM each), 22.23 µl of sterile nuclease-free water (Qiagen, Toronto, ON,

Canada), and 0.25 µl of Pfu polymerase (2.5 U μl^{-1}) (Life Technologies Inc. Burlington, ON, Canada). PCR amplification was performed in a Veriti™ 96-Well Fast Thermal Cycler (Life Technologies Inc. Burlington, ON, Canada). The conditions for the first round of PCR were the following: an initial denaturation at 95ºC for 5 minutes, followed by 10 cycles of touchdown PCR (denaturation at 95ºC for 30s, 30s at a primer annealing temperature of 60ºC for the first cycle and subsequent decrements of 0.5ºC per each subsequent cycle, and 30s of elongation at 72ºC), followed by 30 cycles with constant annealing temperature (denaturation at 95ºC for 30s, 30s at a primer annealing temperature of 55ºC, and 30s of elongation at 72ºC), and a final extension at 72ºC for 5 minutes. The PCR products were visualized on a 2% agarose gel using a fluorescent dye, SYBR Safe DNA Gel Stain (Life Technologies Inc. Burlington, ON, Canada). The PCR products were purified using Omega Mag-Bind RXNPure Plus beads (VWR, Edmonton, AB, Canada) according to the manufacturer's instructions. In the second round of PCR, initial PCR products were barcoded. For this, 5 µl bead-purified initial PCR products were mixed with 5 µl of forward barcoded primer (1 µM), 5 µl of reverse barcoded primer (1 µM), and 35 µl of KAPA HiFi HotStart ReadyMix (VWR, Edmonton, AB, Canada) to adjust total volume to 50 µl. Cycling conditions were the following: an initial denaturation at 95ºC for 3 minutes, followed by 8 cycles with denaturation at 95ºC for 30s, 30s at a primer annealing temperature of 55ºC, and 30s of elongation at 72ºC, plus a final extension at 72ºC for 5 minutes. After the second PCR round, the final PCR products of 516 bp were visualized on a 2% agarose gel and purified again with Omega Mag-Bind RXNPure Plus beads. Concentrations of all 18S rRNA gene amplicons were quantified using a DNA quantification kit, sDNA HS Assay Kit, and a small benchtop Qubit fluorimeter (Life Technologies Inc. Burlington, ON, Canada). For sequencing, quantified amplicons were diluted to 4 nM and pooled. Illumina sequencing was done at the University of Calgary, Alberta, Canada, using the MiSeq instrument and a sequencing kit, MiSeq® Reagent Kit v3 (600 cycle) (Illumina Canada Ulc, Vancouver, BC, Canada).

2.3.3 OTU clustering

The Illumina.fastq reverse and forward reads were first trimmed using seqtk 'trimfq' with default parameters (https://github.com/lh3/seqtk.git), paired using the USEARCH package[181], with a minimum overlap of 30 bp, and merged into a single file. The paired-end reads were quality controlled using the following conditions: minimum length of 300 bp, no uncalled base pairs, and

a maximum expected error of 0.75. Reads were dereplicated using default parameters, and clustered at 97% similarity using the UPARSE algorithm as implemented in USEARCH[181]. We chose a threshold of 97% to ensure that downstream phylogenetic conclusions regarding novel diversity would be robustly supported though, as a result, species-level diversity may be clustered within the same OTU. To ensure that this broad OTU clustering parameter did not result in substantial underestimation of the overall microbial eukaryote community, we also denoised the dataset into amplicon sequence variants (ASVs) using the dada2 pipeline. We used this dataset concurrently with the OTU dataset when carrying out ordination analyses to ensure that the results were supported by both techniques (Figure S2.6).

2.3.4 OTU identification

Preliminary OTU identification was provided by automatic BLAST of OTUs against the SILVA database[182,183], and discarding any that were derived from bacterial or from organellar genomes, or from multicellular organisms (Metazoa or Embryophyta). Each OTU was then manually analysed via BLAST against the GenBank database, and if classification via GenBank was associated with a single taxon (top 10 BLAST hits all that taxa with clear separation from other taxa with a difference in E values of at least three orders of magnitude) then the classification was considered robust. Also, any sequences for which the 5' and 3' ends showed different evolutionary derivations were discarded as chimeric. For OTUs that BLAST analyses identified as uncultured taxa or could not be clearly identified as a single organism, classification was based on their placement in a pan-eukaryote phylogeny. Sequences were aligned to the reference backbone used in Aguilar et al. (2016)[149] and a Maximum Likelihood tree was generated using RAxML BlackBox[184]. Once a tree was obtained with a robust resolution of phyla, OTUs classified within those phyla were compared to the classifications obtained from BLAST results. For each OTU, classification was determined down only to the level that could be robustly assigned, with the same classifications using all four reference methods (automated BLAST against SILVA and PR2, manual BLAST against GenBank, and an association with the relevant clade in a pan-eukaryote phylogeny). These classification levels were visualised using the KronaTools package in Krona[185].

2.3.5 Phylogenetic placement of OTUs

Pplacer[186] trees were generated using the SILVA-filtered OTU list and a pan-eukaryotic backbone alignment generated in Aguilar et al. (2016)[149] for classification of unknown V4 OTUs. The backbone tree for this alignment was generated using the RAxML BlackBox algorithm on the CIPRES web server[184]. OTUs were added using pplacer default parameters and converted to PhyloXML format using guppy[186]. The tree was visualised using Archaeopteryx[187].

2.3.6 Phylogenetics

We produced an initial pan-eukaryotic tree using the eukaryote backbone from Aguilar et al. (2016)[149], and sorted the OTUs into the higher level heterotroph groupings of interest (Amoebozoa, Cercozoa, Ciliophora, Excavata, Fungi) based on both their classification in the reference databases and their position within this preliminary phylogenetic alignment[184]. We produced each specific phylogenetic tree using a backbone alignment containing full ribosomal regions from representative species across the diversity of the queried group, OTUs from the BML samples classified to the group, and the entire ribosomal region top hit obtained from BLASTing the OTU sequences against GenBank, including unclassified and uncultured species. We aligned the sequences using the MAAFT algorithm with E-ins-i and otherwise default parameters[188]. We constructed Maximum Likelihood trees from the alignments using the RAxML BlackBox algorithm on the CIPRES phylogenetics web server[184,189].

2.3.7 Ordination

Using the OTU abundance tables for the Base Mine Lake samples, we used the metaMDS function with default parameters in the R package vegan 2.2.4 to create an NMDS plot of the variation of OTUs across samples[190]. We ensured that all samples included in the analysis had at least 100,000 reads. The samples converged in 2 dimensions with an overall stress of 0.224, meaning that the analysis provided a useful representation of the communities. The results of this analysis were visualised using the ggplot2 package in R[191]. We also estimated the percentage of the variation explainable by any given metadata variable using the PERMANOVA function in vegan with default parameters[190]. We also completed this ordination using an ASV table with the same parameters, and this dataset converged in 3 dimensions with an overall stress of 0.160.

2.4 Results

2.4.1 Presence of eukaryotes

To assess seasonal changes in the eukaryotic community of BML, water samples were taken monthly from the lake during the ice-free period (May to October). The V4 region of the 18S ribosomal rRNA gene was amplified, and amplicons were sequenced. Each sample ranged between 573 and 255,365 reads, with a total of approximately 5.5 million reads of 18S V4 region sequence. After clustering all the V4 sequences at 97% identity as a single dataset, filtering out nontarget (i.e. organellar and bacterial sequences), and performing quality control such as read trimming and chimera removal, 565 microbial eukaryotic OTUs remained for analysis. We explicitly removed the OTUs representing multicellular organisms from the analysis as our pipeline is optimised for clustering and analysis of microorganisms and therefore would not necessarily produce accurate diversity estimates for these organisms. The discarded OTUs mapped to nematodes and other helminths (Metazoa) or model plant species such as crops (Embryophyta). The remaining sequences represent considerably greater OTU diversity than was detected in a previous study of active tailings ponds, including West In-Pit (the tailings pond that was converted to BML in 2012) and Mildred Lake Settling Basin[149]. When the 565 OTUs were classified using the PR2 database, we observed 271 distinct classifications. The majority of these classifications were down to the genus level. However, just under half of these classifications (111 OTUs) were unclassifiable to this taxonomic resolution or lower using this technique (Figure S2.2). The majority of all sequence reads in the dataset were assigned to phyla in which all known taxa are heterotrophic (Figure 2.1). The sum of the relative abundances of OTUs from these phyla was 67%. OTUs belonging to phyla in which all known taxa are photosynthetic accounted for only 6% of the relative abundance, while OTUs classified into groups containing both heterotrophic and photosynthetic species accounted for 27% of the relative abundance (Figure 2.1, Table 2.1). Substantial abundance differences were observed between heterotrophic groups. For example, 34% of the sequence reads from all samples were classified into fungal OTUs, while 13% were classified into ciliate OTUs and 26% into cercozoan OTUs (Figure 2.1). However, OTUs from other majority heterotroph supergroups such as Amoebozoa and Excavata together accounted for less than 1% of the relative abundance (Figure 2.1 and supplementary files). These results, derived from BLAST comparison to the SILVA and PR2 databases, were supported by a pplacer tree

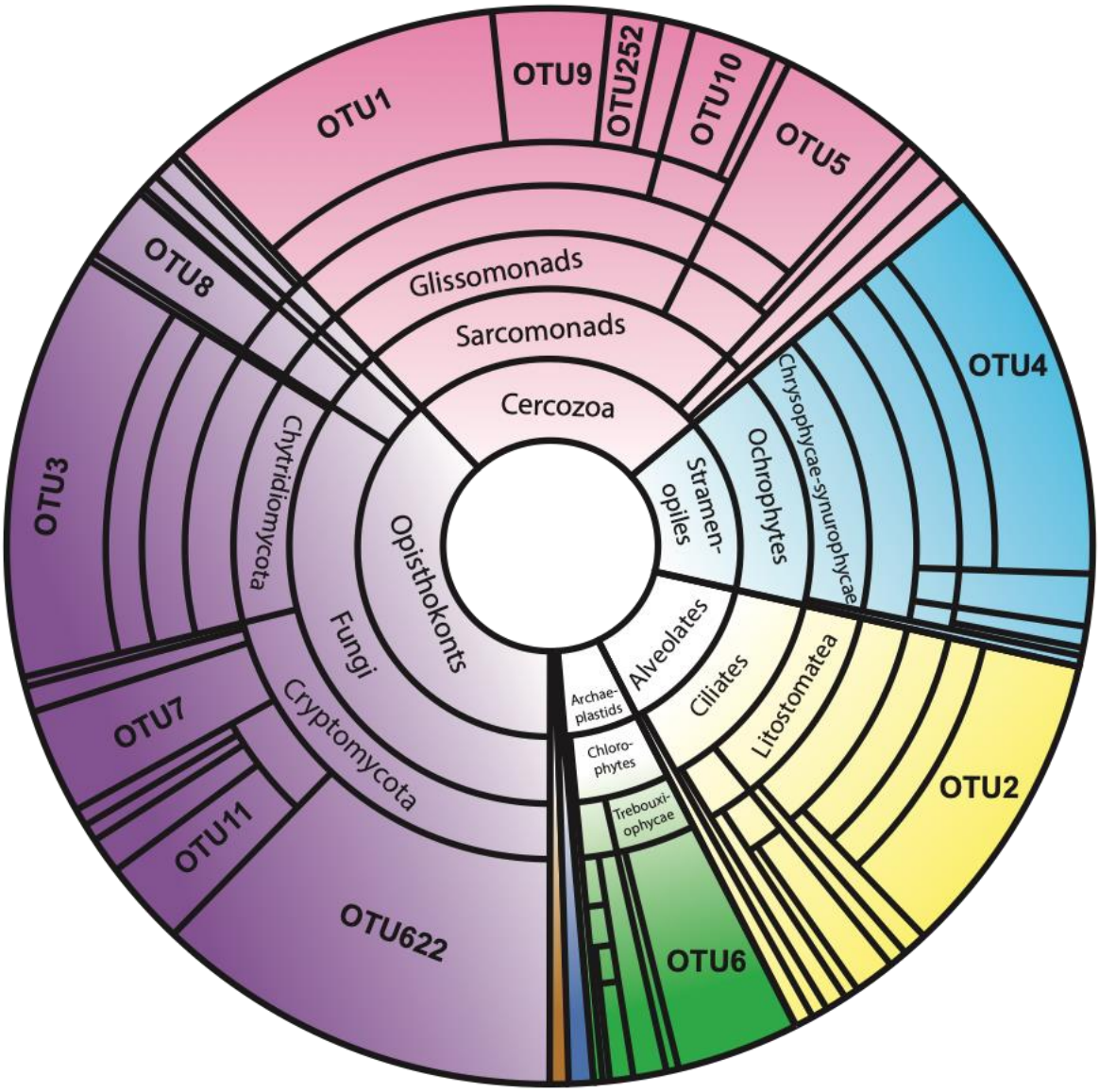


Figure 2.1: KronaPlot of overall eukaryotic diversity in the summer of 2015. Classifications of all microbial OTUs from Base Mine Lake based upon comparison to reference databases. The OTU positions are based on BLAST comparisons to reference databases (GenBank, PR2 and SILVA). In this plot, the proportion of the total abundance of the OTU in all samples is indicated by the size of the wedge. Wedges with fewer classification levels were less well-classified by BLAST comparisons, and no confident classification could be made at a higher taxonomic resolution.

showing a phylogenetic placement of OTU sequences in a backbone tree broadly corresponding to the classifications assigned by BLAST to SILVA, PR2 and GenBank (Figure S2.3).

2.4.2 Heterotroph groups show extensive novel diversity

To better determine the classification of the OTUs, we created phylogenetic trees of each of these groups, including backbone sequences from reference databases and full sequences of close hits from the GenBank database. The groups we identified for further analyses were Amoebozoa, Cercozoa, Ciliophora, Fungi and Excavata, as three of these groups (Fungi, Ciliophora, and Cercozoa) cover a large proportion of microbial heterotroph diversity. One of the OTUs in the Amoebozoa tree, OTU243, consistently grouped with the metazoan sequences in the outgroup rather than with any Amoebozoa sequences—even the one that was retrieved as its closest hit in the GenBank database (Figure S2.4). When investigating this OTU more closely, we discovered it was related to *Syssomonas multiformis,* an organism that was first entered into the GenBank database in July 2017 as part of a newly discovered clade of holozoans.

We used the phylogenetic trees of heterotrophs to determine whether we could identify additional novel diversity in Base Mine Lake. In the context of this study, we defined 'novel diversity' as groups of OTUs that were positioned together to the exclusion of every reference sequence, whether from the reference backbone or GenBank (including the top scoring match from the nr database), with high bootstrap and MrBayes support (higher than 0.75 posterior probability and a bootstrap value of higher than 75). The amount of novel diversity varies between groups and does not appear to be related to the number of OTUs found within that group. In Cercozoa, novel diversity accounts for a large abundance of the OTUs (26%) while also having the greatest proportion of those OTUs classified as novel (33 out of 91) (Figure 2.2); contrarily, in ciliates, every OTU is closely related to another known species (Figure 2.3). It is also notable that the cercozoan OTUs that make up the majority of the diversity (OTUs 1, 5, 9, 10 and 252), are grouped within the Glissomonada (Figure 2.2) with robust bootstrap support; though the resolution of these OTUs within the Glissomonada is less clear, the higher level taxonomic classification is likely accurate due to the high support values. This cercozoan order contains mostly heterotrophic gliding zooflagellates. It is noted as being particularly lineage-rich, and the majority of these lineages have not been cultured or studied.

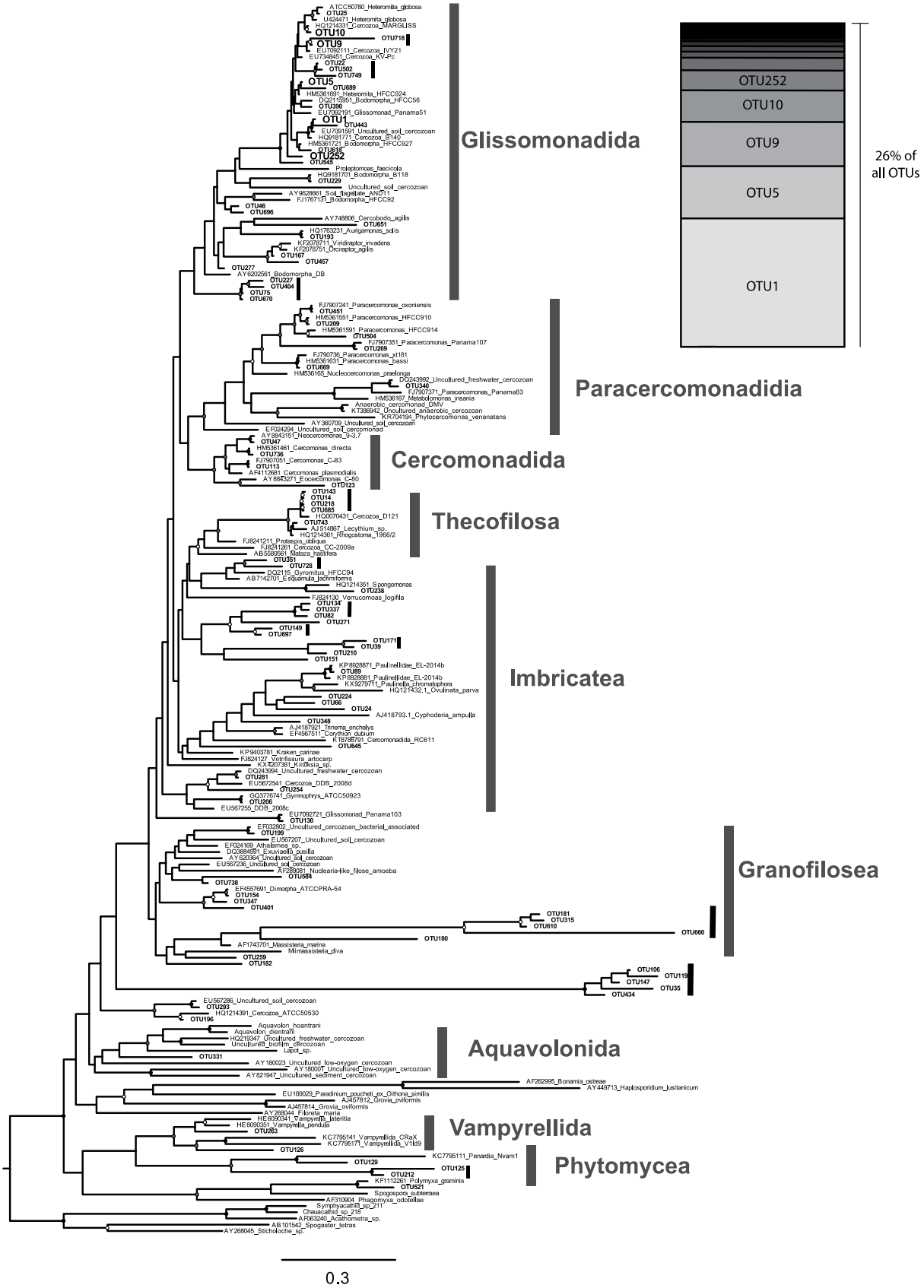

Glissomonadida
Paracercomonadidia
Cercomonadida
Thecofilosa
Imbricatea
Granofilosea
Aquavolonida
Vampyrellida
Phytomycea
OTU252
OTU10
OTU9
OTU5
OTU1
26% of all OTUs
0.3

Figure 2.2: Phylogeny of Cercozoa.
The OTUs shown in this figure make up a total of 26% of all 18S rRNA gene sequence reads from Base Mine Lake, and the majority of abundant OTUs (depicted in larger font) are found in the Glissomonada clade. There are also notable long-branching clades within and adjacent to the Grandofilosea, that contain only Base Mine Lake sequences in strongly supported clades. For this and all subsequent figures displaying phylogenetic analyses, the phylogeny is based on MrBayes and RAxML trees, mapped onto the MrBayes topology. Node support is indicated by the circles on each node: black indicates 1 / 100 MrBayes / RAxML support, grey indicates 0.9 / 90% or higher, and white indicates 0.75 / 75 support. High-level taxonomic groupings are indicated with grey bars. Black bars indicate OTUs defined as novel diversity by the methods described in the paper. Particularly abundant OTUs are indicated in a larger font on the trees, and in the bar graph to the right of the phylogeny as a proportion of total OTUs.

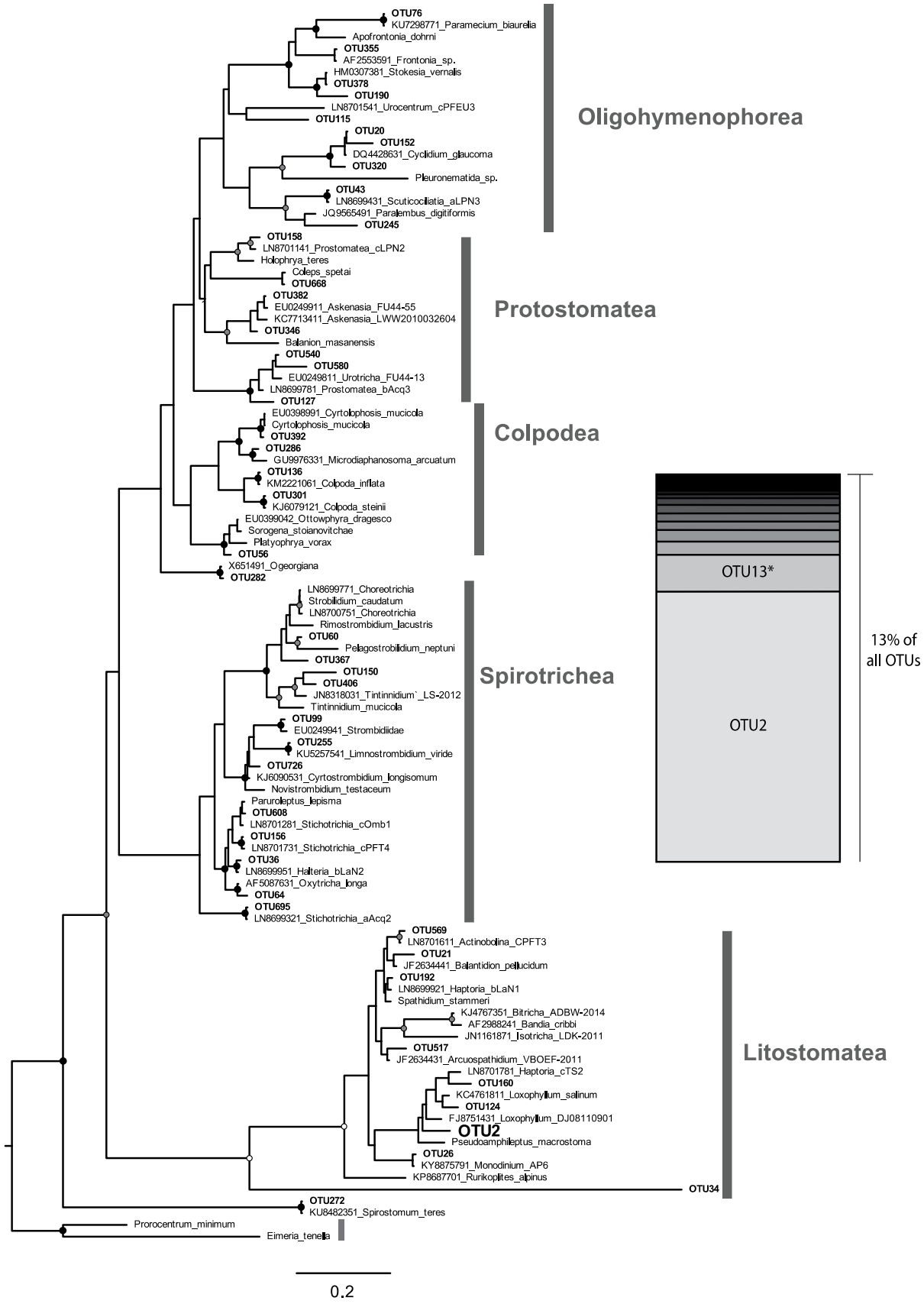

OTU76
KU7298771_Paramecium_biaurelia
Apofrontonia_dohrni
OTU355
AF2553591_Frontonia_sp.
HM0307381_Stokesia_vernalis
OTU378
OTU190
LN8701541_Urocentrum_cPFEU3
OTU115
OTU20
OTU152
DQ4428631_Cyclidium_glaucoma
OTU320
Pleuronematida_sp.
OTU43
LN8699431_Scuticociliatia_aLPN3
JQ9565491_Paralembus_digitiformis
OTU245
Oligohymenophorea
OTU158
LN8701141_Prostomatea_cLPN2
Holophrya_teres
Coleps_spetai
OTU668
OTU382
EU0249911_Askenasia_FU44-55
KC7713411_Askenasia_LWW2010032604
OTU346
Balanion_masanensis
OTU540
OTU580
EU0249811_Urotricha_FU44-13
LN8699781_Prostomatea_bAcq3
OTU127
Protostomatea
EU0398991_Cyrtolophosis_mucicola
Cyrtolophosis_mucicola
OTU392
OTU286
GU9976331_Microdiaphanosoma_arcuatum
OTU136
KM2221061_Colpoda_inflata
OTU301
KJ6079121_Colpoda_steinii
EU0399042_Ottowphyra_dragesco
Sorogena_stoianovitchae
Platyophrya_vorax
OTU56
X651491_Ogeorgiana
OTU282
Colpodea
LN8699771_Choreotrichia
Strobilidium_caudatum
LN8700751_Choreotrichia
Rimostrombidium_lacustris
OTU60
Pelagostrobilidium_neptuni
OTU367
OTU150
OTU406
JN8318031_Tintinnidium`_LS-2012
Tintinnidium_mucicola
OTU99
EU0249941_Strombidiidae
OTU255
KU5257541_Limnostrombidium_viride
OTU726
KJ6090531_Cyrtostrombidium_longisomum
Novistrombidium_testaceum
Paruroleptus_lepisma
OTU608
LN8701281_Stichotrichia_cOmb1
OTU156
LN8701731_Stichotrichia_cPFT4
OTU36
LN8699951_Halteria_bLaN2
AF5087631_Oxytricha_longa
OTU64
OTU695
LN8699321_Stichotrichia_aAcq2
Spirotrichea
OTU13*
OTU2
13% of
all OTUs
OTU569
LN8701611_Actinobolina_CPFT3
OTU21
JF2634441_Balantidion_pellucidum
OTU192
LN8699921_Haptoria_bLaN1
Spathidium_stammeri
KJ4767351_Bitricha_ADBW-2014
AF2988241_Bandia_cribbi
JN1161871_Isotricha_LDK-2011
OTU517
JF2634431_Arcuospathidium_VBOEF-2011
LN8701781_Haptoria_cTS2
OTU160
KC4761811_Loxophyllum_salinum
OTU124
FJ8751431_Loxophyllum_DJ08110901
OTU2
Pseudoamphileptus_macrostoma
OTU26
KY8875791_Monodinium_AP6
KP8687701_Rurikoplites_alpinus
OTU34
Litostomatea
OTU272
KU8482351_Spirostomum_teres
Prorocentrum_minimum
Eimeria_tenella
0.2

Figure 2.3: Phylogeny of Ciliophora.
The most abundant ciliate OTU is OTU2, grouped within the Litostomatea.

Individual OTUs (representing a single unit of genus-to-species-level diversity) also made up a considerable proportion of the total abundance. Particularly notable OTUs are OTU3 (Microsporida), OTU622 (Microsporidia), OTU4 (Chrysophycae-Synurophycae) and OTU1 (Glissomonada), which made up 13, 12, 12, and 10 percent of the overall relative abundance respectively. All of the OTUs that accounted for at least 10% of the overall relative abundance were classified as heterotrophic; though OTU4 was from the class Synurophycae, comparison to BLAST databases showed it was most similar to the heterotrophic *Spumella*-like flagellates. Both OTU622 and OTU3 were classified as fungi (Figure 2.1, Table 2.1). These OTUs were classified only down to a phylum level by BLAST comparisons (Figure 2.1, Figure S2.2). The phylogenetic classification showed that both OTU3 and OTU622 were found in the same small, deeply branching clade of Microsporidia containing only OTUs found in this study and V4 regions from unknown or uncultured organisms, with high bootstrap and Bayesian support (Figure 2.4). The abundance of these high-abundance OTUs varied dramatically from month to month (Figure 2.5); for example, OTU622 was barely detectable in the early summer months before becoming the most abundant overall OTU in September, comprising over 35% of the total reads (Figure 2.5).

2.4.3 The majority of OTU variation appears to be determined by month of sampling

Though the sample set was too small to make any major ecological conclusions, we used multivariate statistics to determine what percentage of the variation could be explained by the depth the sample was taken, the platform position within Base Mine Lake, and the month the sample was taken. We carried out two separate rarefaction analyses using PERMANOVA, and in both cases most of the variation (41.6% and 39.4%, respectively) was explained by month, with a small percentage of the variation (2.5% and 2.4%, respectively) explained by sampling platform. This clustering was also observed in an NMDS plot produced using the metaMDS function in vegan ($k = 2$ with a stress of 0.224), where the samples were strongly organized by month (Figure 2.6). Minor contributors to the sample diversity included the platform and the depth, but these factors explained less than 3% of the variation. To ensure that these results were not an artefact of the 97% clustering threshold (which may mask some of the species-level diversity), we carried out the same NMDS analysis with denoised ASV data. The same trends in month-to-month turnover were observed, while no trends in depth or month were seen. Further analyses to determine whether novel OTUs were associated with a particular aspect of the metadata (such as depth or month) did

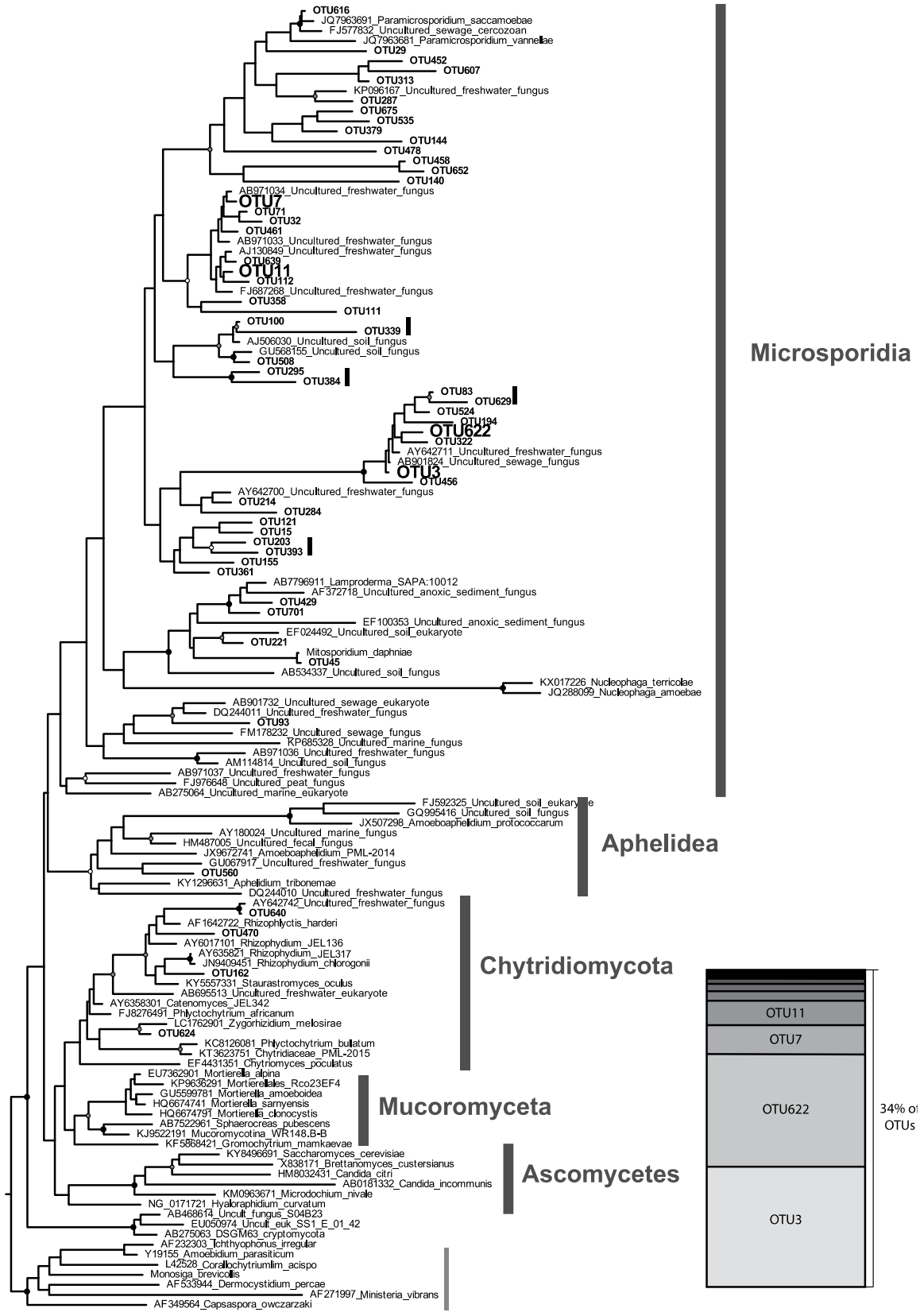

OTU616
JQ7963691_Paramicrosporidium_saccamoebae
FJ577832_Uncultured_sewage_cercozoan
JQ7963681_Paramicrosporidium_vannellae
OTU29
OTU452
OTU607
OTU313
KP096167_Uncultured_freshwater_fungus
OTU287
OTU675
OTU535
OTU379
OTU144
OTU478
OTU458
OTU652
OTU140
AB971034_Uncultured_freshwater_fungus
OTU7
OTU71
OTU32
OTU461
AB971033_Uncultured_freshwater_fungus
AJ130849_Uncultured_freshwater_fungus
OTU638
OTU11
OTU112
FJ687268_Uncultured_freshwater_fungus
OTU358
OTU111
OTU100
OTU339
AJ506030_Uncultured_soil_fungus
GU568155_Uncultured_soil_fungus
OTU508
OTU295
OTU384
OTU83
OTU629
OTU524
OTU194
OTU622
OTU322
AY642711_Uncultured_freshwater_fungus
AB901824_Uncultured_sewage_fungus
OTU3
OTU456
AY642700_Uncultured_freshwater_fungus
OTU214
OTU284
OTU121
OTU15
OTU203
OTU393
OTU155
OTU361
AB7796911_Lamproderma_SAPA:10012
AF372718_Uncultured_anoxic_sediment_fungus
OTU429
OTU701
EF100353_Uncultured_anoxic_sediment_fungus
EF024492_Uncultured_soil_eukaryote
OTU221
Mitosporidium_daphniae
OTU45
AB534337_Uncultured_soil_fungus
KX017226_Nucleophaga_terricolae
JQ288099_Nucleophaga_amoebae
AB901732_Uncultured_sewage_eukaryote
DQ244011_Uncultured_freshwater_fungus
OTU93
FM178232_Uncultured_sewage_fungus
KP685328_Uncultured_marine_fungus
AB971036_Uncultured_freshwater_fungus
AM114814_Uncultured_soil_fungus
AB971037_Uncultured_freshwater_fungus
FJ976648_Uncultured_peat_fungus
AB275064_Uncultured_marine_eukaryote
FJ592325_Uncultured_soil_eukaryote
GQ995416_Uncultured_soil_fungus
JX507298_Amoeboaphelidium_protococcarum
AY180024_Uncultured_marine_fungus
HM487005_Uncultured_fecal_fungus
JX9672741_Amoeboaphelidium_PML-2014
GU067917_Uncultured_freshwater_fungus
OTU560
KY1296631_Aphelidium_tribonemae
DQ244010_Uncultured_freshwater_fungus
AY642742_Uncultured_freshwater_fungus
OTU640
AF1642722_Rhizophlyctis_harderi
OTU470
AY6017101_Rhizophydium_JEL136
AY635821_Rhizophydium_JEL317
JN9409451_Rhizophydium_chlorogonii
OTU162
KY5557331_Staurastromyces_oculus
AB695513_Uncultured_freshwater_eukaryote
AY6358301_Catenomyces_JEL342
FJ8276491_Phlyctochytrium_africanum
LC1762901_Zygorhizidium_melosirae
OTU624
KC8126081_Phlyctochytrium_bullatum
KT3623751_Chytridiaceae_PML-2015
EF4431351_Chytriomyces_poculatus
EU7362901_Mortierella_alpina
KP9636291_Mortierellales_Rco23EF4
GU5599781_Mortierella_amoeboidea
HQ6674741_Mortierella_sarnyensis
HQ6674791_Mortierella_clonocystis
AB7522961_Sphaerocreas_pubescens
KJ9522191_Mucoromycotina_WR148.B-B
KF5868421_Gromochytrium_mamkaevae
KY8496691_Saccharomyces_cerevisiae
X838171_Brettanomyces_custersianus
HM8032431_Candida_citri
AB0181332_Candida_incommunis
KM0963671_Microdochium_nivale
NG_0171721_Hyaloraphidium_curvatum
AB468614_Uncult_fungus_S04B23
EU050974_Uncult_euk_SS1_E_01_42
AB275063_DSGM63_cryptomycota
AF232303_Ichthyophonus_irregular
Y19155_Amoebidium_parasiticum
L42528_Corallochytriumlim_acispo
Monosiga_brevicollis
AF533944_Dermocystidium_percae
AF271997_Ministeria_vibrans
AF349564_Capsaspora_owczarzaki
Microsporidia
Aphelidea
Chytridiomycota
Mucoromycota
Ascomycetes
OTU11
OTU7
OTU622
OTU3
34% of OTUs
0.08

Figure 2.4 Phylogeny of Fungi.
There is a long-branching, well-supported clade containing only Base Mine Lake and uncultured fungal sequences, which also contains two of the most abundant OTUs within the overall dataset, OTU3 and OTU622. This group represent potentially novel diversity.

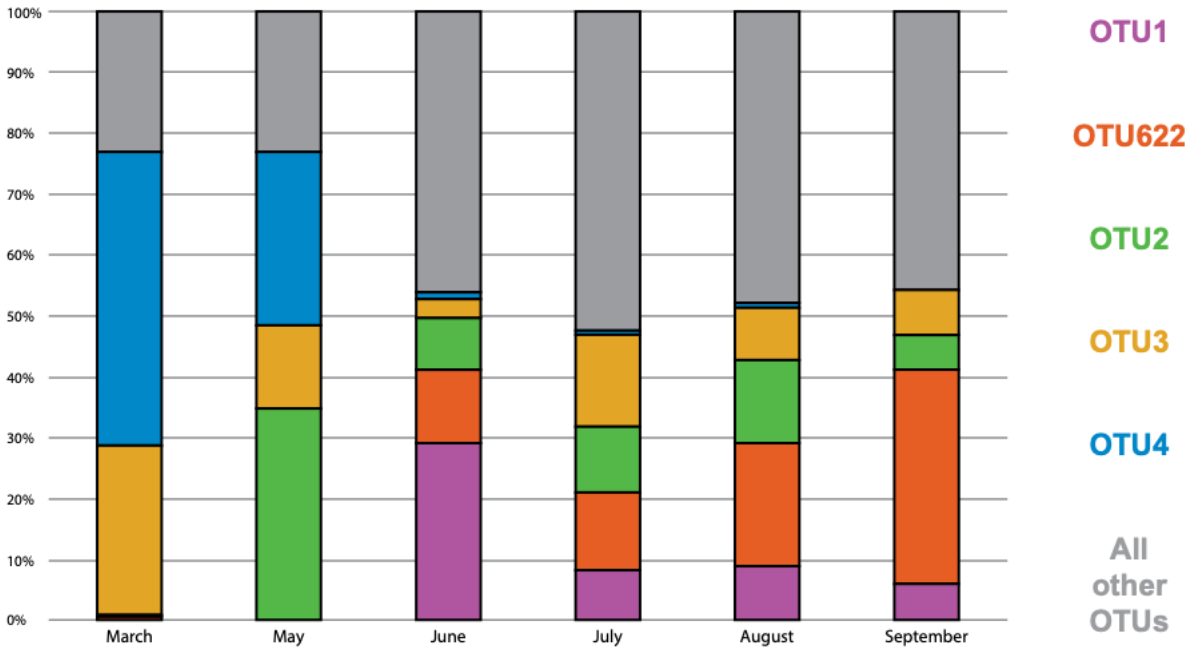


Figure 2.5: Distribution of highly abundant OTUs across the ice-free period of 2015.
OTUs are represented as a proportion of the overall 18S rRNA gene reads for each month, with their taxonomic affiliation to the lowest confidently assigned level from both comparison to reference databases and phylogenetic analyses indicated. The grey section in the bar chart represents all 561 other OTUs. Taxonomic affiliations: OTU1, Glissmonada. OTU2, Litostomatea. OTU3, Microsporidia. OTU4, Spumella. OTU622, Microsporidia.

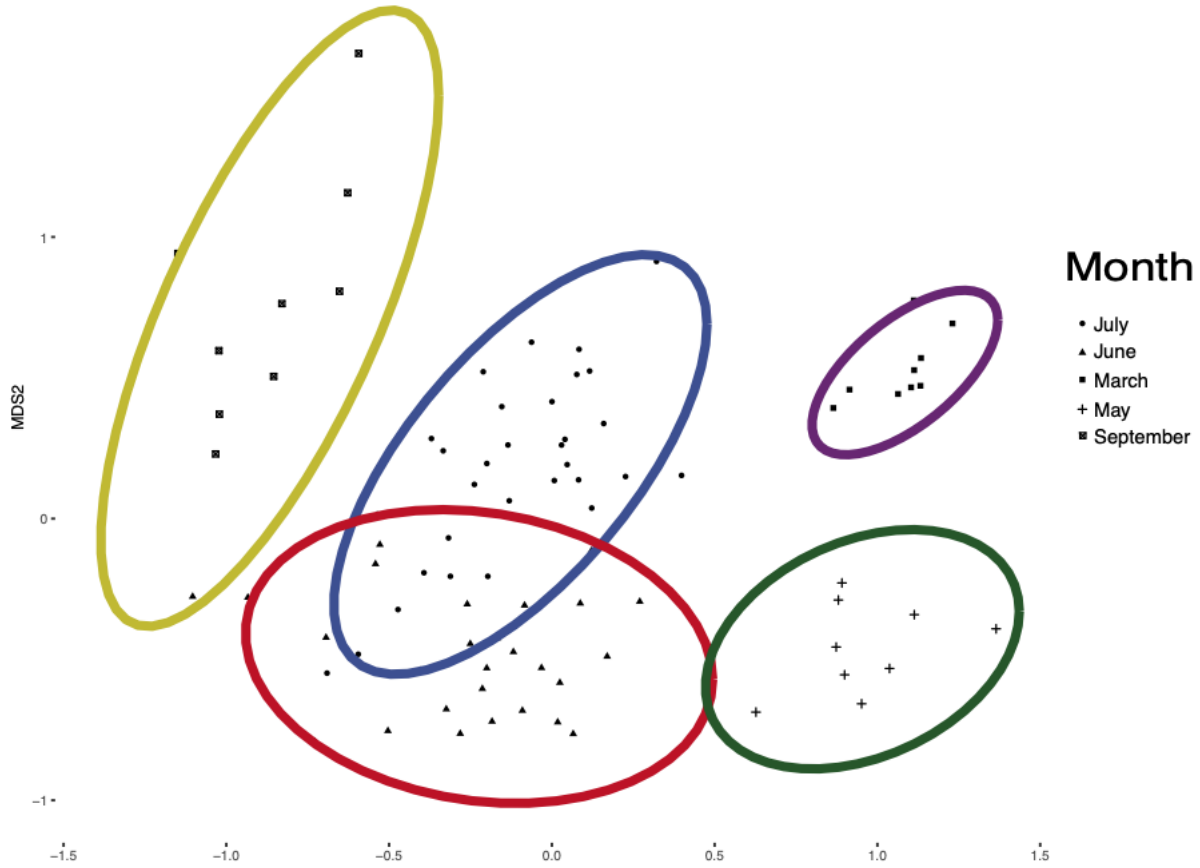


Figure 2.6: Ordination of OTUs from Base Mine Lake.
NMDS analysis of BML OTU abundances using Bray-Curtis distance and default metaMDS analysis in R(vegan) of the total OTU abundance table by sample for all samples of each species taken from Base Mine Lake. Month is indicated by the shape of the data point for each sample, and ellipses show the clustering of the points for each month.

not show any statistically significant patterns associated with the novel diversity, and OTUs associated with novel diversity were found in all heterotrophic phyla we studied at all months and at all depths.

2.5 Discussion

2.5.1 Eukaryote presence

Eukaryotic microbial community structures are understudied compared to bacterial communities, particularly in the important context of reclamation microbiology[192–194]. In Base Mine Lake, after three years of reclamation efforts, OTUs detected included members of all eukaryotic supergroups.

Further, while the majority of the OTUs were from heterotrophic taxa, photosynthetic taxa were also present (Figure 2.1, Figure S2.3). Abundant taxa that were well represented in BML, such as short-branched Microsporidia and other microbial fungi, were also abundant in both reclaimed and undisturbed Albertan soils, suggesting that there was likely a repopulation of Base Mine Lake from the local environment[149]. Since microbial community composition is often highly heterogenous even within a relatively uniform environment, it is important that local diversity is taken into account when evaluating reclamation efforts. There was also a great deal of month-on-month turnover of OTUs in Base Mine Lake, and the month the sample was collected explained the majority of the variation (Figure 2.5, 2.6).

The monthly turnover was substantial, particularly considering the organisms were heterotrophic. While turnover of photosynthetic taxa is common in seasonal lakes in summer due to the dynamics of algal blooms[195], the extreme changes in abundance of heterotrophs (and particularly the domination of fungi in the later summer months) is not usually noted in studies that use presence of taxon-distinctive photosynthetic pigments to identify the of lake microeukaryotes[158]. We also did not observe any stratification of the microbial community by depth, which is highly unusual for a lake environment, especially considering that thermal stratification has been observed in Base Mine Lake[174] is consistent with other boreal lakes. Thermal stratification is known to be a cause of changes in microbial community composition[195], and it will be interesting to see whether this changes over time.

2.5.2 High relative abundance of heterotrophs

The most notable feature of the OTU species-level classification was the number of OTUs classified as heterotrophs within the samples, which were taken from the spring to fall and therefore at peak productive time for phototrophs in seasonally ice-covered lakes[196] (Figure 2.1, Figure S2.3). There are several factors that could account for the apparent imbalance between the relative abundance of heterotrophs and photosynthesisers. Endogenous bacteria may be acting as a food source for these heterotrophs, many of which are bacteriovores; ciliates commonly feed on bacteria, and fungi are often detrivores that will consume a variety of dead cellular materials[197,198]. Soils in the Albertan boreal forest close to the Base Mine Lake site also appear to have an unusually high abundance of fungi, particularly short-branch Microsporidia, which are known to include parasites of amoebae, ciliates, and planktonic crustaceans[199]. The high abundance of fungi in Base Mine Lake may simply be a reflection of the local environment[198]. Also, other high-throughput sequencing studies have determined that small heterotrophs, particularly parasitic taxa, are much more abundant globally in soils than was previously thought[200]. The rapid month-to-month turnover in these environments, particularly of the most abundant OTUs, suggested that the eukaryotic community was changing over the summer season (Figure 2.5). The extent to which this was caused by the seasonal variation remains to be determined. We also note that some of the heterotrophs present in this environment in the summer of 2015 may be transient members of this community. As reclamation progresses, we will be able to determine which members of the community are recurrent from season to season and which are transient, as has been done in other long term ecological studies [201]

2.5.3 Low relative abundance of photosynthesisers

While photosynthetic taxa were detected both in the reference database classification and the pplacer tree, they accounted for only a small percentage of the abundance of classified OTUs (Figure 2.1, Figure S2.3). There are several technical biases that could partially account for this finding. For example, certain taxa are known to be relatively poorly amplified by universal V4 primers[202]. These include the mostly heterotroph taxa Amoebozoa and Excavata, consistent with the low numbers that we observed (Figure S2.4, Figure S2.5). Euglenoids, a mostly photosynthetic excavate clade, have been shown to be highly represented in other mine tailings-contaminated water due to their high metal tolerance[166]. However, we only saw one euglenoid sequence with

low abundance (OTU525, Figure S2.5), and this may be an example of PCR bias in our dataset. However, these primers have not been noted to fail to amplify phototrophs apart from the group Prymnesiophycae [203]. The likely biological explanation for the relatively low abundance of photosynthesisers is the turbidity of the water cap causing low light penetration into Base Mine Lake. The measured turbidity during the summer of 2015 was high, particularly early in the spring when algal blooms usually occur[174]. Turbidity was also strongest over the summer at Platform 3, which was the most different to the other two platforms in terms of community composition; this also supports a scenario where turbidity is the major driving force behind the microbial eukaryote community in BML[174]. However, Tedford et al. (2018)[204] also notes that Platform 3 is closest to the freshwater inflows to BML, which may also affect eukaryotic community composition. There are also additional biological factors that could be at play; there is evidence that shelled or scaled algae, such as diatoms, are less able to survive in environments containing hydrocarbons as the increased surface area of the scales causes proportionately increased exposure when compared to the cell size[158]. Also, most studies of phytoplankton interactions with hydrocarbons describe marine environments where the environment is naturally saline. Salinity of BML was elevated compared to its freshwater inputs[205]. There may be an interacting effect between salinity and hydrocarbon presence on the phytoplankton species in the local environment—this could be an interesting area for future research.

2.5.4 Abundant taxa

Three taxonomic groups appeared to be unusually highly represented in these samples: one OTU within the class Litostomatea (ciliates), multiple OTUs within the order Glissomonadida (cercozoans), and multiple OTUs within the group Microsporidia (fungi) (Figure 2.2-2.4). Litostomatea is a large and diverse class with a variety of lifestyles; the majority are free-living predators of bacteria and smaller protists, and some are anaerobic commensals[206]. Oil sands tailings ponds contain an extremely abundant bacterial community[207], suggesting that bacterivores could thrive there. That the majority of Litostomatea present were from a cluster of species-level diversity represented by OTU2 suggested that this may represent an organism with very high tolerance to hydrocarbon-contaminated environments (Figure 2.3). Glissomonads were the group with the most OTUs classified as novel diversity from the phylogenetic analysis (Figure 2.2); the organisms themselves are relatively morphologically similar, but recent studies of their genetic

diversity show them to be an extremely diverse group[208,209]. Similar to the Litostomatea, all known glissomonads are heterotrophs and could therefore potentially find a niche feeding on the hydrocarbon-degrading bacteria (Figure 2.2). The presence of Microsporidia is more puzzling since these organisms are traditionally understood to be parasitic and should be less represented in an environment which has relatively few trophic levels[210]. However, more recent analysis of the Microsporidian group shows that it is far more diverse and wide-ranging than was previously thought and contains many species that were originally considered to be Crypto-/Rozellomycota, which are suspected to contain a large diversity of protist-parasitizing lineages[199]. It is possible that the microsporidia in BML are parasites of other protists in the lake. If potential metazoan microsporidian hosts are absent from sites such as BML, these habitats would be ideal sources of cells to determine whether the short-branched microsporidian diversity detected in the environment are all parasites of other protist, or whether some are free-living [199]. Other fungi were relatively very poorly represented: only four chytrid OTUs were detected, in contrast to many freshwater plankton diversity studies where they are often more diverse and abundant, and Dikarya were only present at low levels (Figure 2.1, 2.4). Therefore, parasitic fungi appear to be relatively more abundant and diverse in BML than saprotrophs. Interestingly, the eukaryvorous cercozoan flagellates Aquavolonida have recently been shown to be diverse and very frequently detected in freshwater plankton[199], but only one OTU with this classification was detected in BML. This is in line with the relatively low levels of eukaryotic diversity detected in BML, where the communities were dominated by a relatively small number of abundant OTUs (Figure 2.1). OTU4, another abundant OTU most visible in the early summer months, belongs to the *Spumella*-like microflagellates[211]. *Spumella*-like microflagellates are functionally similar to other heterotrophic flagellates such as Litostomatea and Glissomonadida, and likely fill a similar ecological niche in environments in which they are present [211].

The relative abundance of single-celled fungi, most of which were Microsporidia, in later months (June, September) was particularly striking. In the boreal forest of northern Alberta, coniferous pollen matures and disperses in late May/early June [212]. This pollen, which causes thick, matted layers on many bodies of water in the region, is filled with microbial fungi, mostly chytrids, which could then disperse into the water column[213]. Our results and this seasonal event are correlated, but we cannot draw a causal link between them at this time. Recent high-throughput sequencing

studies of soil environments in northern Alberta have also shown that the presence of microheterotrophs such as Fungi and Cercozoa represent a considerable proportion of the diversity in high-throughput sequencing studies, which is consistent with our results[200] (Figure 2.1 and S2).

2.5.5 Novel diversity

The clades we identified as 'novel diversity' were defined as such due to their grouping with other BML-derived OTUs to the exclusion of every other published 18S RNA gene sequence. This means either that these OTUs and the surrounding clade are representatives of known taxonomic groups but have diverged so much they cannot be conclusively classified as those groups by V4 18S rRNA gene sequencing, or that they potentially represent previously unknown protist diversity. The relatively poor taxonomic resolution of a considerable proportion of OTU sequences demonstrates substantial levels of novel diversity within these samples, confirmed by BLASTn searches against GenBank and phylogenetic analyses, which also showed that some of this novel diversity is long-branched (Figure 2.2-2.4, S2.3, S2.4). OTUs associated with this novel diversity were not merely relegated to the 'rare biosphere' but represented some of the most abundant OTUs within the samples. Northern Alberta is characterized by many natural hydrocarbon seeps and exposures, so it is possible that these OTUs are also abundant in natural habitats near BML and act as a seed bank[214]. Heterotrophic cercozoans showed high levels of novel diversity, with 34% of the OTUs grouping with other BML OTUs to the exclusion of any reference sequences (Figure 2.2). Notably, the relatively abundant glissomonad OTU9 is in this grouping and represents 3% of the sequencing reads (Figure 2.2). These small heterotrophs are relatively understudied compared to the larger and more morphologically diverse ciliates, and this lack of reference information may explain why so many of the cercozoan OTUs are novel[208,209] even though group-specific PCR and culturing studies have previously revealed high levels of diversity within several cercozoan clades. Single-celled eukaryotes have also shown remarkable adaptability to challenging environments, such as ciliates in both industrially contaminated environments and urban wastewater [163,215]. It is also worth noting that a single OTU is not necessarily a single species; in this study, an OTU represents V4 18S rRNA regions at least 97% similar to each other, which may represent a single described species or multiple described species. A single described species may also be split over multiple OTUs[216].

2.5.6 Sample heterogeneity

The metadata we had available for these samples was depth, platform, and month of sampling. NMDS analysis of sample variation shows that the most important of these to explain sample variation was the month of sampling (2.6). There were no significant community differences when testing either the depth or platform variables. This is unusual for a dimictic boreal lake in the summer where thermal stratification results in gradients of temperature and oxygen penetration and algal bloom turnover usually results in extreme community stratification by depth. In such lakes high photosynthetic diversity often occurs in the oxygen-rich surface waters and heterotrophs and detrivores in the relatively anoxic deep waters[196]. This lack of depth stratification could be explained by water turbidity; since when these samples were taken the lake was still in the process of settling, photosynthesis would have been inhibited. The low abundance of photosynthesisers, and therefore algal blooms, which are a vital part of lake turnover in summers, may be a reason for the unusual abundance and turnover of heterotrophs (Figure 2.5 and 2.6) which we hope to explore further in later years of sampling.

2.6 Conclusions

We have characterized a complex and highly distinct community of microbial eukaryotes in Base Mine Lake in the summer of 2015. This community was dynamic and showed clear changes in diversity and abundance of taxa structured over time. Our study most notably identified novel diversity within the heterotroph community, speculatively due to a combination of the abundant food sources for bacteria and bacteriovores and the unique environmental challenges providing ecological pressure against photosynthetic algae. This work provides a solid basis to explore the biology of new and diverse heterotrophic lineages as well as their potential role in the important process of reclamation of hydrocarbon-impacted environments.

2.7 Afterword

This study began with sample collection in the summer of 2015, and sequencing of the samples was carried out as part of a larger industrial collaboration at the University of Calgary in 2016. OTU clustering and preliminary analysis of the data was carried out in 2017, with additional ecological analysis in 2018 before final manuscript preparation, review, and publication in 2019.

The latter half of the 2010s was a period of rapid acceleration, both in sequencing technology and in refinement of eDNA analysis techniques for microbial eukaryotes[35,65]. Accordingly, there are now some inconsistencies between the analysis completed in this study and the current best practices for microbial community analysis.

Firstly, it is noted in this paper that there is a relatively low abundance of OTUs from the Excavata or the Amoebozoa. While this may be a result of a low abundance of species from these groups in the samples, it has also been shown through extensive analysis of mock communities and other benchmarking studies that the Stoeck et al. (2010) primers used in this study are less efficient at amplifying the Amoebozoa or Excavata 18S rRNA V4 region than other taxa[202]. Though this is discussed in Section 2.5.3, it is worth noting that since the publication of this study, a new user guide of best practices for eDNA assessment of the microbial eukaryote community has been published that describes sets of primers that can be used together to maximise the effectiveness of diversity assessments[35]. An additional line of evidence that Excavata may be more strongly represented than this study implies has also been uncovered; analyses of the phototrophic component of the BML community post-2016 has uncovered a substantial euglenid community which is not represented here. However, this may also be explained by the relative lack of photosynthesisers overall.

Secondly, the clustering threshold used for production of OTUs in this study is 97%, which more likely equates to genus-level diversity in microbial eukaryotes[70]. As discussed in section 1.2, current opinion suggests that there is no single clustering threshold that would suffice to explain all of eukaryotic diversity; most current studies suggest a clustering threshold of at least 99%, with many researchers using zero-radius OTUs or ASVs[216]. Despite this, we have chosen to maintain a 97% clustering threshold for phylogenetic analysis in both Chapter 2 and Chapter 3. The reasoning for this, including benchmarking data, is explained in detail in Appendix I; briefly, 97% is a threshold which allows us to capture trends in diversity while still clustering the OTUs to a depth manageable for phylogenetic community assessment. Additionally, there is now substantially more sequence data available in public databases for comparison, mostly due to new curation efforts such as the development of the EukBank database[21]. This is also noted in section 2.4.2; an OTU that in the first round of classification was classified as a divergent Amoebozoa was in fact

most closely related to *Syssomonas multiformis*, a basal holozoan not identified until two years after the initial samples were taken[217]. No doubt repeating comparison of these samples to current databases would change the number of sequences defined as novel diversity by this study. However, the substantial number of OTUs which are defined as novel diversity indicates this is likely a good source for unknown eukaryotic diversity. Since the publication of this study, BML has proven itself to be a good resource for bioprospecting: a novel type of SAR11-infecting phage was identified via metagenome-assembled genomics of a sample from BML's water cap[218]. A novel mixotrophic bacteria associated with nitrogen cycling has been identified from the tailings/water interface[219]; and novel genes encoding enzymes that are able to oxidate short-chain alkanes, methane and ammonia were identified in bacterial genomes from BML[122].

This chapter, which describes the eukaryotic community in Base Mine Lake across the summer of 2015, describes a diverse and apparently thriving community of eukaryotes in this reclamation environment. Studies of the geophysics, bacterial community, and biogeochemistry of BML now include many multiyear studies, recognising the importance of understanding the changes in a reclamation environment over time[119,120,220]. Accordingly, the next chapter of this thesis describes a multiyear study of the eukaryotic community of BML over time and the community's interactions with biotic and abiotic variables and environmental disturbances.

Chapter 3
BASE MINE LAKE 2015-2018 DEMONSTRATES A EUKARYOTIC HETEROTROPH MICROBIOME IN FLUX

3.1 Preface

This chapter takes the form of a manuscript in preparation for submission, created in collaboration with the Dunfield Lab at the University of Calgary and Dr. David Bass at the London Museum of Natural History. It is a continuation of the monitoring programme that produced the data presented in Chapter 2; thus, some redundancy exists in the introductory information and the methodology. In this chapter, Chapter 2 is referred to as its published form, Richardson et al. (2020).

3.2 Introduction

3.2.1 Tailings management and Base Mine Lake

The Athabasca Oil Sands are one of the largest nonconventional oil reserves in the world. This deposit of bituminous sands, where the heavy hydrocarbon bitumen is combined with rocky sand and silt anywhere from hundreds of metres under the ground to bitumen outcrops in streams, stretches across northern Alberta, Canada, and into the neighbouring province of Saskatchewan[91]. The current proven oil reserve in this region is approximately 165 billion barrels, with another 250 billion barrels potentially recoverable with improved extraction technology[91]. Exploitation of the oil sands as a source of fossil fuels is a major economic driver of the province of Alberta and is projected to be a growing industry for decades to come[221].

To extract usable oil from oil sands, the sands must be washed with a combination of hot water and detergents to remove the bitumen from the sandy matrix[91]. Open-pit mining currently accounts for approximately 20% of bitumen extraction in the province and is the most economically feasible method of nonconventional oil production[88]. Open-pit mines have an extremely large ecological footprint, currently covering 500km^2 of northern Alberta, and the process is water intensive; it requires 1-2 barrels of water for every barrel of oil produced[88]. Tailings are a by-product of bitumen extraction and contain a mixture of coarse solids (sands), fine solids (clays), residual bitumen, and other organic compounds including naphthenic acids, heavy metals, and salts, some of which are byproducts of the extraction process and some of which are byproducts of the bitumen

deposits themselves. Tailings are held in pit settling basins (tailings ponds) to segregate the coarse solids from the water and fine solids. The water in the tailings pond is recycled for use in extraction and upgrading. Tailings settle over time, gradually transforming from the highly mixed fluid fine tailings (FFT) to stratified, thickened mature fine tailings (MFT) and oil sands process water (OSPW)[94]. This OSPW can be reused in the bitumen extraction process; currently, about 90% of the water used for bitumen extraction is recycled[90]. Some MFT has also been successfully reclaimed via sand cap, landform design, and soil cover and vegetation planting, such as Sandhill Fen on the Syncrude site[222]. While this reclamation strategy appears successful, it does not fully address the issue of FFT stored in the oil sands areas. The Canadian boreal forest, a vast circumpolar region stretching from Yukon and the Northwest Territories across northern Canada into Ontario and Quebec, contains more lakes than the rest of the world combined; robust lake ecosystems are essential to the health of this environment[89]. Any reclamation strategy for tailings will therefore have to incorporate lakes as well as terrestrial environments, due to the volume of tailings produced by the extraction process. However, the technology in this area is far less established, and it is a priority that lakes are incorporated into the reclaimed landscape of the boreal forest overlaying the Athabasca Oil Sands deposit[97].

Base Mine Lake (BML) is the first full scale pilot project for creating end-pit lake (EPL) environments in former mine pits, a strategy proposed for reclaiming. The mine pit West In-Pit, formed the basis for BML, had been filled with tailings for approximately 17 years until December 2012. At this point, the lake was commissioned as a demonstration of Water Capped Tailings Technology, and no further tailings solids entered the lake. In 2013, water from a freshwater reservoir and OSPW was used to "cap" the tailings and bring the lake water to the final design elevation. At commissioning, Base Mine Lake consisted of approximately 45m of FFT, capped with a combined OSPW and freshwater cap of approximately 8-10m[220]. The two components remain separate due to their different viscosities, with the thickness of the FFT largely preventing their intrusion into the water cap[220]. The water cap of BML was also designed to prevent resuspension of fine tailings by wind[223]. Wet-capping of BML was completed in 2013, and the lake is operated as a flow-through system. Freshwater is imported annually during open-water season from a freshwater reservoir and outflow returns to Sycnrude's Recycle Water System[224]. However, tailings reclamation is a complex process, and it will take decades to fully assess the

success of the mine closure plan; modelling of the settling of BML's tailings suggests that it will not reach its ultimate settled point until the end of the 21st century[224].

3.2.2. Base Mine Lake, 2013-2018

Since Base Mine Lake is the only full-scale pilot project of wet-capping technology in the Athabasca Oil Sands region, there is considerable interest in the biogeochemical processes at work in the system. BML has been intensively monitored throughout its reclamation life cycle, and the first multiyear studies of the state of its reclamation are beginning to be released. Throughout the early years of its reclamation (2014-2018), there have been multiple disturbances to the ecosystem, both natural and artificial. The two most notable disturbances in this period happened within months of each other around early 2016. In the winter of 2016, an active reclamation intervention was undertaken to address the ongoing issue of excessive turbidity in the water cap[204]. This turbidity, which had been noted in several studies as a potential barrier to reclamation success[124,225], was reduced through alum (as aluminum sulfate) addition and was found in the next summer to be substantially reduced[204]. This reduction in turbidity has persisted seasonally in later years; while turbidity is high in the spring months after ice thaw, as the water cap settles and stratifies after this mixing event, the water clears and light penetration increases[204].

The other ecological disturbance that substantially affected the Athabasca Oil Sands region was the Horse River Fire of May 2016. This wildfire was the costliest natural disaster in Canadian history, with $9.9 billion of direct and indirect costs and approximately 1.5 million hectares of burned forest[146]. The fire caused damage to many oil sands production facilities as well as the nearby city of Fort McMurray[146]. Wildfire has short-term and long-term effects on the ecosystem. Ash and other chemicals (such as polycyclic aromatic hydrocarbons) emitted from burning forests can be deposited into lakes, resulting in both a transient increase in turbidity and shifts in the chemical composition of the lake[226]. Wildfires can also result in flooding (due to increased erosion in burned areas) and increased hydrophobicity of burned soils, preventing water absorption. This will increase the quantity of water within a lake but can also result in the reduction of water quality. Runoff water will contain chemicals from whatever has burned in the region; in areas of industrial or urban development, this can include materials from burned buildings and their contents as well as any chemicals used as fire retardants by emergency personnel[227]. The changes in nutrient

balance afforded by wildfires can also cause multiyear trophic effects on local watersheds by increasing the availability of phosphorus and bioavailable carbon[228]. This can, in turn, lead to unusual microbial dynamics such as algal blooms, which can have substantial effects on biodiversity. The substantial effect that the Horse River fire had on the ecology of the environment around Fort McMurray and the Syncrude site may be a potential source of variation in the microbial complement of BML.

Geophysical and chemical changes in BML have largely been the focus of early reclamation management. As previously noted, the first aspect of BML to be addressed via an active management strategy was the extremely high turbidity found after BML was established[204]. Dompierre and Barbour found that, in 2013 and 2014, there were two major contributors to mass transport between the FFT and the water cap. Dewatering of the FFT where, as the tailings settle into MFT, the water it contains is displaced into the water cap results in a release of sediment, and direct mixing from wind and waves in the water cap causes disturbance of the tailings/water interface (TWI) between the water cap and the FFT[117]. Dompierre et al. (2016) also completed a detailed geochemical study of BML during the summer of 2013 and found that porewater from the FFT had a distinct chemical composition from the water cap, indicating that dewatering of FFT would likely cause a substantial shift to the water chemistry at the TWI over time, with approximately 0.73m of tailings settlement per year[229]. In 2018 there was no acute toxicity in BML's water cap, though some compounds such as ammonium, chloride, phenolic compounds, boron, and short-chain hydrocarbons were still frequently above Alberta's Protection of Aquatic Life guidelines[224]. It was determined that most of these compounds were entering the water cap from the underlying tailings porewater. Assessments of the toxicological risk of BML between 2014 and 2016, focusing on compounds which may cause ecological concern, found that these had largely been substantially diluted from freshwater inputs. The concentrations of potentially harmful chemicals such as heavy metals had reduced to negligible quantities well within the range considered safe for freshwater in the region by 2016[118]. However, salinity in BML remained elevated compared to the local watersheds (approximately 10x the salinity of the nearby Athabasca River) and this was projected to remain elevated for decades to come[118,224].

It was noted in these geochemical studies that biogeochemical processes are likely responsible for increasing the rate of dewatering in BML[229]. Further studies of BML more specifically targeted the bacteria responsible for these biogeochemical processes and identified a diverse and thriving microbiome, highly stratified both taxonomically and functionally in the FFT and the water cap layers[121,230]. Additionally, the water cap shows distinct monthly patterning across the ice-free summer[230]. The most important biogeochemical processes currently observed in BML from a reclamation perspective are methanotrophy and nitrification—the abundance of long-chain hydrocarbons at the TWI create an ideal environment for growth of methanotrophic bacteria[230]. This process, along with fixation of nitrogen from ammonia, are the two major oxygen-consuming processes in BML. Though oxygen is able to diffuse into the surface of BML, the abundance of bacteria engaged in oxygen-consuming processes at the TWI mean that, for the majority of the year, the TWI is anoxic. This trend becomes even more pronounced in later years of reclamation; in 2016, the anoxia persisted throughout the year and is projected, based on bacterial modelling, to continue consuming oxygen at a greater rate each year until the anoxia in the lower 3.5m of BML above the TWI is permanent[120].

3.2.3 Assessing microbial diversity

Environmental DNA (eDNA) sequencing is a common method of using high-throughput sequencing to assess diversity and has been successfully used to assess the microbial communities in Base Mine Lake. The majority of eDNA sequencing in BML involves metagenomics, where DNA is extracted directly from samples and sequenced[120,230]. This gives a variety of genes and genomes, though they are mostly fragmentary due to the short lengths of the sequences (most Illumina sequencing runs produce contigs of approximately 400nt). These contigs can then be assembled into longer reads which are analysed by comparison to known sequences. With enough sequencing depth, entire operons or even genomes can be assembled[231].

Though this technique is extremely effective in analysing bacterial microbiomes and has been used to provide numerous fascinating results from BML (including the identification of new bacterial clades and alkane-degrading enzymes[122,232]), its use in assessing eukaryotic diversity is limited in many environments. The vast majority of environmental samples will be overwhelmingly dominated by bacteria, which also have advantages when it comes to the DNA amplification and

sequencing. Since their genomes are not condensed and bound in a nucleus, they are more readily accessible for sequencing; additionally, eukaryotes are more likely to resist the cell lysis stages of a DNA extraction protocol[13]. A preliminary study of metagenomics versus primer-directed amplification of barcoding genes in tailings pond overwater demonstrated that over 90% of contigs obtained from metagenomic studies consisted of bacteria or organelle genes[233].

Despite these drawbacks, eDNA has been used to survey protist diversity in the oil sands area for nearly a decade. Amplicon sequencing, unlike metagenomic sequencing, involves PCR-mediated amplification of genes of interest before sequencing of the environmental sample. By amplifying only specific genes, more sequencing depth can be devoted to the process of interest[35]. For example, for taxonomic identification, the ribosomal RNA genes are commonly used as they are conserved across the tree of life but have enough variation between organisms that they can act as a molecular barcode to identify specific species. 16S sequencing has been used in assessing the bacterial microbiome of many environments of northern Alberta, including undisturbed and disturbed streams and rivers, tailings ponds, tailings deposits, and BML itself[99,102,103,233]. 18S amplicon sequencing, which specifically targets eukaryotes, was used in 2014 to evaluate the protist communities in soil samples from reclaimed and undisturbed environments[234] and in 2016 to determine the eukaryotic component of tailings and overwater from active tailings ponds[233]. The latter study, by Aguilar et al. (2016), was the first published example of eukaryotic diversity in aquatic tailings environments and found a small but diverse and apparently thriving microbiome[233].

18S sequencing has also been used to examine protist diversity in the water cap of BML in the ice-free summer of 2015[235]. Richardson et al. (2020) found a diverse community of eukaryotes which appeared to have seasonal bloom dynamics and formed distinct communities in each of the surveyed months. Additionally, careful phylogenetic analysis of the isolated operational taxonomic units (OTUs, clusters of amplicons that represent genus-level diversity at the given level of divergence) indicated that there was substantial novel diversity in the protist communities of BML[235]. The eukaryotic microbiome described in this study was also dominated by heterotrophic taxa with a noted absence of phototrophs such as algae; this is particularly unusual for a thermally stratified boreal lake in the summer as these environments are usually dominated by sequential algal blooms[235].

3.2.4 Heterotrophs in a reclamation context

One of the most unusual findings from the Richardson et al. (2020) paper was the abundance of heterotrophs identified in BML in the summer of 2015, a period usually dominated by photosynthesisers[235]. Whether this is due to the lack of photosynthetic diversity or an unusual abundance of photosynthetic taxa is still under investigation. However, the heterotrophic eukaryotic component of the microbiome is an understudied and potentially extremely informative niche for lake environments—particularly in the context of an end-pit lake still in the early stages of reclamation.

Heterotrophic eukaryotes, also known as heterotrophic flagellates (HFs), are extremely diverse and come from across the spectrum of eukaryotic diversity. Most species are small and have few distinguishing morphological or ultrastructural features. HFs occupy a trophic level above the bacteria in BML; rather than directly degrading hydrocarbons as a feeding source, they consume hydrocarbon-degrading bacteria as well as other HFs[236]. Richardson et al. (2020)'s study of the eukaryotic community of BML in the summer of 2015 found that many distinct clades of HFs were present, with blooms of different taxa dominating each month: May was dominated by the heterotrophic Ochrophyte group Synindales, while June was dominated by ciliates and September by Fungi[235]. These analyses also showed that there was considerable diversity within these groups, particularly within the ciliate class Litostomatea, the cercozoan class Glissomonadida, and the fungal class Microsporidia[235]. Unusually, all three of these classes represent some of the most poorly understood classes within these phyla. The Litostomatea are one of the few ciliate classes with no model organism representatives despite containing the only example of a ciliate capable of causing disease in humans, *Balantidinium coli*[237]. Litostomatean ciliates have been noted for their difficulty to grow in culture, and very little genetic information is available compared to most other ciliate classes particularly the more speciose examples[237]. Glissomonadida are a class of heterotrophic cercozoans, an enigmatic group within the Rhizaria. Recent DNA-based taxonomic analysis has gone a long way towards increasing our understanding of the class, but ultimately, additional sampling and culturing efforts will be necessary to understand the scope of glissomonad diversity[32,142]. Microsporidia, on the other hand, used to be understood as a solely parasitic, small class of fungi. However, DNA-based taxonomy of fungi has illustrated that in fact, many nanoflagellate Microsporidia thought to be sister to the class actually fall squarely within it, and

other species originally classified as Rozellids, a related lineage, are also within the Microsporidia[33].

It is somewhat ironic that the most abundant and speciose lineages detected within the newly established reclamation environment of Base Mine Lake belong to enigmatic and poorly understood clades. However, it also provides an exciting opportunity for bioprospecting, particularly since Richardson et al. (2020) highlighted the novel diversity indicated by the BML amplicon phylogenies[235]. The biogeochemical conditions indicated by the bacterial microbiome are also potentially attractive for HFs; blooming bacteria provide an abundant prey source and many HF lineages, particularly within the ciliates, are noted for their resistance to anoxia and toxic compounds such as naphthenic acids and residual hydrocarbons[236]. HFs also show considerable promise as reclamation indicators in other anthropogenically influenced environments, including those associated with hydrocarbon contamination[141,238]. Since heterotrophs seem to be the first to establish themselves in the eukaryotic microbiome of EPLs, finding HFs that could act as potential early indicators of reclamation in new EPLs would be an extremely useful tool for future reclamation efforts.

3.1.5 Scope of chapter

In this chapter, I investigate the heterotrophic eukaryote microbiome of BML in the four ice-free summers between 2015 and 2018 using 18S amplicon assessments. I determined that the three groups identified as particularly speciose in 2015 (Glissomonada, Litostomatea, and Microsporidia) were also highly speciose in the later years, though other heterotroph species, such as the Oligohymenophorean ciliates and Vampyrellids from within the cercozoa also reached comparable richness. I also identified a substantial increase in the eukaryotic photosynthetic microbiome of BML post-2016, indicating that the reduction in turbidity was potentially successful in establishing a photosynthetic community.

Using the classifications that I obtained, I then analysed statistical trends in the heterotrophic eukaryote microbiome of BML over time. I found that in the vast majority of the available metadata, including platform, depth, and month of sampling, no clear trends could be observed. However, for both the overall heterotroph community and for each of the majority heterotrophic

clades included in the phylogenetic analysis, the community changed year-on-year. These changes, when compared to proposed reclamation start and end comparison points Mildred Lake Settling Basin and the freshwater reservoir Beaver Creek Reservoir, showed that BML developed a distinct HF community from either of these environments. Additionally, time point clustering of abundant species revealed that, at least taxonomically, the communities pre- and post-disturbance in 2016 had a remarkably similar distribution despite containing completely different OTUs. This indicates potential for community resilience to disturbance, though the functional consequences of this are unknown.

Finally, as it was clear that the variable with the most explanatory power in the BML environment is time, I examined the community dataset for individual OTUs for any with particularly strong correlations with the time series. I identified a core heterotrophic microbiome and persistent heterotrophic microbiome in Base Mine Lake and used phylogenetics to determine the taxonomic assignment of these OTUs. I also identified OTUs that had unusual correlations with time and found two examples of OTUs which were highly abundant post-disturbance and may therefore be particularly responsive to wildfire.

3.3 Methods

3.3.1 Sampling

We took samples from three platforms located at: the centre (Platform 1), the north eastern section (Platform 2), and the south western section (Platform 3) of the lake (Figure S3.1) at intervals from below the surface to the tailings/water interface. This took place from March to September 2015. I have included metadata from each sample in Table S3.1. We rinsed 1 L polycarbonate containers for water samples three times with water before filling to the top and capping them. We transported the containers to the University of Calgary on ice in coolers at about 5$_{o}$C and processed them within 7 days of sampling. Water chemistry information was not available. We harvested the biological materials from 1 L of water samples at 8,000 × *g* for 10 min at 4ºC using the Avanti J-E high-performance centrifuge (Beckman Coulter Life Sciences, Indianapolis, USA) and stored at -20ºC until processed for DNA extraction.

3.3.2 DNA extraction, PCR amplification, and sequencing

We extracted the DNA using FastDNA Spin Kit for Soil (MP Biomedicals, Solon, OH, USA). Short fragments (380 bp) of the V4 region of the eukaryotic 18S rRNA gene were amplified using universal primers for eukaryotes (bolded)[60] with Illumina adapters attached to the 5'-ends (not bolded) : Ill_18S_F 5' - TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCAGCA(G/C)C(C/T)GCGGTAATTC C-3'; Ill_18S_R 5'- GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGACTTTCGTTCTTGAT(C/T)(A/G)A - 3'. Each 30 µl PCR mixture contained 1 µl of DNA template, 1.56 µl of primer solution (10 µM), 3 µl of 10x PCR buffer (Life Technologies Inc. Burlington, ON, Canada), 2 µl of deoxynucleoside triphosphate (dNTPs) (25 mM each), 22.23 µl of sterile nuclease-free water (Qiagen, Toronto, ON, Canada), and 0.25 µl of Pfu polymerase (2.5 U μl^{-1}) (Life Technologies Inc. Burlington, ON, Canada). We performed PCR amplification in a Veriti™ 96-Well Fast Thermal Cycler (Life Technologies Inc. Burlington, ON, Canada). The conditions for the first round of PCR: an initial denaturation at 95ºC for 5 minutes, followed by 10 cycles of touchdown PCR (denaturation at 95ºC for 30s, 30s at a primer annealing temperature of 60ºC for the first cycle and subsequent decrements of 0.5ºC per each subsequent cycle, and 30s of elongation at 72ºC), followed by 30 cycles with constant annealing temperature (denaturation at 95ºC for 30s, 30s at a primer annealing temperature of 55ºC, and 30s of elongation at 72ºC), and a final extension at 72ºC for 5 minutes. We visualized the PCR products on a 2% agarose gel using a fluorescent dye, SYBR Safe DNA Gel Stain (Life Technologies Inc. Burlington, ON, Canada). We purified the PCR products using Omega Mag-Bind RXNPure Plus beads (VWR, Edmonton, AB, Canada) according to the manufacturer's instructions. In the second round of PCR, the initial PCR products were barcoded. For this, we mixed 5 µl bead-purified initial PCR products with 5 µl of forward barcoded primer (1 µM), 5 µl of reverse barcoded primer (1 µM), and 35 µl of KAPA HiFi HotStart ReadyMix (VWR, Edmonton, AB, Canada) to adjust total volume to 50 µl. Cycling conditions were: an initial denaturation at 95ºC for 3 minutes, followed by 8 cycles with denaturation at 95ºC for 30s, 30s at a primer annealing temperature of 55ºC, and 30s of elongation at 72ºC, plus a final extension at 72ºC for 5 minutes. After the second PCR round, we visualized the final PCR products of 516 bp on a 2% agarose gel and purified again with Omega Mag-Bind RXNPure Plus beads. Concentrations of all 18S rRNA gene amplicons were quantified using a DNA quantification kit,

sDNA HS Assay Kit, and a small benchtop Qubit fluorimeter (Life Technologies Inc. Burlington, ON, Canada). For sequencing, we diluted quantified amplicons to 4 nM and pooled. Illumina sequencing was done at the University of Calgary, Alberta, Canada, using the MiSeq instrument and a sequencing kit, MiSeq® Reagent Kit v3 (600 cycle) (Illumina Canada Ulc, Vancouver, BC, Canada).

3.3.3 OTU clustering

I clustered OTUs using the QIIME2 microbial analysis platform[54]. I paired reads with a minimum length of 300 and minimum overlap of 30, with no uncalled bases. I dereplicated samples, and chimeras and singeton sequences were removed before clustering. I used two clustering thresholds as outlined in Appendix I: 99% and 97% identity. I classified these sequences via comparison to the SILVA database[239], and any that were determined to belong to bacteria or organellar genomes were removed.

3.3.4 OTU identification

I used automatic BLAST of OTUs against the SILVA database, discarding any derived from bacterial or from organellar genomes, or from multicellular organisms (Metazoa or Embryophyta)[239], for the initial classification. OTU sequences were then classified via BLAST into the PR2 and SILVA reference databases[19,239]. I then compared these classifications to ensure that the results were the same; for any phylum-level discrepancies between the classifications, the classification was determined using a third BLAST search into the GenBank database[18]. If all three classifications differed, I classified the sequence as Eukaryota with no further exposition. I visualized these classifications to the lowest level the sequences could be confidently classified, using a KronaPlot created in the KronaTools package[240].

3.3.5 Phylogenetic placement of OTUs using pplacer

I generated pplacer trees using the SILVA-filtered OTU list and a pan-eukaryotic backbone alignment generated in Aguilar et al. (2016) for classification of unknown V4 OTUs[233,241]. I generated the backbone tree for this alignment using the RAxML BlackBox algorithm on the CIPRES web server[242,243]. I added OTUs using pplacer default parameters and converted the tree to PhyloXML format using guppy, implemented in pplacer[241]. I completed further pplacer

alignments with the same methods, but with the lineage-specific heterotroph backbones used in Richardson et al. (2020)[235].

3.3.6 Phylogenetics

I produced an initial pan-eukaryotic tree using the eukaryote backbone from Aguilar et al. (2016)[233] and sorted the OTUs into the higher-level heterotroph groupings of interest (Amoebozoa, Cercozoa, Ciliophora, Excavata, Fungi) based on both their classification in the reference databases and their position within this preliminary phylogenetic alignment. I produced each specific phylogenetic tree using a backbone alignment containing the following full ribosomal regions from representative species across the diversity of the queried group as defined in Richardson et al. (2020)[235], OTUs from the BML samples classified to the group, and the entire ribosomal region top hit obtained from a BLAST search of the OTU sequences against GenBank, including unclassified and uncultured species. I aligned the sequences using the MAAFT algorithm with E-ins-i and otherwise default parameters[244]. I constructed Maximum Likelihood trees from the alignments using the RAxML BlackBox algorithm on the CIPRES web server[242,243]. I constructed MrBayes trees from the same alignments using MrBayes 3.2.6 on the CIPRES web server [243].

3.3.7 Ordination, time cluster analysis and identification of time correlated OTUs

When examining the ordination of the total dataset, I aimed to determine whether the overall differences in the communities between BCR, BML, and MLSB would potentially mask any additional trends associated with the environmental variables in each lake. Accordingly, I ran a combined ordination of all of the samples, clustered at a 97% identity, to determine how much of the variation could be explained by the provenance of the sample. I used the metaMDS function in the R package vegan 2.2.4 to carry out NMDS and bioenv analyses[245]. I ensured that each OTU in the community was found in at least 3 samples and with an overall abundance of at least 10 sequences, otherwise it was discarded. The results of these analyses were visualised using the ggplot2 package in R[246].

I generated time clusters using the Temporal Insights into Microbial Ecology (TIME) online platform by construction of a tree based upon the similarity of the graphs produced when mapping

the abundance of that OTU against time[247]. I also used this platform to generate the core microbiome and persistent microbiome using the criteria defined in Caporaso et al. (2011)[248]. Briefly, the core microbiome are OTUs present in every sample, whereas the persistent microbiome are OTUs "present in 20% or more timepoints, with at least 90% of those observations being consecutive"[248].

I identified OTUs highly correlated with time using Maximal Information Coefficient analysis via the MICtools program suite[249]. I classified trends into linear positive, linear negative, and nonlinear trends based on comparisons between the Spearman-rho and Pearson correlations and the MICtools coefficient as calculated by MICtools[249].

3.4 Results

3.4.1 Taxonomic assessments of OTUs by comparison to reference databases

I used two methods of producing taxonomic assessments of the 97% clustered OTUs from Base Mine Lake: comparison to known reference databases, using the databases SILVA, PR2 and GenBank, and phylogenetic assessment of representative OTUs against a reference backbone. The crossreferenced classifications, and the overall community contributions from each of them, are indicated in Table S3.2, and the distribution of classification outcomes is shown in Table S3.3. The taxonomic assessment for all OTUs obtained from BML between 2015 and 2018 obtained via comparison to reference databases is shown in Figure 3.1. Overall, this community shows a much greater biodiversity than the 2015-only data, particularly with regard to photosynthetic heterotrophs. While the 2015 data showed only a small percentage abundance of Archaeplastida and very little representation overall from other majority photosynthetic taxa, Figure 3.1 shows a much more substantial representation from Archaeplastida, Cryptophyta, and photosynthetic Ochrophyta. However, similar to the 2015 data, single OTUs are dominant in BML diversity. Five OTUs comprise 5% or more of overall diversity, and one fungal OTU accounts for 25% of overall abundance alone. This suggests that BML may still dominated by a relatively small core microbiome. Many of the same taxa identified as particularly speciose in 2015 are also present in the overall BML data set; Glissomonadida and Litostomatea account for 10% and 4% of the overall diversity, respectively, though neither are as dominant as they were in 2016. It is more difficult to

Figure 3.1: KronaPlot of taxonomic distributions of OTUs across Base Mine Lake samples between 2015 and 2018.
The plot represents the relative abundance of OTUs across this dataset. Classifications are based on cross-referencing the PR2, SILVA and GenBank databases. Sections marked with an asterisk (*) represent single-OTU level diversity at a 97% clustering threshold. Sections are described down to the level to which they can be confidently classified.

identify the diversity of the Microsporidia from database comparisons as the taxonomic referencing for this clade is currently under revision, but Figure 3.1 does demonstrate a substantial fungal component. Finally, two OTUs were not assigned a taxonomy based on comparison to any of the 3 databases, which means that either the three databases gave three different kingdom-level classifications or that the deciding classification, based on a comparison to the entire GenBank nt database, revealed top hits that were only uncultured eukaryotes. Given that these OTUs cumulatively represent 10% of the overall diversity in BML, the phylogenetic assessment of these sequences is particularly important.

3.4.2 Phylogenetics

As well as comparison to taxonomic databases, I used phylogenetics to compare the OTUs from BML to reference alignments. I used a two-stage process; first, I created a phylogenetic tree of all OTUs in a pan-eukaryotic alignment (Supplementary Figure 3.1) and, using the phylum-level classification obtained from this tree, split the OTUs into the three heterotrophic groups which made up the majority of the diversity in 2015: Cercozoa, Ciliophora, and Fungi (Figures 3.2, 3.3, and 3.4 respectively) . In all the trees, the number of 97%-level OTUs is approximately 10x higher than those detected in Richardson et al. (2020). This indicated that many more heterotrophic groups are present, and overall the biodiversity of BML has increased since 2015. In the case of the Fungi, the majority of the diversity is still concentrated within the Microsporidia, accounting for 70% of the detected OTUs including the highly abundant fungal sequence observed in the KronaPlot that makes up 25% of the overall diversity. However, in the Ciliophora and the Cercozoa, the overall abundance and the species richness of the different groups has shifted. While the 2015 samples from BML were dominated by a single ciliate OTU from the class Litostomatea, there is now much more diversity both within the Litostomatea and observed in other clades. Indeed, similar species richness (at a 97% clustering threshold) is observed in the rest of the spirotrichea, and protostomatea, and the oligohymenophorea; while 22% of the OTUs are classed within the Litostomatea, these other groups account for 22%, 22% and 23% respectively. For the Cercozoa, the distribution between classes is less even, but the Glissomonadida are also no longer the sole dominant group. There is roughly equivalent richness in the Vampyrellidae, another enigmatic Cercozoan lineage that is only recently being explored in detail. While the Glissomonadida account for 31% of the overall OTU richness, the Vampyrellidae account for 28%.

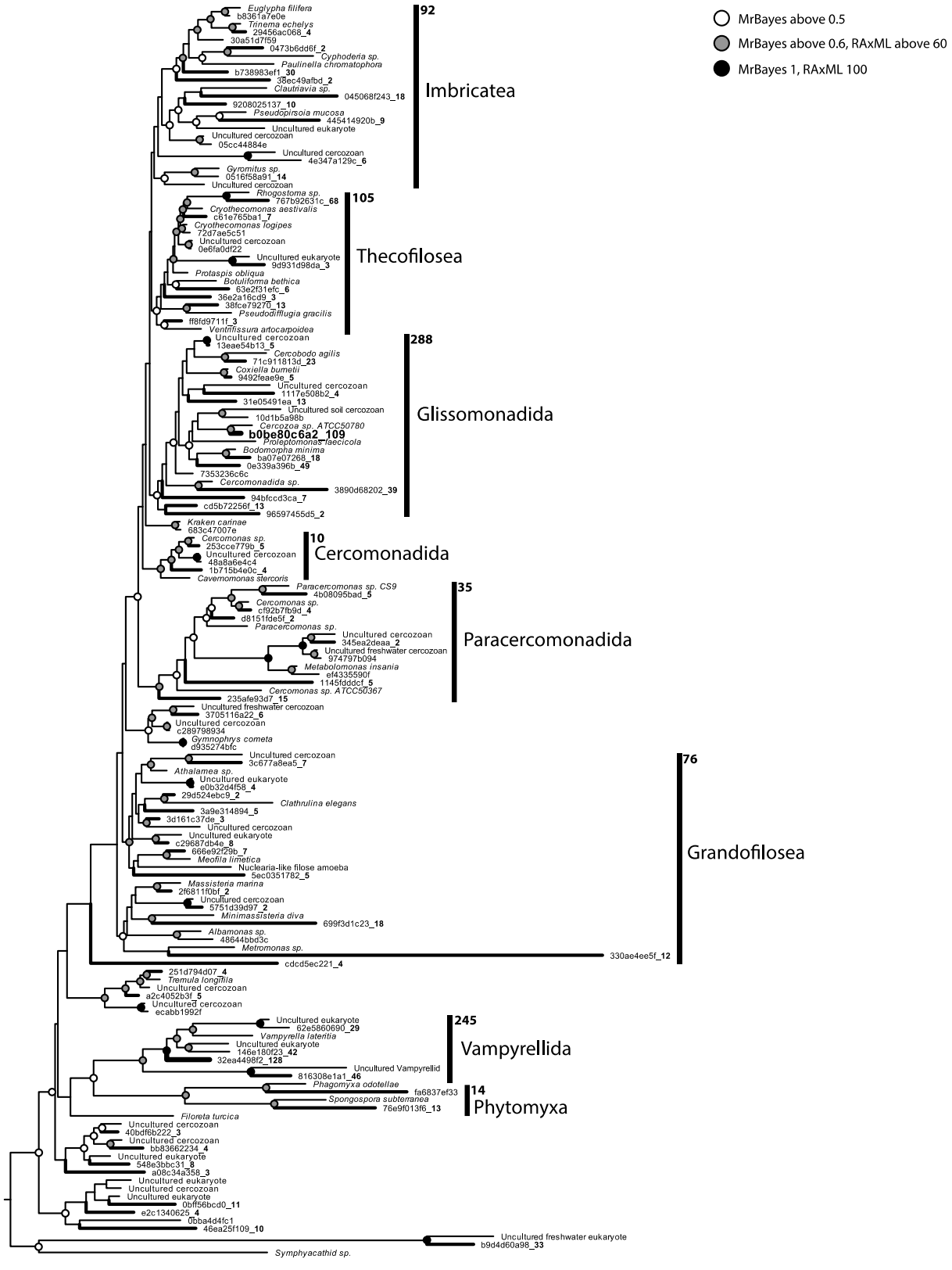

MrBayes above 0.5
MrBayes above 0.6, RAxML above 60
MrBayes 1, RAxML 100
Imbricatea
92
Thecofilosea
105
Glissomonadida
288
Cercomonadida
10
Paracercomonadida
35
Grandofilosea
76
Vampyrellida
245
Phytomyxa
14
Euglypha filifera b8361a7e0e
Trinema echelys 29456ac068_4
30a51d7f59
0473b6dd6f_2
Cyphoderia sp.
Paulinella chromatophora b738983ef1_30
38ec49afbd_2
Clautriavia sp. 045068f243_18
9208025137_10
Pseudopirsoia mucosa 445414920b_9
Uncultured eukaryote
Uncultured cercozoan 05cc44884e
Uncultured cercozoan 4e347a129c_6
Gyromitus sp. 0516f58a91_14
Uncultured cercozoan
Rhogostoma sp. 767b92631c_68
Cryothecomonas aestivalis c61e765ba1_7
Cryothecomonas longipes 72d7ae5c51
Uncultured cercozoan 0e6fa0df22
Uncultured eukaryote 9d931d98da_3
Protaspis obliqua
Botuliforma bethica 63e2f31efc_6
36e2a16cd9_3
38fce79270_13
Pseudodiffugia gracilis ff8fd9711f_3
Ventrifissura artocarpoidea
Uncultured cercozoan 13eae54b13_5
Cercobodo agilis 71c911813d_23
Coxiella bumetii 9492feae9e_5
Uncultured cercozoan 1117e508b2_4
31e05491ea_13
Uncultured soil cercozoan 10d1b5a98b
Cercozoa sp. ATCC50780 b0be80c6a2_109
Proleptomonas faecicola
Bodomorpha minima ba07e07268_18
0e339a396b_49
7353236c6c
Cercomonadida sp. 3890d68202_39
94bfccd3ca_7
cd5b72256f_13
96597455d5_2
Kraken carinae 683c47007e
Cercomonas sp. 253cce779b_5
Uncultured cercozoan 48a8a6e4c4
1b715b4e0c_4
Cavernomonas stercoris
Paracercomonas sp. CS9 4b08095bad_5
Cercomonas sp. cf92b7fb9d_4
d8151fde5f_2
Paracercomonas sp.
Uncultured cercozoan 345ea2deaa_2
Uncultured freshwater cercozoan 974797b094
Metabolomonas insania ef4335590f
1145fdddcf_5
Cercomonas sp. ATCC50367
235afe93d7_15
Uncultured freshwater cercozoan 3705116a22_6
Uncultured cercozoan c289798934
Gymnophrys cometa d935274bfc
Uncultured cercozoan 3c677a8ea5_7
Athalamea sp.
Uncultured eukaryote e0b32d4f58_4
29d524ebc9_2
Clathrulina elegans
3a9e314894_5
3d161c37de_3
Uncultured cercozoan
Uncultured eukaryote c29687db4e_8
666e92f29b_7
Meotiia limetica
Nucleraria-like filose amoeba 5ec0351782_5
Massisteria marina 2f6811f0bf_2
Uncultured cercozoan 5751d39d97_2
Minimassisteria diva 699f3d1c23_18
Albamonas sp. 48644bbd3c
Metromonas sp. 330ae4ee5f_12
cdcd5ec221_4
251d794d07_4
Tremula longifila
Uncultured cercozoan a2c4052b3f_5
Uncultured cercozoan ecabb1992f
Uncultured eukaryote 62e5860690_29
Vampyrella latentia
Uncultured eukaryote 146e180f23_42
32ea4498f2_128
Uncultured Vampyrellid 816308e1a1_46
Phagomyxa odontellae fa6837ef33
Spongospora subterranea 76e9f013f6_13
Filoreta turcica
Uncultured cercozoan 40bdf6b222_3
Uncultured cercozoan bb83662234_4
Uncultured eukaryote 548e3bbc31_8
a08c34a358_3
Uncultured eukaryote
Uncultured cercozoan
Uncultured eukaryote 0bff56bcd0_11
e2c1340625_4
0bba4d4fc1
46ea25f109_10
Uncultured freshwater eukaryote b9d4d60a98_33
Symphyacathid sp.
0.2

Figure 3.2: Phylogenetic distribution of OTUs classified as Cercozoa from Base Mine Lake.
Node support is based on RAxML (bootstrap) values and MrBayes (probability) values. Numbers by each broader classification line indicate the overall number of OTUs within that clade, and the bolded numbers next to each OTU ID indicate the number of OTUs represented by that sequence. The bolded OTU in a larger font, found within the Glissomonadida, is representative for the group that contains the most abundant OTUs, accounting for approximately 77% of the overall cercozoan abundance and approximately 10% of overall protist abundance. The Vampyrellida and Glissomonadida account for most of the species richness in the detected Cercozoa, with 245 and 288 OTUs respectively.

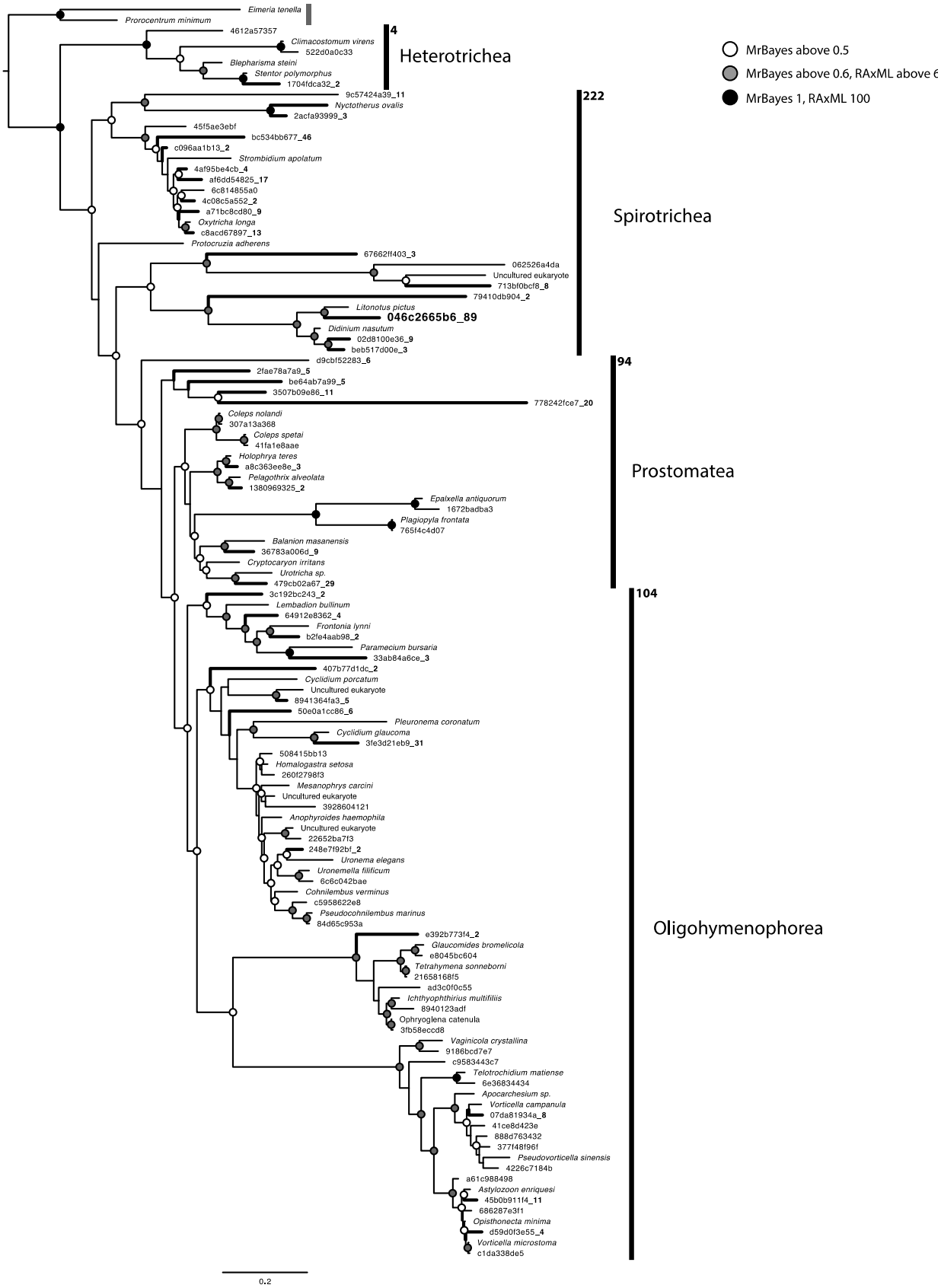

MrBayes above 0.5
MrBayes above 0.6, RAxML above 6
MrBayes 1, RAxML 100
Eimeria tenella
Prorocentrum minimum
4612a57357
Climacostomum virens
522d0a0c33
Blepharisma steini
Stentor polymorphus
1704fdca32_2
4
Heterotrichea
9c57424a39_11
Nyctotherus ovalis
2acfa93999_3
222
45f5ae3ebf
bc534bb677_46
c096aa1b13_2
Strombidium apolatum
4af95be4cb_4
af6dd54825_17
6c814855a0
4c08c5a552_2
a71bc8cd80_9
Oxytricha longa
c8acd67897_13
Protocruzia adherens
Spirotrichea
67662ff403_3
062526a4da
Uncultured eukaryote
713bf0bcf8_8
79410db904_2
Litonotus pictus
046c2665b6_89
Didinium nasutum
02d8100e36_9
beb517d00e_3
d9cbf52283_6
94
2fae78a7a9_5
be64ab7a99_5
3507b09e86_11
778242fce7_20
Coleps nolandi
307a13a368
Coleps spetai
41fa1e8aae
Holophrya teres
a8c363ee8e_3
Pelagothrix alveolata
1380969325_2
Prostomatea
Epalxella antiquorum
1672badba3
Plagiopyla frontata
765f4c4d07
Balanion masanensis
36783a006d_9
Cryptocaryon irritans
Urotricha sp.
479cb02a67_29
104
3c192bc243_2
Lembadion bullinum
64912e8362_4
Frontonia lynni
b2fe4aab98_2
Paramecium bursaria
33ab84a6ce_3
407b77d1dc_2
Cyclidium porcatum
Uncultured eukaryote
8941364fa3_5
50e0a1cc86_6
Pleuronema coronatum
Cyclidium glaucoma
3fe3d21eb9_31
508415bb13
Homalogastra setosa
260f2798f3
Mesanophrys carcini
Uncultured eukaryote
3928604121
Anophyroides haemophila
Uncultured eukaryote
22652ba7f3
248e7f92bf_2
Uronema elegans
Uronemella filificum
6c6c042bae
Cohnilembus verminus
c5958622e8
Pseudocohnilembus marinus
84d65c953a
Oligohymenophorea
e392b773f4_2
Glaucomides bromelicola
e8045bc604
Tetrahymena sonneborni
21658168f5
ad3c0f0c55
Ichthyophthirius multifiliis
8940123adf
Ophryoglena catenula
3fb58eccd8
Vaginicola crystallina
9186bcd7e7
c9583443c7
Telotrochidium matiense
6e36834434
Apocarchesium sp.
Vorticella campanula
07da81934a_8
41ce8d423e
888d763432
377f48f96f
Pseudovorticella sinensis
4226c7184b
a61c988498
Astylozoon enriquesi
45b0b911f4_11
686287e3f1
Opisthonecta minima
d59d0f3e55_4
Vorticella microstoma
c1da338de5
0.2

Figure 3.3: Phylogenetic distribution of OTUs classified as Ciliophora from Base Mine Lake.
Node support is based on RAxML (bootstrap) values and MrBayes (probability) values. Numbers by each broader classification line indicate the overall number of OTUs within that clade, and the bolded numbers next to each OTU ID indicate the number of OTUs represented by that sequence. The bolded OTU in a larger font, found within the Litostomatea, is representative for the group that contains the most abundant OTUs, accounting for 59% of the Ciliophora diversity and 4% of the overall diversity. The Oligohymenophorea and the Litostomatea contain the most species richness, with 104 and 111 OTUs respectively.

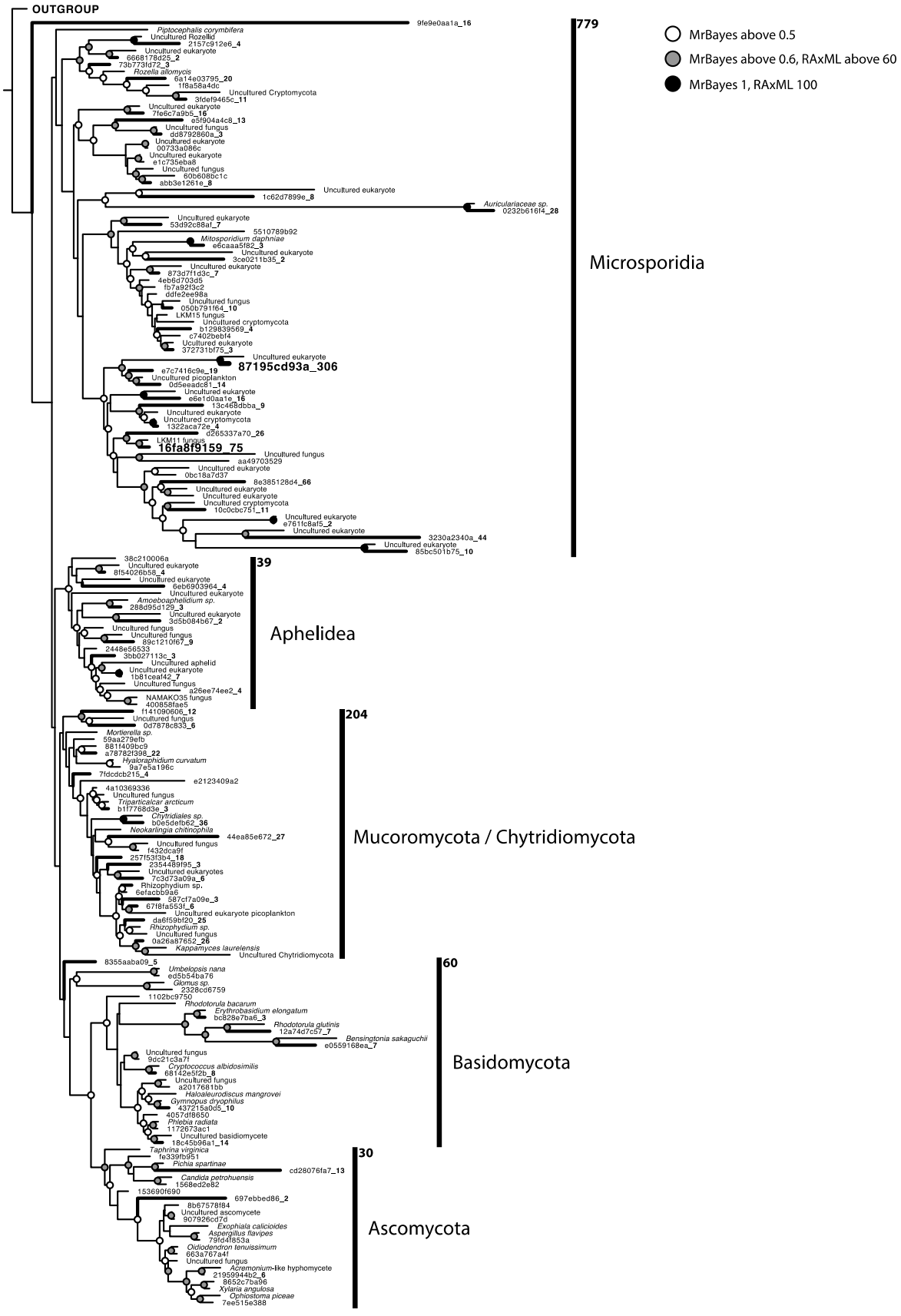

OUTGROUP
MrBayes above 0.5
MrBayes above 0.6, RAxML above 60
MrBayes 1, RAxML 100
779
Microsporidia
87195cd93a_306
16fa8f9159_75
39
Aphelidea
204
Mucoromycota / Chytridiomycota
60
Basidomycota
30
Ascomycota

Figure 3.4: Phylogenetic distribution of OTUs classified as Fungi from Base Mine Lake.
Node support is based on RAxML (bootstrap) values and MrBayes (probability) values. Numbers by each broader classification line indicate the overall number of OTUs within that clade, and the bolded numbers next to each OTU ID indicate the number of OTUs represented by that sequence. The bolded OTUs in a larger font, found within the Microsporidia, are representative for the groups that contains the most abundant OTUs, accounting for 88% of the fungal diversity and 25% of the overall protist diversity for 87195c93a, and 4% of the fungal diversity and 1% of the overall protist diversity for 16fa8f9159. The Oligohymenophorea and the Litostomatea contain the most species richness, with 104 and 111 OTUs respectively.

3.4.3 Ordination

Alongside the sequencing reads, I had access to metadata concerning the month, depth, and platform of sampling for each of these sequences, as well as reference samples collected more infrequently from two other water bodies in the Base Mine Lake area. These comparison points consisted of Mildred Lake Settling Basin (MLSB), an active tailings pond on the Syncrude mine

site and the source of much of the FFT that was eventually incorporated into BML, and Beaver Creek Reservoir (BCR), an artificial reservoir close to Base Mine Lake. BCR is used as the freshwater input for BML, as the water level is regularly adjusted to account for evaporation.

The NMDS ordination converged in 3 dimensions with a maximal stress of 0.184 and samples grouped by lake (Figure 3.5A). I then used the bioenv function to determine which environmental variables correlated to observed groups, and the two significant variables were determined to be Lake and Year with a correlation coefficient of 0.3393. This was consistent with the initial hypothesis described in the Methods, that Lake would be a substantial source of the variation in biodiversity between the microbial communities and that the lake communities may need to be analysed individually to determine the within-lake trends.

Repeating these analyses with the dataset clustered at 99% uncovered the same overall trends. When I carried out a bioenv analysis of this dataset, only Year came out as an explanatory variable, with a correlation of 0.278. I was unable to detect any significant relationship between any of the other metadata variables (month of sampling, platform). As it seemed that both the overall community across the 3 lakes and the community within BML changed over time, I separated the samples from each year and analysed the relationships between the lakes. Over time, the communities in BML, BCR and MLSB became even more distinct. An NMDS analysis of the summer of 2018, which converged in 3 dimensions with a stress of 0.14, demonstrated complete separation of these communities (Figure 3.5C). In this dataset, the Lake variable had a correlation coefficient of 0.54 with the data when analysed using bioenv.

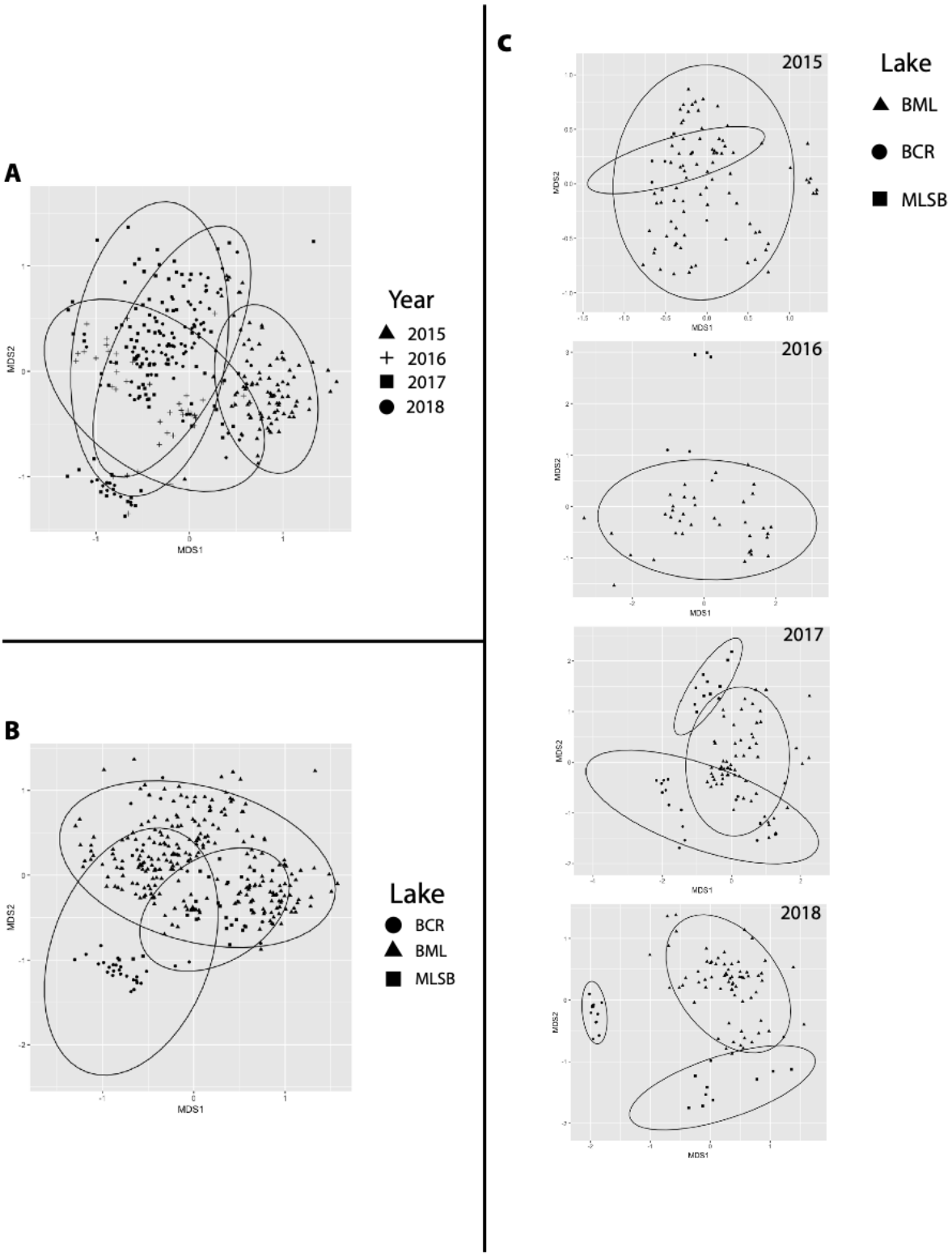


Figure 3.5: Ordination of OTUs from 2015-2018.

For all images, the freshwater reservoir Beaver Creek Reservoir is abbreviated as BCR, the reclamation site Base Mine Lake is abbreviated as BML, and the reference active tailings pond Mildred Lake Settling Basin is abbreviated as MLSB. **A:** Ordination of OTUs from across the three lakes, with year indicated by the geometry of the points and by ellipses. **B:** Ordination of the overall dataset, with the sampled lake indicated by the geometry of the points and by ellipses. **C:** Ordination of the overall dataset, across the years of sampling (2015-2018), with the sampled lake indicated by the geometry of the points and by ellipses.

3.4.4 Evaluating pre- and post-2016 OTU changes

The fact that the year of sampling had the most explanatory power of any environmental variable in the Base Mine Lake dataset, that the ecological history of the BML environment included two major disturbances in early 2016, and that previous studies had noted shifts in the environment and bacterial communities from 2016 onwards, suggested that the pre-and post-2016 environment may have fundamental differences in both taxonomy and function. To evaluate the pre-and post-2016 eukaryotic heterotroph community in BML, I used the pan-eukaryotic phylogenetic trees generated in Section 3.3.2 to identify the OTUs in each of the heterotrophic clades. I extracted these sequences from the overall OTU dataset to generate a heterotroph abundance table and used the online platform Temporal Insights in Microbial Ecology (TIME) to generate a report of trends across time for the heterotrophic OTUs based on months from Time 0 (January 2015, the start of the current sampling programme). This protocol uses dynamic time warping to identify groups of taxa with a similar abundance pattern over time. The resulting clustered tree showed two large nodes which, when plotted on a graph of abundance over time, correlated to taxa abundant pre-2016 and taxa abundant post-2016 (Figure 3.6A). I then used the previously obtained taxonomic assessments for the 97% clustered datasets to determine the taxonomy of these pre- and post-2016 time correlated OTUs and found that the distribution of taxonomies was incredibly similar both in abundance and distribution (Figure 3.6B).

3.4.5 Core and persistent heterotrophic microbiome of BML

One can infer that the core microbiome of Base Mine Lake is small based on the dominance of single OTUs from the overall abundance across the dataset. I used the TIME platform to assess the core and persistent microbiome of BML: the ‘core’ microbiome is defined as being consistent across all time points whereas the ‘persistent’ microbiome is defined as OTUs present in extended runs of time points, but not all. In BML, only five OTUs were consistently detected in every sample (Figure 3.7).

There are two fungal sequences in this core microbiome; the first corresponds to the OTU that was found to comprise 25% of the overall abundance. Contrary to its classification through database comparisons as Candida, this OTU is grouped within the Microsporidian clade and corresponds to an OTU determined to be potentially novel diversity in the Richardson et al. (2020) study of 2015

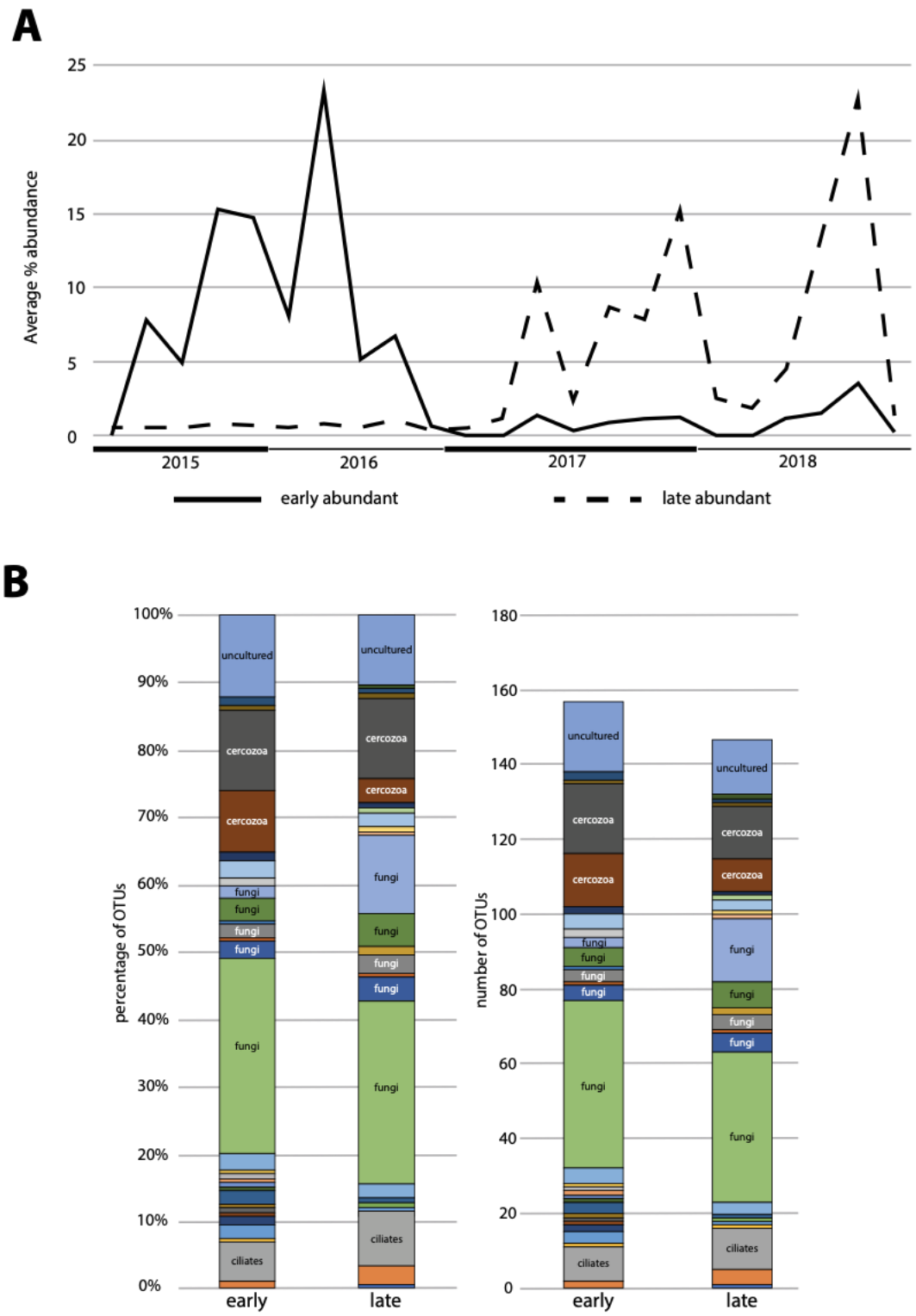

A
Average % abundance
25
20
15
10
5
0
2015
2016
2017
2018
early abundant
late abundant
B
percentage of OTUs
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
uncultured
cercozoa
fungi
ciliates
early
late
number of OTUs
180
160
140
120
100
80
60
40
20
0

Figure 3.6: OTU distribution and classification pre- and post-2015 in Base Mine Lake.
A: average abundance of OTUs that have distinct early-abundant and late-abdundant distribution as determined by TIME analysis. The two clusters of OTUs show two clear average distributions. **B:** Comparative classifications of OTUs within the two clusters, to the class level, with both relative and absolute abundance. Phylum-level classifications of groups of interest are labelled on each bar chart, which account for the majority of the abundant clades.

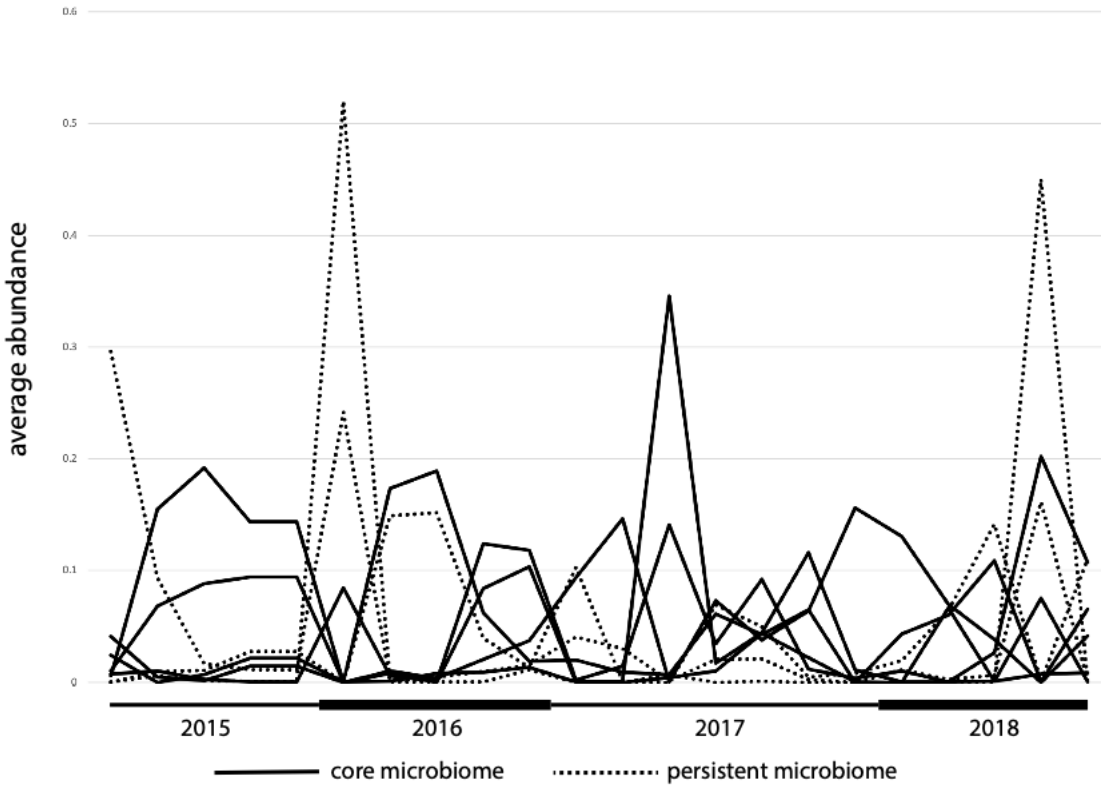


Figure 3.7: Core and persistent microbiome of BML.
The distribution of abundances of the eight core and persistent OTUs is indicated across the 48 months of sampling. The average abundance is based on the average abundance of each OTU at each time point, averaged across all the samples obtained from that time point. The definitions of "core" and "persistent" microbiome are taken from Caporaso et al. (2011), where, briefly, the core microbiome are OTUs present in every sample, whereas the persistent microbiome are OTUs "present in 20% or more timepoints, with at least 90% of those observations being consecutive".

(OTU622). The second fungal OTU is classified as Candida by database comparisons and, in the phylogenetic analysis, is also found within the Ascomycota. Despite being found in every sample, both OTUs have very different abundances across time; the Microsporidian OTU was highly abundant before the disturbances of 2016 and minimal post-2016. On the other hand, the Ascomycotan OTU has sporadic spikes in abundance but no consistent relationship with time.

There is a single ciliate found in the core microbiome of BML that was phylogenetically and database classified to the Oligohymenophorea. The detected abundance of this OTU is much more sporadic and variable, though abundance was consistently minimally detected pre-2016. There is also an Oligohymenorphorean ciliate identified as persistent. A single cercozoan, defined phylogenetically and database-classified to the Glissomonadida is part of the core microbiome; it is minimally abundant pre-2016.

Interestingly, the two uncultured OTUs identified in the taxonomic assessment of BML (each accounting for 5% of the overall diversity) were also identified in the assessment of the core and persistent microbiome—one as a core member and one as a persistent member. Both of these OTUs also had extremely interesting distributions with respect to the known ecological history of Base Mine Lake; both are minimally abundant in all samples except for those taken in July 2016, the first samples to be obtained after the mine site was reopened following the Horse River Fire. These OTUs are incredibly abundant in these samples and make up approximately 20% of the overall detected diversity in BML. I used a pplacer analysis to determine the position of the OTUs identified in the core and persistent microbiome and the MICtools analysis in a pan-eukaryotic backbone (Supplementary Figure S3.5). Though the broad taxonomic sampling of this backbone means that the classifications are not as precise as those from the heterotrophic clade-specific phylogenies in Figures 3.2-3.4, this suggests that the unidentified sequences without an assigned heterotrophic clade as described above may be basal opistokonts; in this tree, they group with the unranked holozoan slime mould *Fonticula alba*.

To determine whether the observed results with respect to time were statistically significant, I used Maximal Information Coefficient (MIC) analysis using the programme MICtools. The benefit of MIC over other statistical methods is that it is able to detect nonlinear as well as linear relationships, and since many of the OTUs in BML were hypothesised to respond to seasonal

dynamics, both nonlinear and linear relationships with time were expected. I used the entire OTU dataset clustered at 97%, along with the abundance data by month from Time = 0, to determine which OTUs showed a statistically significant relationship with time when considering the number of OTU comparisons. This analysis identified 35 OTUs with significant MICs. I further subset this data using a comparison between the MIC, Pearson correlation, and Spearman correlation into OTUs with a linear positive, linear negative, and non-linear relationships with time (Table S3.4). Those with a linear positive relationship were largely photosynthetic, which corresponds to the return of photosynthetic taxa post-2016. Of the OTUs identified in the core and persistent microbiome, all were also detected in this analysis as statistically significant.

3.5 Discussion

3.5.1 Eukaryotic community assessment

I carried out an overall eukaryotic community assessment of Base Mine Lake between 2015 and 2018 using both a crossreferencing protocol to compare the obtained OTUs to three major taxonomic databases (GenBank, PR2 and SILVA) and by phylogenetic placement within pan-eukaryotic and heterotrophic backbones as described in Richardson et al. (2020)[18,19,235,239]. This cross-database comparison revealed substantial heterogeneity between databases for classifications of the same sequence. Though the majority (83%) of the classifications were identical to phylum level between PR2 and SILVA, 17% showed substantial, often kingdom-level, classification discrepancies between these two major classification databases and, of the overall dataset, 13% of OTUs generated at this clustering threshold could not be confidently assigned to the domain Eukaryota (Table S3.3, Figure 3.1). This could be due to several factors. The two databases are maintained by different teams, which means that different sections of the database could have been more or less up to date depending on the curation effort in that taxonomic clade[250]. The BLAST algorithm itself is inherently stochastic and may not always obtain the global optima when aligning two sequences[251]. Undersampling of certain areas of the eukaryotic tree may lead to sequences not adequately represented in any database and the BLAST algorithm locating the nearest top hit[21]. Ultimately, this is an issue with any database classification method of taxonomic assignment and can only be addressed with greater curation efforts for protist taxa. There are numerous efforts to increase taxonomic resolution underway, and various researchers have different methods of combatting this, either by using environment-specific curated databases or by

increasing global efforts of taxonomic curation[252]. These include the EukRef project, which has recently joined forces with PR2 to increase the resolution and diversity of protist reference data[21].

The KronaPlot of overall taxonomic assessment has a very similar overall community composition to the 2015-only data in heterotrophic regions—in particular, I identified a very substantial fungal component with large segments from Glissomonadida. There are also extremely dominant OTUs, with multiple OTUs comprising more than 5% of the overall diversity. This may be due to the clustering threshold used when creating the OTUs; while there is no single OTU clustering threshold that adequately reflects species-level diversity, it is generally agreed that a 97% clustering threshold is more reflective of genus-level than species-level diversity which would result in artificially high OTU abundances[71].

Conversely, there are some substantial differences between the community assessment from 2015 in Richardson et al. (2020)[235] and the community assessment taking into account later years. The first is the recovery of many more photosynthetic taxa, which appear from 2016 onwards in much higher abundances. This is likely related to the reduction in turbidity due to the addition of alum in early 2016 which allowed for greater light penetration into the water cap[224]. The biodiversity of ciliates has also increased; while the 2015 ciliate diversity was dominated by a single Litostomatean OTU[235], the overall ciliate abundance across the four years shows a much greater variety of taxa with a higher Oligohymenophorean representation. Ciliates have been extensively studied in reclamation environments and have been found in many studies of hydrocarbon-based disturbances to increase in biodiversity as the disturbance decreases[83,238,253–255]. The apparent increase in biodiversity in ciliates over time in Base Mine Lake suggests that ciliate biodiversity may be a useful proxy for reclamation progress, or that bioindicator species could be identified from the ciliate phylum for end-pit lake environments.

Incorporation of phylogenetic taxonomic assessments for these OTUs also revealed that there is still substantial diversity in the groups that demonstrated the most novel diversity in 2015 (Glissomonadida, Litostomatea and Microsporidia). In fungal, cercozoan, and ciliate backbone trees, 33, 26, and 70% of OTUs obtained from BML across the four-year study were grouped into these classes compared to the overall dataset, respectively. For the fungal and cercozoan trees, the relatively poorly understood Glissomonadida and Microsporidia still appear to be abundant and

important parts of the BML ecosystem and demonstrate the importance of continual study of these taxa. Conversely, the domination of Litostomatean ciliates observed in 2015 and noted as reduced in the KronaPlot analysis is also corroborated by the phylogenetic evidence, underlining the importance of ciliate biodiversity as a potential indicator of reclamation success. Additionally, Vampyrellidae demonstrate equivalent OTU richness to Glissomonadida in the phylogenetic analysis of the four-year dataset. Understanding of the diversity of this clade is also developing, and it may play an expanded role in the ecology of BML in future years.

It is important to note that the diversity of heterotrophs obtained from the protocol outlined in 3.2.2 may result in systemic biases against particular groups of eukaryotes. In particular, the V4 regions amplified by the Stoeck et al. (2010) primers used in this study are known to introduce biases to communities through inconsistent amplification, particularly in oxygen-deficient environments[63]. This may result in their relative lack of abundance in the community assessment (though phylogenetic analysis indicates that, despite their low abundance, the Amoebozoa appear to be particularly diverse). eDNA assessments can also be affected by the cell biology of the organism; for example, ciliates contain multiple copies of their rRNA genes, and their diversity may be artificially amplified[256].

3.5.2. Ordination

For each sample taken from Base Mine Lake, I had information on the month of sampling, platform from which sampling occurred (see 3.2.1 for details on platforms), the month the sample was taken, and the depth from which the sample was taken. I also had access to reference data from two nearby water bodies, both of which represent different points along the reclamation timeline. Mildred Lake Settling Basin, an active tailings pond, contains both unreclaimed FFT and MFT and an OSPW water cap; Beaver Creek Reservoir, conversely, is an artificial freshwater body next to BML that acts as a source of dilution of the water cap. Reid et al. (2019)[101] have suggested that, due to similarities in the construction, biogeochemistry and microbiome of artificial reservoirs in the Athabasca Oil Sands region, these are a useful comparison point for the potential end result of reclamation[101].

To determine the extent to which the heterotrophic microbiome of BML is correlated with known metadata for Base Mine Lake, I used the BML dataset clustered at 97% similarity for NMDS ordination analysis, consistent with Richardson et al. (2020)[235]. Since ordination does not require taxonomic classification of OTUs, I was able to include a clustering threshold more representative of real-life species diversity than that used for the other analyses in this chapter. However, to ensure the results were in line with the other analyses on this dataset, I repeated the ordination analyses with the same 99% clustered dataset used in the taxonomic assessment and obtained the same trends (Supplementary Data). The primary result obtained from these analyses was that, for the entire dataset, the metadata variables with the most explanatory power were the lake the sample came from and the year of sampling. Both of these results were to be expected as the sampled lakes have very different compositions, ecological histories, and biogeochemistry[101]. The fact that the year of sampling was so significant is also not surprising as the microbiome of the Base Mine Lake environment (which makes up the majority of the samples) has varied sharply over these times, both in its geophysical parameters and biogeochemistry[118,120,229,230]. The NMDS analysis of all samples separated by year (Figure 3.5B) showed that the community most strongly separated from the others was that of 2015; this is consistent with the rest of the data and with the results of Richardson et al. (2020), which described a highly unusual heterotroph-dominated community[235].

To ensure that any additional variation correlated with metadata was not obscured by the changes in lake community composition, I also ran an ordination analysis with only the BML samples. In this analysis, the only variable that significantly correlated with community composition was the year of sampling. This is unusual as, in most lakes in the region, month of sampling and depth of sampling are highly correlated with community composition, particularly in the summer months[257]. The fact that there is still no apparent correlation in the overall BML dataset with the depth of sampling is also highly unexpected; like many circumpolar lakes that seasonally freeze, BML is highly stratified by temperature in the summer months and studies of lake turbidity have shown a seasonal pattern of settling that should also affect the community composition at different levels of light penetration[120,121,204]. There are several potential explanations for this observation. The first is that BML is still simply too early in the reclamation process to have a stable heterotrophic community. All other analyses of the lake, whether biogeochemical or physical, have observed that the environment is still changing (for example, the transient anoxia at the base of the

water cap has had dramatically different values throughout the years of observation)[120,230]. The ecological history of BML has also demonstrated two substantial disturbances in early 2016, almost exactly halfway through the sampling period; the response to this disturbance and gradual recovery of the community from both alum addition and wildfire may have obscured any seasonal trends[146,204,224]. Alternatively, this may be due to a sampling issue; the Base Mine Lake sampling effort may not fully capture the depths of the heterotrophic community, which may also explain the tiny core and persistent microbiome. Even in the BML models where Year was the only significant variable detected with bioenv, the correlation was fairly low at 0.278; the majority of the variation in the dataset was unexplained. This points to either an as-yet-unidentified variable being responsible for the majority of this variation, or that BML is still subject to substantial stoachastic variation over time.

To determine to what extent different heterotrophs were affected by both alum addition and wildfire, I used a time cluster analysis to identify a community of OTUs which were negatively affected by these disturbances (i.e. that were high abundance before the winter and early summer of 2016 and minimal afterwards) and a community that were positively affected (i.e. were minimally present in the environment before 2016 and then abundant afterward). The taxonomic affiliations of these communities are strikingly similar and suggest that there may be some level of resilience to disturbance within the heterotroph community; it also suggests that the dominant organisms in the BML microbiome may be determined somewhat stochastically. If the disturbances in 2016 caused a loss or substantial shift of heterotroph biodiversity, the fact that the microbial community that was reduced in abundance and the one that arose after the disturbance are so similar in taxonomic affiliation suggests that the organisms filling the niches left by these disturbances are whichever are able to obtain abundance in that particular season. It is tempting to draw a link between taxonomic affiliation and ecosystem functionality, and while this may be the case generally in bacteria and in particular eukaryotes, this cannot be inferred across the board[258]. Functional redundancy, where multiple taxa carry out the same metabolic processes within in environment, has been shown to be incredibly important in structuring microbial communities in multiple environments, and functional profiles of microbes are often more closely associated with environmental variables than taxonomy[259,260]. While these results may imply some level of functional redundancy in the microbiome of BML when it comes to heterotrophs, more careful

analysis would be required (for example, amplification of functional genes or manual annotation of taxa based on the relevant literature) before these conclusions could be drawn.

3.5.3 The core and persistent microbiome of eukaryotic heterotrophs in Base Mine Lake

The ordination and time-clustered results indicate that there is substantial turnover in the heterotrophic eukaryote populations of Base Mine Lake. This is corroborated by the extremely small core and persistent microbiome: the OTUs that are present in every sampling month (core) or present in extended sampling months (persistent). One would expect that heterotrophic species would be less strongly seasonal or susceptible to bloom dynamics than photosynthetic species, whose abundance is much more strongly correlated with external environmental variables[261]. However, only five OTUs were identified in every sample, and only a further three were common enough to be classed as 'persistent'. Many of these OTUs have a very similar classification to those found as abundant in the 2015 samples, including two Microsporidian, two Litostomatean, and two Glissomonadidan OTUs. As noted in the 2015 study, these taxa are natural fits for the BML early reclamation environment; they can feed on bacteria and many of the heterotrophic lineages are omnivorous, which is a useful trait in an environment with a high turnover in its microbiota[32,33,142,237]. Resistance to anoxia is also common in protist heterotrophs which may be a useful trait in BML, particularly at the fluid/water interface[120]. As also noted in Chapter 2 and in Richardson et al. (2020), heterotrophic protists may also be able to play a more active role in reclamation; there are numerous examples of hydrocarbon-degrading fungi, and ciliates have been noted for their role in dispersing hydrocarbons in aquatic environments[102,137,138].

The reasons for the small core and persistent microbiome are similar to those for why there are no strong correlations with any of the metadata variables. The reclamation environment may currently be changing too fast to establish a seasonal microbiome, and the disturbances in 2016 may have obscured any evidence of a seasonal microbiome or potentially a systematic under-sampling of these taxa. Ultimately, the only way to determine which factors are responsible, and to what extent, is a continuation of the monitoring programme in BML. Since the water cap of BML has been finalised for the next several years and will likely not involve any other substantial disturbances (unless there are more ecological disturbances in the region due to natural causes), there is less likely to be any obscuring factors[204]. Ideally, the biodiversity and species richness of the core and

persistent microbiome in BML will increase over time. It is an established ecological principle that increased biodiversity leads to environmental resilience[262,263]; as the current biogeochemistry of BML may promote eutrophication[120], it is essential to ensure that the heterotrophic eukaryotes that keep this process in check are able to thrive.

The presence of two uncultured OTUs that are particularly abundant immediately after the Horse River Fire, and minimally abundant elsewhere, are also a point of particular interest. There are several potential reasons for the pattern observed. Wildfires reduce the amount of water absorption by soil and promote stormwater runoff into streams and lakes, which can cause localised flooding after a rainstorm[226]. This can also result in physical mixing of water bodies and movement of organisms around in the 3D space of a lake—in the case of BML, this could mean that these OTUs are generally more abundant in the lake but are somehow missed by the sampling programme. Since the current BML microbial monitoring programme uses three midlake platforms in different regions of the lake[220], this is a possibility; however, it is unlikely as the organism is still present in other samples, just at a much smaller concentration. Alternatively, these organisms are particularly abundant after wildfires. This would be an interesting possibility as wildfires do cause a lot of biological and chemical shifts in a water body that would favour the growth of some microorganisms over others[228].

3.5.4 Conclusions

Base Mine Lake is an extremely important environment from an ecological and an economic perspective; for the successful reclamation of mining by-products (tailings and OSPW) from the Athabasca Oil Sands, it is essential that a protocol for reclamation of FFT and the integration of their water caps into the local watersheds is established. Heterotrophic eukaryote diversity in the early stages of BML's reclamation do not appear to have well established stratification either by depth or seasonally, which does not correspond to the diversity observed in the bacterial microbiome. Heterotrophic eukaryotes may therefore offer an early indicator of reclamation success, as they appear to grow established communities earlier in the reclamation process than photosynthetic eukaryotes and respond to the changing environment in end-pit lakes under reclamation. Many heterotrophic eukaryotes in BML appear to be from understudied clades like the Microsporidia and the Glissomonadida, and two uncultured eukaryotes show strong response

to wildfire. This suggests that that BML may also be a valuable environment for protist bioprospecting. Ultimately, only a continuation of the monitoring programme will determine how heterotrophs can contribute to reclamation success in the Athabasca Oil Sands.

3.6 Afterword

These data represent the heterotrophic protist community of BML between 2015 and 2018. Concurrently to this analysis, the Dunfield group has been analysing the photosynthetic component of the protist community, briefly discussed in Syncrude's 2019 report to the Alberta Energy Regulator[224].

Chapter 4

MEMBRANE TRAFFICKING AND EVOLUTIONARY MECHANISMS OF ADAPTATION IN THE ECOLOGICALLY RELEVANT HETEROTROPHIC PROTIST PHYLUM CILIOPHORA

4.1 Preface

The work in this chapter has been drawn from multiple sources, some of which have been previously published. The bulk of this work originates from 2015 and was presented at the 2017 International Congress of Protistology (ICOP) in Prague. The functional work on the HOPS/CORVET complexes was carried out in 2016 and published as Sparvoli et al. (2017)[264], and much of introduction was drawn from my contributions to Guerrier et al. (2017)[265]. Since the initial publication of this research, additional sampling points have emerged; most notably, the newly sequenced genomes *Uroleptopsis citrina*[266] and *Euplotes octocarinatus*[267]. This chapter is a synthesis of the overall survey of membrane trafficking components containing annotated genome information from all available ciliate sources up to 2019, functional work carried out in collaboration with Dr. Daniela Sparvoli and Dr. Aaron Turkewitz at the University of Chicago and published in 2017, and bioinformatic analysis of protein complexes of interest carried out in 2019. Because of the disparate nature of this work, I have therefore subdivided the results and discussion by the type of membrane trafficking machinery under discussion. The first section concerns the heterotetrametric adaptin complexes (HTACs) and includes an updated version of the overall survey of ciliate HTAC diversity presented in Prague. The second section concerns the endocytic machinery and include a functional domain analysis and phylogenetic analysis of the protein Eps15R which, based on the initial survey of endocytic machinery, has been substantially duplicated in several ciliate genomes. The third section concerns the membrane tethering complexes and includes a functional and bioinformatic analysis of the HOPS/CORVET membrane tethering complexes as initially presented in Sparvoli et al. (2017)[264].

For the work in this chapter, I would like to thank Aaron Turkewitz and Daniela Sparvoli for their *T. thermophila* work and providing an initial introduction to ciliate genomics and ciliates in general. I would also like to thank Eleni Gentekaki, Anatasios Tsaousis, and Denis Lynn for providing the ciliate transcriptomic data I used in Chapter 4.

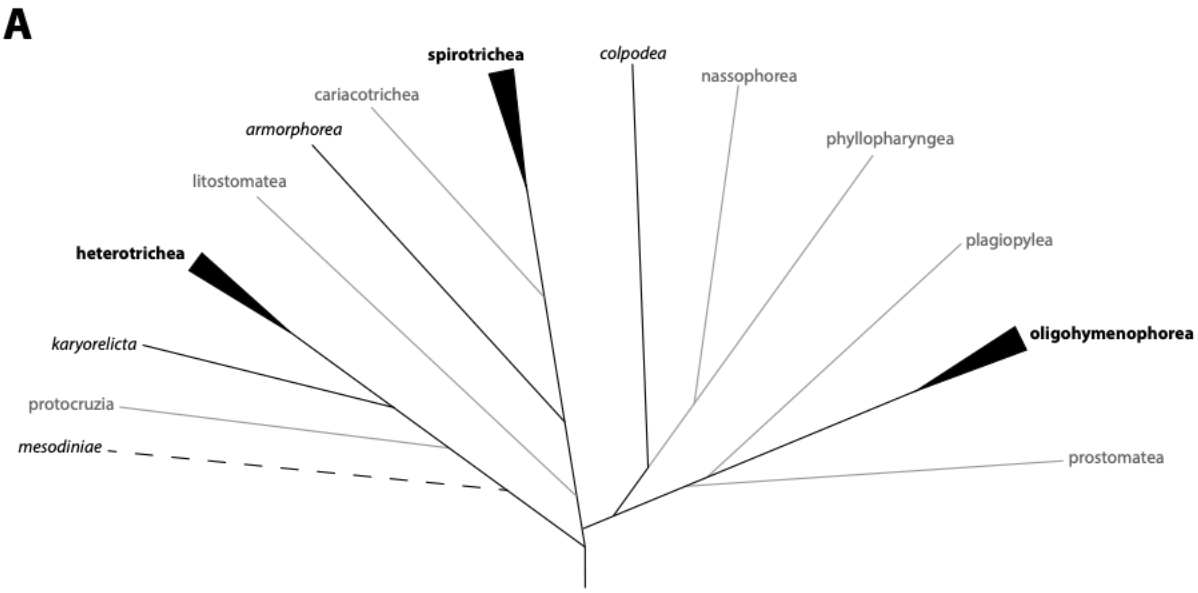
4.2 Introduction

4.2.1 "The ciliated protozoa"

Many of the most notable heterotrophic protists in surveys of anthropogenically influenced environments and in the literature surrounding the cell biological effects of hydrocarbons on protists come from the phylum Ciliophora, or ciliates[268]. This group of organisms is found within the alveolate superphylum, which is within the TSAR supergroup (telonemids, stramenopiles, alveolates, and rhizarians) (Figure 4.1A, B)[269]. Ciliates have been studied for centuries due to their large size and morphological diversity; some of the earliest identified protozoa were ciliates, a phylum named for the cilia that cover the outside of the cell surface in each species. These characteristic cell structures provide assistance for ciliates in movement and prey capture and have allowed them to become some of the most successful microscopic hunters known across the diversity of eukaryotes[269].

Though the ciliate phylum exhibits a wide array of morphologies, biogeographies, and life histories, none are exclusively photosynthetic; unlike the closely related dinoflagellates, no ciliate has fully undergone endosymbiosis[269]. All ciliates are some combination of parasitic or heterotrophic, are generally large (at the extreme, *Stentor coeruleus* can grow up to 2mm in size), and have a highly regulated cellular structure (Figure 4.1C)[269]. Ciliate species are found in almost every environment on earth and are particularly noted for their abundance in freshwater and marine ecosystems[270]. They are highly ecologically relevant, both as regulators of trophic webs and in microbial nutrient cycling[271,272]. Since they are such highly successful bacteriovores, ciliates are extremely important for preventing eutrophication through cyanobacterial blooms in many lakes and regulating the size of algal blooms in the open ocean[273,274]. They are also extremely important in the management of wastewater; it has been extensively noted that without the presence of ciliates, it is impossible to clean urban wastewater of harmful bacteria so it can be released into local watersheds[141,275]. However, ciliate abundance in ecosystems is not always beneficial. Ciliates

are also parasites and epibionts of multicellular organisms from rotifers to fish and have various roles in these ecosystems; in some cases, this can result in disease[276,277]. Ciliates such as *Ichthyopthirius multifilis* are causative agents of fish diseases such as white spot and whirling disease and are responsible for substantial morbidity and mortality in fish stocks[278].


A
spirotrichea
colpodea
nassophorea
cariacotrichea
armorphorea
phyllopharyngea
litostomatea
plagiopylea
heterotrichea
oligohymenophorea
karyorelicta
protocruzia
prostomatea
mesodiniae

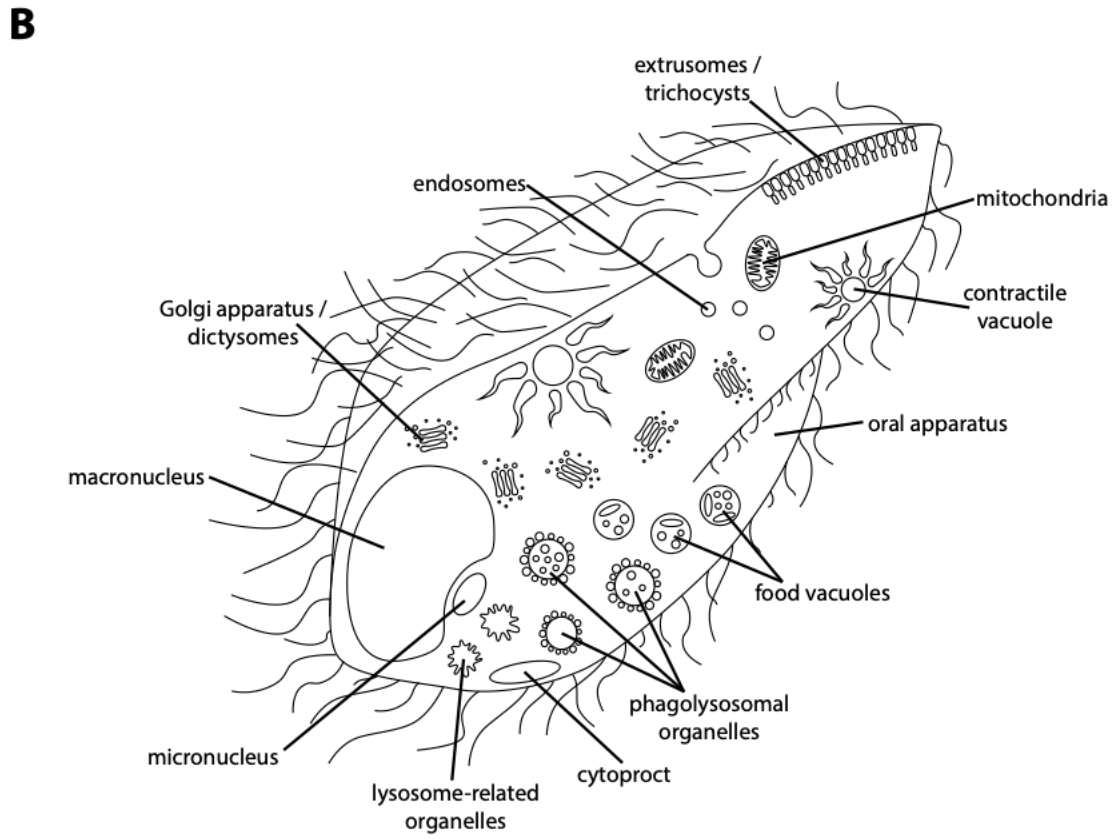

B
extrusomes /
trichocysts
endosomes
mitochondria
contractile
vacuole
Golgi apparatus /
dictysomes
oral apparatus
macronucleus
food vacuoles
phagolysosomal
organelles
micronucleus
cytoproct
lysosome-related
organelles

Figure 4.1: Diversity and cell biology of the phylum Ciliophora.
A: Phylogeny of ciliates based on Gao et al. (2015). Lineages with at least one genome or transcriptome representative are in black, while lineages with multiple representatives are in bold and have an expanded terminal node. **B:** Schematic of ciliate organelle diversity based on the morphology of the model ciliate *Paramecium caudatum*. Different membrane-bound organelles are described.

There is also increasing evidence that the microbiome of ciliates themselves provides a substantial contribution to their lifestyles in various environmental niches. As well as acting as epibionts, symbionts, and parasites, ciliates can have their own epibionts, symbionts, and parasites[279]. The epibiotic and symbiotic communities of ciliates often include bacteria which contribute to their survival in the local ecosystem; for example, sulphur-reducing bacteria are often found as epibionts of anaerobic ciliates[279]. Ciliates are highly important and highly relevant ecosystem and microbial community contributors in the bacterivorous niche.

4.2.2 Ciliates as bioindicators and in anthropogenically-influenced environments

The relative ease with which one can distinguish between different ciliates has made them popular candidates for bioindicators, and there are many studies on the distribution of ciliates within disturbed environments[238,268,280,281]. Ciliates are also amenable to many types of identification, and studies on the biodiversity of ciliates in anthropogenically influenced environments have used varied techniques such as microscopy, expressed sequence tagging, and environmental DNA (eDNA) barcoding[83,280,282]. There is also an equally impressive array of compounds and contaminants which have had their environmental impact assessed using ciliates: these include heavy hydrocarbons[283], heavy metals[284], eutrophication[285], naphthenic acids[286], salinity[287] and general urban pollution[281]. Most of these studies of ciliate communities in anthropogenically influenced environments conclude that there is a sharp decline in biodiversity in disturbed environments, and that biodiversity of ciliates is an excellent indicator of ecosystem health.

Though there is evidence of many species of ciliates having documented resistance to hydrocarbons in these bioindicator studies, the evidence is conflicting, and the mechanism of this resistance is unknown. The fact that species richness is significantly reduced in anthropogenically influenced environments suggests that some, if not most, species of ciliates are sensitive to these disturbances. However, many ciliate communities also show notable resistance to anthropogenic stresses and are essential to the health of disturbed ecosystems, as in wastewater and sewage. It is evident that there must be substantial variation in resistance to disturbance across the diversity of ciliates; however, the extent to which different lineages are resistant to disturbance, and the extent to which this correlates with either their ecological niche or phylogeny, is unknown. There have been some *in vitro* studies on the interactions between ciliates and hydrocarbons in the local

environment which have provided some speculation into the mechanisms of this interaction; for example, inhibition of protist movement (and therefore grazing on bacteria) appears to reduce hydrocarbon degradation[137]. In this study, the authors hypothesised that the observed effects may be due to: protists having a direct capacity to degrade hydrocarbons, the consumption of bacteria without degradation capacity allowing the overgrowth of bacteria with degrading capacity, or consumption of hydrocarbon-degrading bacteria keeping the bacterial growth in log phase and, therefore, requiring additional hydrocarbons as fuel for growth[137].

In the specific case of Base Mine Lake, the overwhelming abundance of heterotrophs identified in the early years of reclamation make ciliates an extremely interesting and attractive target as early bioindicators of reclamation success. So far, ciliate diversity in the year 2015 has been assessed in the paper Richardson et al. (2020)[235] and across the four summers of 2015-2018 in Chapter 3. These studies identified a single OTU belonging to the Litostomatean clade that dominated the ciliate diversity in 2015, exhibiting a bloom in the early summer season and relatively modest abundances outside of this bloom. The biodiversity of ciliates from 2015 to 2018 increased substantially, particularly in the Oligohymenophorea. Across the entire dataset from 2015-2018, Litostomatea OTU abundance was approximately equal to Oligohymenophorean OTU abundance. However, the abundance of individual ciliate lineages was still extremely variable and the overall heterotroph community of BML, including the ciliates, was not stable and not correlated with any of the expected seasonal or environmental variables such as depth; the substantial correlation across the dataset with the year of sampling also implied that this environment is still under considerable change. Ciliates may well act as good bioindicators in this system, either overall or in the responses of specific OTUs. However, due to two extreme environmental perturbations in 2016, the system will need to be followed for more years before any sort of causal relationship can be established.

Efforts to understand the overall effects of hydrocarbons on ciliate communities are also hampered by undocumented morphological and cell biological diversity within the ciliate phylum. The two model ciliates, *T. thermophila* and *P. tetraurelia*, are both from the clade Oligohymenophorea[269]. Though new ciliate models from outside this class (such as *Stentor coeruleus* from the Heterotrichea and *Oxytricha trifallax* from the Spirotrichea) are now more commonly being used,

the vast majority of the cell biological information known for ciliates has been obtained from the Oligohymenophorean class[269]. Given that these species represent one class out of the fourteen that comprise the ciliate phylum and the ciliates observed in eDNA studies or via microscopy represent a much larger group of classes, it is first essential to expand our understanding of the cell biology of these organisms across ciliate diversity.

The genomes of both *T. thermophila* and *P. caudatum* were sequenced in 2006[288,289]. As ciliates are so large and easy to isolate, the sequencing revolution that has allowed the expansion of eDNA studies into mainstream eukaryotic microbial ecology has also resulted in an exponential increase in the available genetic data in the ciliate phylum. There are now several published, annotated genomes from ciliates in three of the fourteen classes (Heterotrichea[290], Spirotrichea[266,267,291,292], and Oligohymenophorea[293,294]), and a variety of data from environmental studies, transcriptomes, and metagenomes[295,296]. This makes them an excellent candidate for further investigation, both for their potential as bioindicators and the potential for identifying novel diversity or their cellular adaptations to the extreme environment of the Athabasca Oil Sands.

4.2.3 Ciliate cell biology and the membrane trafficking system

As well as their cilia, ciliates are known for two other major cell biological innovations: their double nuclei, which have different roles within the cell and in gene expression, and their idiosyncratic membrane trafficking system, which includes the alveoli, small subcellular membranous sacs for which the Alveolata lineage is named[28,269].

The genetic organisation of ciliates is by far the better understood of these two phenomena. Ciliates have two nuclei, a macronucleus (MAC) and micronucleus (MIC). The MAC contains all of the expressed genes in thousands of copies of fragmented minichromosomes, while the MIC contains germline genetic information and is held in a highly condensed state outside of cell division[269]. The MAC must be reconstructed for gene expression after every cell division, which made ciliates an early model for examining condensation and structuring of nuclear DNA; the most notable example of this was the discovery of telomeres in *Tetrahymena thermophila* by Elizabeth Blackburn which earned her, along with Carol W. Grieder and Jack W. Szostak, a Nobel Prize[297]. The germline genomes of many ciliates, most notably in the Spirotrichea, have highly fragmented

and rearranged genes[292]. A single gene may be split into several sequences and scattered throughout the MIC, only reassembled when the new MAC is assembled after cell division. In the MAC, the situation is reversed; each gene is assembled into its coding sequence in often thousands of copies throughout the thousands of MAC chromosomes, leading to high copy numbers of almost every gene within the single cell[289]. This phenomena has implications both for comparative genomics and for eDNA assessments of protist diversity: in the case of comparative genomics, it can lead to the misidentification of copy numbers or false paralogues[288], and in the case of eDNA assessments, the large copy number of 18S rRNA genes may lead to an inflated abundance estimate of ciliates within a DNA sample[298]. The advent of high-throughput genomics and, in particular, single-cell genomics and transcriptomics techniques has led to a renewed interest in the genetic diversity and organisation of ciliates, and we now have information on the nuclear arrangements of both MIC and MAC in *O. trifallax*[299] and *T. thermophila*[300].

The membrane trafficking system of ciliates, another cell structure which has exhibited notable unique adaptations across this phylum, is far less studied outside of a handful of model organisms[301]. The membrane trafficking system (MTS) is, broadly, the machinery within a cell which transfers cargo from one membrane-bound compartment to another, usually mediated via membranous vesicles[302]. The MTS in eukaryotes has been well studied in model organisms[303] and via comparative genomics in many lineages[304]. However, new genomes, model systems, and cellular machinery are constantly being discovered, and there are still many novel adaptations within clades which have yet to be identified. One aspect of membrane trafficking complexity in the ciliates is a phylum-wide system of rapid regulated exocytosis from intracellular vesicles just below the plasma membrane[305,306]. These organelles, also collectively known as extrusomes, sit just under the cell surface and rapidly release their contents into the extracellular environment in response to an intracellular signal. The exact structure and contents of these organelles varies across ciliates, as does their function: for example, the best studied extrusomes are the mucocysts of *T. thermophila*[307] and trichocysts of *P. tetraurelia*[308]. Despite the relatively close relationship between these organisms given the extent of ciliate diversity, these homologous organelles have different morphology, structures, and associated genes[306]. However, both are integrated into the membrane trafficking system and appear to be related to lysosomes, another membrane-bound cellular component which has a role in recycling of cell contents[309].

4.2.4 The effects of hydrocarbons on ciliate cell biology

The earliest studies of hydrocarbons interacting with ciliate cell biology, which kickstarted the use of ciliates as bioindicators in later studies, took place in the early 1980s. Electron microscopy of ciliates was heavily used to examine the cellular structures of ciliates affected by various compounds, including hydrocarbons. In Rogerson and Berger (1982)[135], crude oil was placed on the ciliate *Colpoda colpodium*, and changes to its cell biology were observed via electron microscope. After the addition of crude oil, the membrane structure of the cell showed considerable morphological deformation. In particular, membrane-bound vesicles which seemed to contain a lipidlike structure became evident in the cell, and the vesicles themselves were larger and less circular than the vesicles in the undisturbed cell[135]. In this case, addition of hydrocarbons obviously changed the membrane trafficking system of the cell; though in this particular paper the resulting effect on the survival rate of ciliates was not tested, other work by the same authors indicated that exposure to high concentrations of crude oil resulted in significant mortality in this lineage and other cultures from across ciliate diversity, with oxidative stress suggested as the likely mechanism of toxicity[310]. Gomiero et al. (2013)[311] tested the effects of polluted sediment, which included high quantities of polyaromatic hydrocarbons (PAHs), on the ciliate *Euplotes crassus*. PAHs are smaller-chain, volatile hydrocarbons, which are a constituent of crude oil along with the longer-chain hydrocarbons used for fuel and heavy bitumen. In this study, they observed substantial toxicity at high concentrations of these contaminants. They also observed sublethal effects of hydrocarbon contamination such as reduced mobility and cell division, which they attributed to diversion of energy from these processes into cellular detoxification. They noted that hydrocarbon contamination also influenced the membrane trafficking system of *E. crassus*, with upregulated endocytosis and food vacuoles resulting in an increased overall cellular pH. The authors also noted an effect on the integrity of the stability of the lysosomal membrane, which also indicates a role in the membrane trafficking system[311].

In *Tetrahymena thermophila* (also known in some literature as *Tetrahymena pyroformis*, a historical name for the species) some of the resistance mechanisms for PAHs have been elucidated. Bamdad et al. (1998 and 1999)[286,312] identified an efflux pump in the outer membrane of the ciliate cell which, when inhibited, led to an accumulation of PAHs in the cell cytoplasm. This implies that *T. thermophila* has as active resistance mechanism for small hydrocarbons; though they can

enter the cell, they are actively removed via this efflux machinery. The authors noted a resistance to PAHs in *T. thermophila*, while other studies have noted that contamination with PAHs can result in toxic and sublethal effects to individual lineages of ciliates[311], and environmental studies have also noted than exposure to PAHs such as naphthenic acids cause overall reductions in biodiversity[83].

Gilbert et al. (2017)[138] also experimentally exposed ciliates to crude oil, which contains longer-chain hydrocarbons as well as the PAHs; however, their interest was in the resulting distribution of the oil as opposed to the effects on the ciliate cell. They found that addition of ciliates to a hydrocarbon-polluted dish resulted in smaller lipid droplets distributed more evenly within the water column, which suggested a potential use in remediation of oil spills; the current best practice for dispersing crude oil settled on the surface of the water is the use of chemicals such as Corexit. Though these chemicals prevent the deaths of macrofauna, they have ecologically devastating effects on the local microbial populations, including ciliates[313]. If ciliates can act as biological dispersants, this would reduce the reliance on chemicals to manage oil spills. As ciliates exhibit grazing behaviour, which involves swimming through these droplets, and as their cilia are able to trap microscopic quantities of this oil, they may be able to physically disperse larger oil droplets without themselves being harmed[138].

The evidence for the effects of hydrocarbons on ciliate cell biology is clearly very lineage-specific; noted effects include lethal and sublethal toxicity, ultrastructural and morphological deformation, and resistance via efflux and natural dispersal[135,138,286]. It is evidently important to understand the extent of ciliate cell biology in order to identify potential mechanisms of resistance when bioprospecting in extreme environments and when identifying species that may potentially be useful as bioindicators. Most of the currently available data on the effects of hydrocarbons on ciliates also implicates the membrane trafficking system in the effects of hydrocarbon contamination; this system is known to have substantial natural diversity across the ciliate phylum[301,306]. It is therefore important to understand this natural diversity before specific changes detected by environmental studies can be classed as significant to the biology of the affected area.

4.2.5 Scope of this chapter

In this chapter, I use comparative genomics and phylogenetics to assess the baseline diversity of the membrane trafficking system across the ciliate phyla. Using genome and transcriptome data obtained from model lineages of ciliates as well as environmental studies, I assess the completeness of the genes necessary to build various aspects of the membrane trafficking system, including coat complexes, tethers, and adaptor complexes. I identify a mostly complete membrane trafficking gene complement and relatively consistent distribution across lineages. However, some protein complexes appear to be patchy or entirely missing across ciliate diversity, including AP3, Dsl1, exocyst, COG, and TRAPPII. I discuss the potential functional consequences of the loss of each of these components and use phylogenetics and literature searches to determine what the likely mechanisms for loss and downstream effects on cell structure may be. Additionally, the HOPS subunit, responsible for membrane trafficking to lysosomes and late endosomes, appears to be lost in some ciliate lineages. Using phylogenetics combined with functional data, my collaborators and I identify a mechanism of diversification and functional compensation in ciliates that represents a previously unknown mechanism of evolutionary plasticity.

4.3 Methods

4.3.1 Transcriptome cleanup protocol

A table of the genomes and transcriptomes used in these analyses and their provenances is given in Supplementary Table S4.1. Two of these transcriptomes are currently unpublished; these were obtained from Dr. William Bourland at Boise State University. Transcriptomes were removed based on the following criteria: too few BLAST hits to the adaptin complexes, a conserved group of canonical membrane trafficking components, or prey organisms too closely related to the target organism to be detected using a high-throughput homology search. To ensure that nonciliate queries are not included in the final dataset (as many of the transcriptomes are contaminated with other eukaryotic sequences), all sequences defined as potential hits through an initial round of homology searching are also searched, via BLAST, in a database of representative organisms across eukaryotic diversity. Any sequences which did not have a ciliate top hit from this BLAST search were discarded.

4.3.2 Homology searching

For the comparative genomic detection of MTS homologues across ciliate diversity, an annotated query dataset from *Homo sapiens* was used for queries. Translated ORF coding sequences from the genomes and transcriptomes of selected ciliate species were used as the target dataset into which the queries were searched using the BLAST algorithm[251]. Positive BLAST hits against each query were those with an E value of less than 0.05, for which reciprocal BLAST against the genome containing the query sequence retrieved either the same sequence or an isoform of the sequence with the same E value or lower. To identify a hit as orthologous to the query, we further required that the E value be at least three orders of magnitude lower than the next lowest hit. Hits that were consistent with the first two criteria but did not show clear superiority over other hits were classified as potential hits, which may be either the query protein or a close homologue. Custom scripts were used to assemble the raw data obtained from reciprocal BLAST searches and the associated statistical data into a usable format (Appendix 3). To ensure that duplicated machinery were not simply multiple fragments of genes, hits that were less than half of the length of the queries were discarded.

To ensure that all possible putative queries were identified from the genomes and transcriptomes, Hidden Markov Models (hmms)[314] were constructed for each protein containing confidently identified sequences as well as other sequences from within the Alveolates. The accessions of the sequences used for hmm constructions are in Appendix 3. These hmms were used to search the genomes for additional sequences. For the transcriptomes, BLAST searches were repeated using confidently classified ciliate hits. Additionally, the scaffolds for each of the genomes were searched using the genome-specific queries and those of closely related species to ensure that no nonannotated proteins were missed in the coding and amino acid files.

4.3.3. Phylogenetic tree construction

Once homologues of the proteins in the MTS had been determined via BLAST[251], the phylogenetic relationships of the patchy homologues were determined using both maximum likelihood (via the RAxML algorithm[242]) and Bayesian inference (via the MrBayes algorithm[315]) run using the CIPRES server[316]. Homologues were aligned using MUSCLE[317] and alignments were manually trimmed to retain regions of unambiguous homology. For the RAxML trees, iqTREE[318] was used

to select an appropriate rate evolution model. Consensus trees were constructed through manual inspection of both RAxML and MrBayes outputs and determination of corresponding nodes, mapped onto the MRBayes topology.

4.3.4 Domain analysis

Domain analysis was carried out in Pfam[319]. Putative hits were searched into the Pfam database, and those with the appropriate domain were extracted and used for phylogenetic analysis as described in Section 4.2.3.

4.4 Results

4.4.1 Cleanup of transcriptome datasets

The transcriptome datasets used in this chapter were obtained from various publicly available sources, including the Marine Microbial Eukaryote Sequencing Project , phylogenomic analyses[321], or other cell process-specific analyses[322]. Because of this, I first sought to establish which transcriptomes were sufficiently complete to use for a whole-genome membrane trafficking survey. I used the adaptin proteins for this initial survey as they are extremely well conserved within the eukaryotes and easy to retrieve using BLAST even with extremely divergent queries. The summary of my results can be viewed in Supplementary Table S4.1. I discarded the transcriptomes where a forward BLAST retrieved hits for less than 10% of the known protein diversity of the adaptin family (2 or fewer of the 20 proteins), as these were likely to be too incomplete to provide useful data. This protocol led to the discard of *Anophyroides haemopilia, Entodinium caudatum,* and *Polyplastron sp.*, which had 2, 1 and 1 detected adaptins, respectively.

After removing extremely incomplete transcriptomes, I determined which transcriptomes were likely to require additional cleanup protocols and whether the methodology I had established for this (described in 4.2.1) would lead to meaningful results. As most ciliates were fed on bacterial prey or prey that was not closely related to themselves (such as chlorophytes), the only transcriptome discarded in this stage of the analysis was *Litonotus sp.* This ciliate was fed on another ciliate, *Euplotes crassus*, before sequencing, and as my experimental technique relied upon the extraction of specific transcripts rather than any further assembly, it would be impossible to

determine which transcripts belonged to *Litonotus sp.* and which belonged to *E. crassus* using the protocol described in 4.2.1.

The summary of the percentage of putative hits determined to be the result of contamination can be seen in Supplementary Table S4.1. The amount of contamination is extremely variable; while most transcriptomes had less than 10% nonciliate sequences, *Mesodinium pulex* and *Platyophrya macrostoma* both have over 40% nonciliate sequences. These sequences were removed before further processing or annotation of any of the putative hits obtained from these datasets.

4.4.2 Heterotetrametric adaptin complexes

The heterotetrameric adaptin complexes (HTACs) are a group of proteins that function in the vesicle formation stage of membrane trafficking, and each HTAC, generally, localises to a different set of organelles on the membrane trafficking pathway[323]. In this study, I surveyed the adaptins (AP1, 2, 3, 4 and 5), and TSET (Figure 4.2). These complexes are necessary for cargo selection and vesicle formation during membrane transport[324]. I also surveyed the structural components of vesicles, COPI, COPII, and clathrin, which act alongside the HTACs in vesicle formation and budding[323]. As different combinations of these protein complexes are associated with vesicles at different areas of the cell and different membrane trafficking processes, surveying the overall diversity of HTACs and coat complexes provides a useful overview of probable organellar diversity. The results of this survey are summarised in Figure 4.2, a Coulson plot which indicates the presence, absence, and copy number of the HTACs and coat complexes.

Of the adaptins, AP1, 2 and 4 are generally conserved in all the sampled ciliate classes. AP5 and TSET are almost completely absent, save a few retained subunits. AP3 appears to have a patchier distribution; it is generally present in the Heterotrichea and some of the other species, such as *O. trifallax, Mesodinium pulex, Platyophyra macrostoma, Paramecium caudatum,* and *Tetrahymena thermophila* also have a full AP3 complement. However, most species do not have a complete AP3, including many of the genome assemblies which would not be missing coding sequences simply because of lack of expression. The structural vesicle components, including clathrin, also appear to be reduced or missing. However, heavy-chain clathrin was detected in every genome and transcriptome, often in multiple copies. COPI and COPII were generally complete, especially

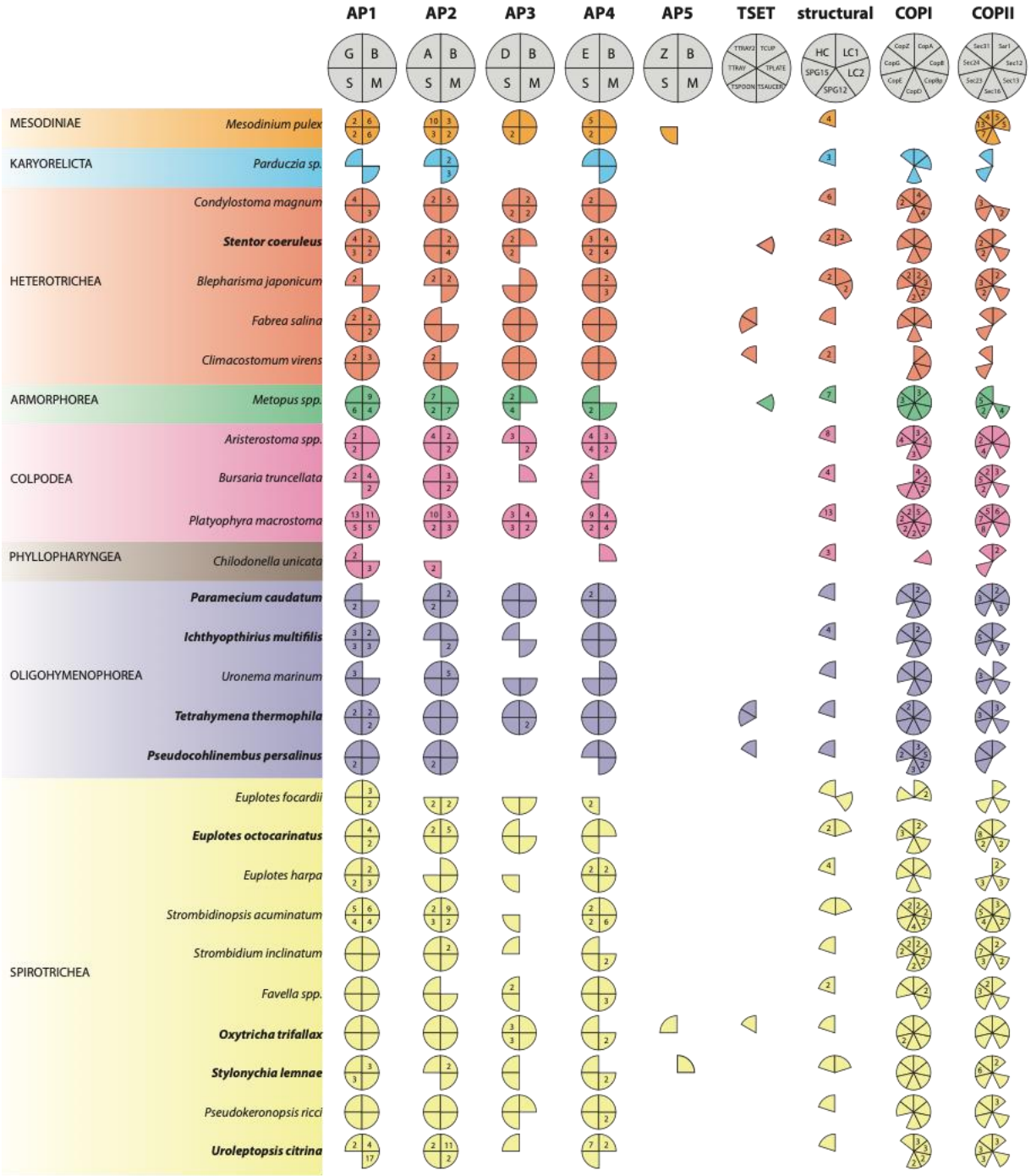


Figure 4.2: Coulson plot of heterotetrameric adaptin complexes across ciliate diversity.

Coloured sections indicate presence of the protein, while numbers on the segment indicate how many paralogues of that protein were identified. Ciliate classes are separated by colour, and bolded taxa are genomes.

in the genome assemblies, missing only Sec12 and Sec16 in most species. These genes are often lost in numerous taxa, so this is consistent with the known distribution of the proteins.

Some genomes appear to have an overall expansion in the membrane trafficking machinery. The ciliate genomes which appear to have generally expanded protein machinery in the sampled species are *Stentor coeruleus* and *Uroleptopsis citrina.* Other species appear to have expanded only specific protein complexes. For example, the transcriptomes *Strombidium inclinatum, Blepharisma japonicum,* and *Bursaria truncellata* have expanded COPI and COPII machinery but a canonical number of adaptin subunits. Though several subunits from otherwise conserved complexes were undetected in this annotated genome assembly, *Pseudocohnilembus persinalis* appears to have duplications in the COPI machinery.

From these results, distribution of the AP3 adaptin complex appeared to be less well conserved than AP1, 2, and 4 across ciliates. To determine whether this was a potential loss of AP3 function across the ciliate clade as a whole, I carried out a phylogenetic analysis of the four subunits of the adaptins (beta, mu, sigma, and epsilon/gamma/alpha/delta/zeta or EGADZ) in the nine sampled genomes (*Euplotes octinarius, Ichthyopthirius multifilis, Oxytricha trifallax, Paramecium caudatum, Pseudocohlinembus persinalis, Stylonychia lemnae, Stenor coeruleus, Tetrahymena thermophila,* and *Uroleptopsis citrina*).

These trees, shown in Figure 4.3A, B, C, and D, show resolution of each subunit into the AP1, 2, 3, and 4 subunits with strong bootstrap and probabilistic support. Any expansion of paralogues in any given organism appear to be lineage-specific; there is no evidence of any ancestral duplications of any of these proteins. However, in some cases, there are multiple duplications; for example, *Stentor coeruleus* has four copies of each of the mu subunits, all closely conserved.

Conversely, almost all the AP5 subunits have been lost in every genome. The only exception is an AP5Z in *O. trifallax* and an AP5B in *S. lemnae*. These sequences form long-branching outgroups to both the beta subunit and EGADZ subunit trees, excluded from the other adaptin clades. They were subsequently removed from the final figures as the long branches affected the resolution of the rest of the tree. However, the full trees including AP5 can be found in Appendix IV as Supplementary Figures S4.5 and S4.6.

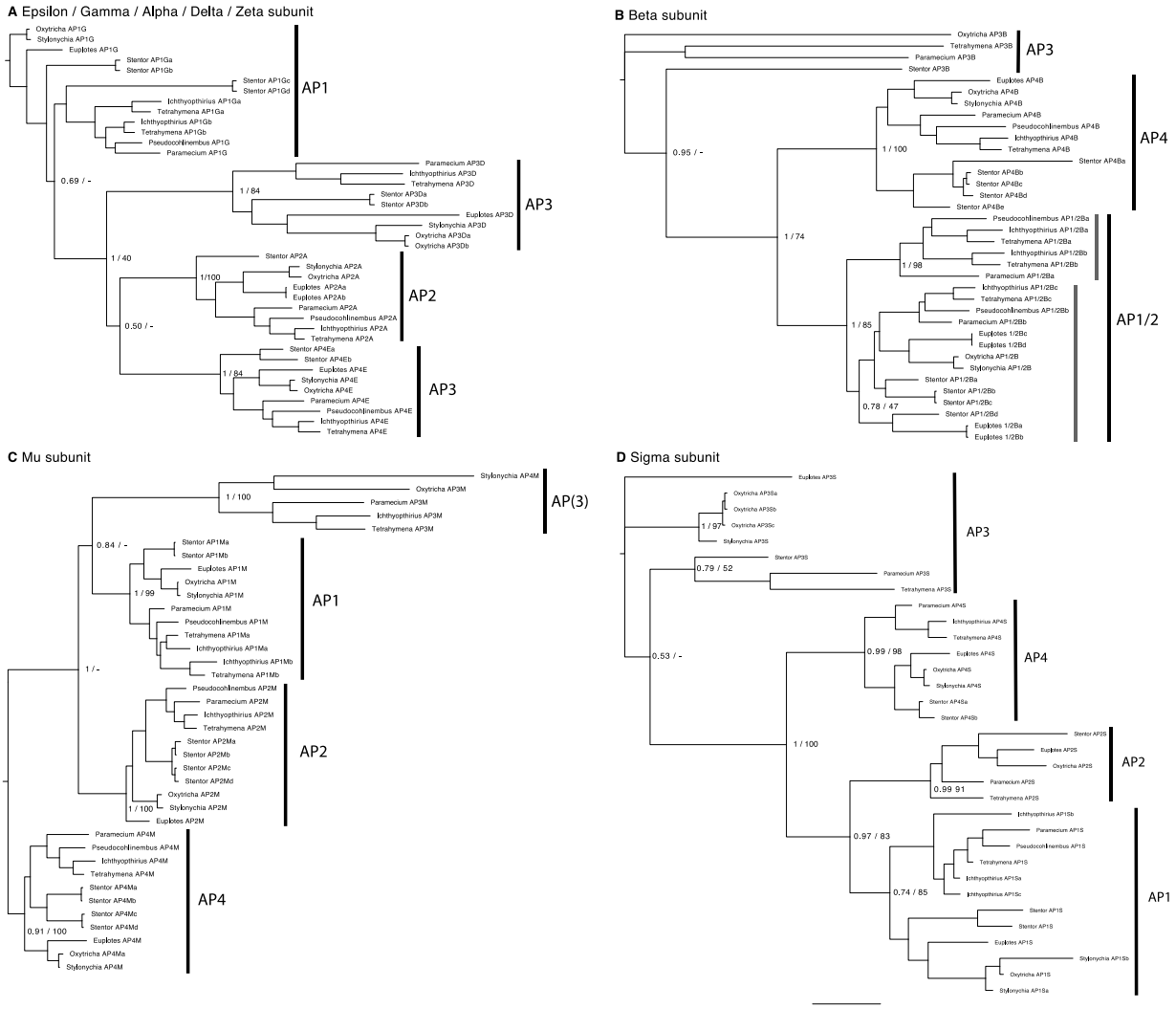


Figure 4.3: Phylogenetic analysis of adaptin subunit distribution across ciliates. Bootstrap values are based upon RAxML, and probability values are based upon MrBayes. All adaptin subunit expansions appear to be lineage-specific, and AP3 subunits in every tree are long-branching.

In each of the subunit trees, AP3 subunits branched with strong support from the other adaptins, with comparatively long average branch lengths to the other adaptins. This is indicative of divergent sequences, which implies that there is either strong selection on the AP3 complex away from the other subunits or relaxed selection on the complex. Strong diversifying selection of the entire complex would usually result in a longer branch at the base of the AP3 grouping; however, generally the leaf nodes for each species were long-branching. This implies that there is altered selection on the AP3 subunits compared to the other adaptin complexes.

4.4.3 Endocytic machinery

The endocytic machinery is responsible for mediating vesicle trafficking of extracellular cargo from the outside the plasma membrane into the membrane-bound compartments of the cell[325]. This is a complex process involving extracellular receptors to initiate vesicle formation on the plasma membrane, structural components for vesicle formation and budding, and cargo adaptors and tethering complexes to ensure correct trafficking of different types of vesicles to various intracellular compartments[325]. Endocytosis is a potential site of resistance mechanisms to hydrocarbon contaminants in ciliates[286], and so I analysed the distribution of endocytosis-specific machinery across ciliate diversity (Figure 4.4). Overall, there is also little correlation between presence of endocytic machinery and the phylogenetic derivation of the sampled ciliates, apart from in the Spirotrichea. This class of ciliates have some of the most highly derived genetic traits of the ciliates and are an extremely divergent lineage, but they contain the two species exhibit the most conserved endocytic machinery: *Stylonychia lemnae* and *Oxytricha trifallax*. However, the related ciliate *Uroleptopsis citrina,* which is also a member of the Spirotrichea, has very little of the endocytic machinery detected despite having a newly sequenced and well annotated genome. The distribution of these components across the sampled ciliates is somewhat uneven; while some components, such as AP180, Dab2 and NPR appear to be entirely undetectable or highly reduced, other components are substantially expanded. The EP15R protein, which is involved in production of clathrin-mediated pits on the cell surface[326], appears to have been expanded in all ciliate genomes. Fab1 and EpsinR are similarly expanded in some lineages, though less substantially than Ep15R. Fab1 and EpsinR are also lost in some genomes—most notably in the parasitic ciliate *Ichthyopthirius multifilis*.

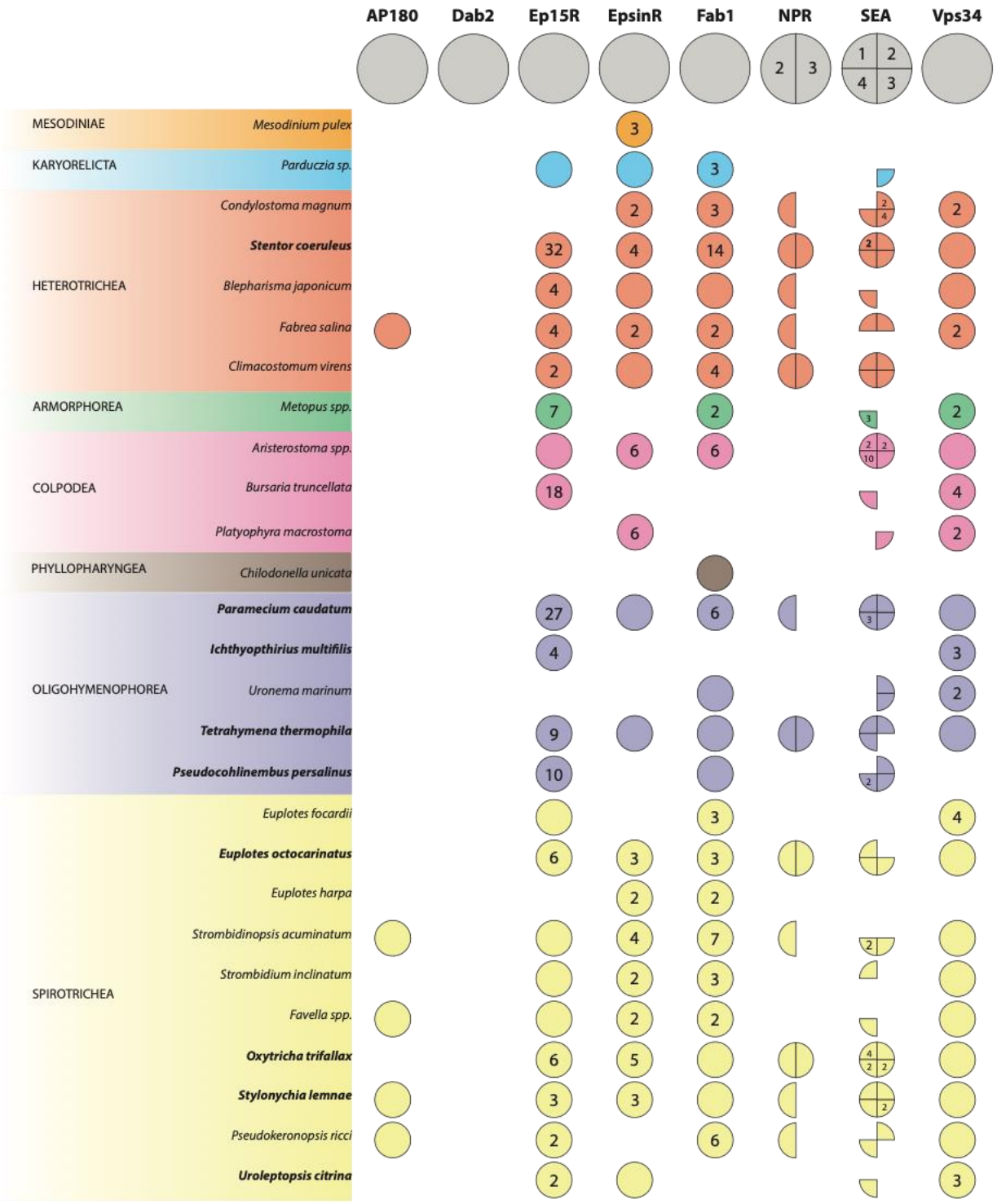


Figure 4.4: Coulson plot of endocytic machinery across ciliates.
Coloured sections indicate presence of the protein, while numbers on the segment indicate how many paralogues of that protein were identified. Ciliate classes are separated by colour, and bolded taxa are genomes.

Conversely, Ep15R was detected in every genome. It appeared to be substantially expanded across multiple ciliate lineages from multiple classes; for example, 18 copies were detected in the partial genome *Bursaria truncellata* from the class Colpodea, 27 in the genome *Paramecium caudatum* from the class Oligohymenophorea, and 32 in the genome *Stentor coeruleus* from the class Heterotrichea. This suggests that there was either an ancestral duplication or expansion of this protein before the diversification of these ciliate classes, or extreme lineage-specific expansions of this protein on multiple occasions. However, from the BLAST results alone, it is difficult to distinguish genuine putative hits from putative hits that arise from common domain structure or sequence motifs rather than genuine homology. Some protein domains, due to biophysical constraints or functional convergence, recur across all organisms in homologous and nonhomologous proteins. In the case of EP15R, the annotated protein is known for repeated EF-hand domains that are essential for protein function[326], and I hypothesised that these conserved domains were causing an artefact in the BLAST results and an inflation of the actual results. Accordingly, I carried out a domain analysis for each protein to determine which were more likely to be the actual putative hit using Pfam to annotate the domains of each putative EP15R. I found that the vast majority did not have any detectable domain architecture, so I restricted my downstream analysis only to hits that had at least one annotated EF-hand domain; this was 12% of the total putative hits (32 out of 263). On a tree rooted on *Emiliania huxleyi*, a haptophyte, the restricted EF-hand only dataset showed that only lineage-specific duplications were present, with the new largest total of potential expansions as 4 in *Mesodinium pulex* (Figure 4.5). This is within the expected range of protein duplication in a comparative genomics survey, and no longer merited further investigation. However, this analysis did identify two potential ancestral duplications of the Eps15R protein: one at the base of Heterotrichea (Heterotrichea A and B in Figure 4.5) and one at the base of Spirotrichea (Spirotrichea A and B in Figure 4.5). These duplications are not extremely well supported, but this differs from the lineage-specific expansions most observed in Figure 4.2.

4.4.4 Comparative genomics in multisubunit tethering complexes

Tethering complexes are an important part of membrane trafficking and act in the docking stage of vesicle movement throughout membrane-bound cellular compartments[327]. Like the heterotetrameric adaptin complexes, different multisubunit tethering complexes (MTCs) act at

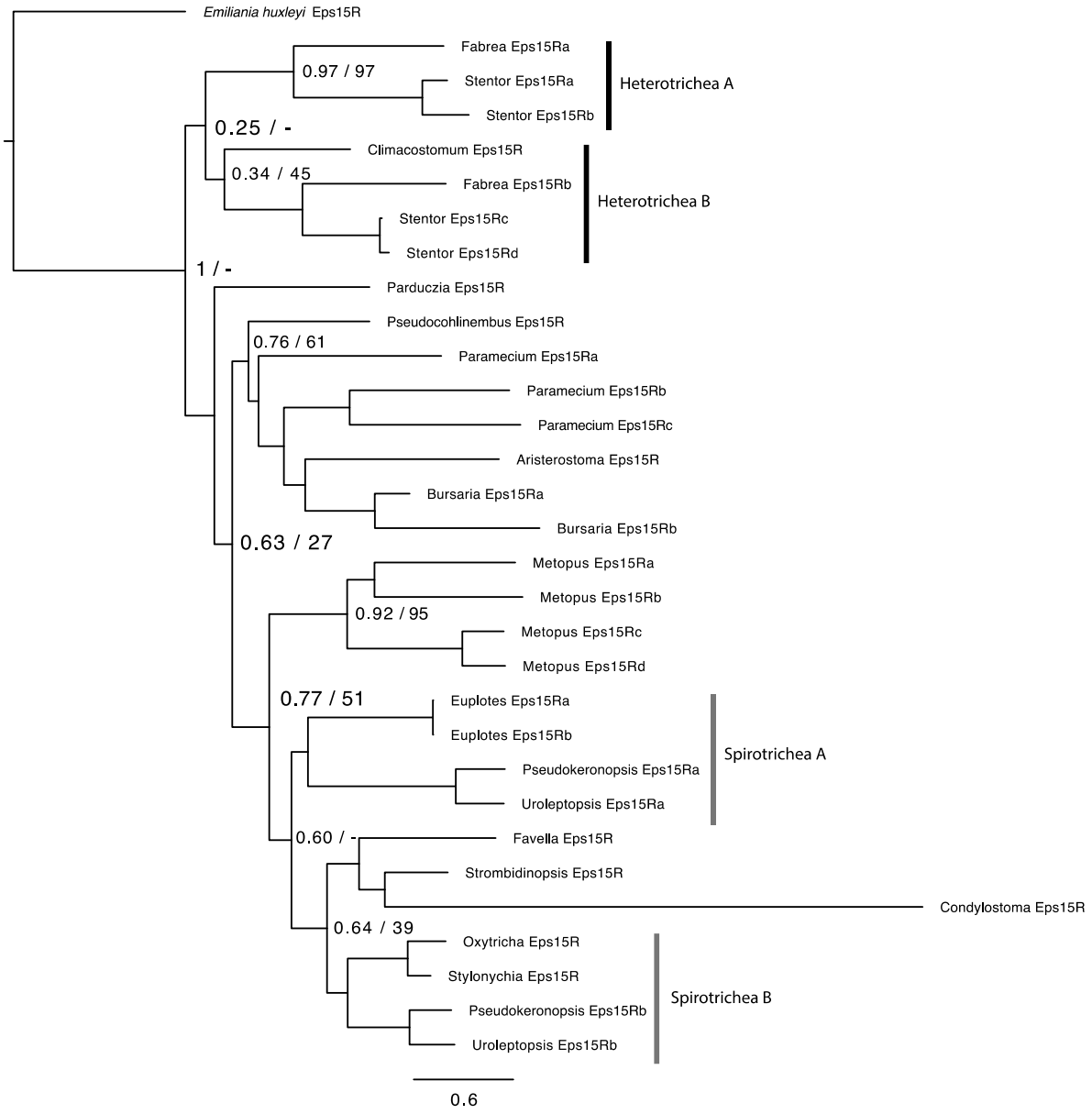


Figure 4.5: Phylogenetic tree of EF-hand containing Eps15R hits across ciliates. Bootstrap values are RAxML, while probability values are MrBayes. Most Eps15R diversity is lineage-specific, though there are two potential (but poorly supported) early duplications in the evolutionary history of the Heterotrichea and the Spirotrichea.

different subcellular compartments[328]. The patterns of gains and losses of these components can act as an indicator of whether there are any divergent aspects to the compartments themselves; for example, if all of the proteins involved in trafficking to the Golgi body are undetected in sampling multiple genomes, it is reasonable to assume that there may be divergence in the morphology or other proteins involved in Golgi structure in that organism. The distribution of MTCs across this ciliate dataset is shown in Figure 4.6. Overall, the MTCs appear to have the most missing sequences from each of these protein groups. Of the surveyed complexes, only GARP, TRAPPI, and the HOPS/CORVET core proteins appear to be generally complete in most of the annotated genomes. Even then, there are notable absences; most TRAPPI subunits are undetected in the two genomes, *Ichthyopthirius multifilis* and *Uroleptopsis citrina,* and several GARP are similarly undetected in *Pseudocohlinembus persinalis*.

The other MTCs show substantial apparent loss of components, across the diversity of genomes and transcriptomes. There are many explanations for a missing gene sequence in a dataset. In the case of the transcriptome data, this could simply be a low level of gene expression preventing the gene from being detected. However, genome assemblies, which extract all DNA sequence and try to detect all coding sequences, usually have a more comprehensive gene complement. In these cases, if a protein sequence is undetected in a BLAST search, even considering hmm and scaffold searches, this usually indicates either a loss of that sequence from the genome or substantial divergence from the canonical sequence. Expanding upon the known cell biological roles of these lost complexes, in the remainder of this section, we have examined the potential consequences of this loss on the complexes Dsl1, HOPS, and CORVET.

Dsl1 is almost totally missing in all of the genomes and transcriptomes. This is notable because Dsl1 is the complex that is necessary for biogenesis of peroxisomes, the membrane-bound organelles responsible for recycling of reactive oxygen species and other potentially damaging cellular components[329]. Generally, protists that do not have a functional Dsl1 complex do not have peroxisomes[329]. However, peroxisomes are well described and characterised in the model organism *Tetrahymena thermophila*, and the peroxisomal complement of closely related organisms such as Apicomplexa has been studied in detail[330,331]. We undertook a BLAST-based

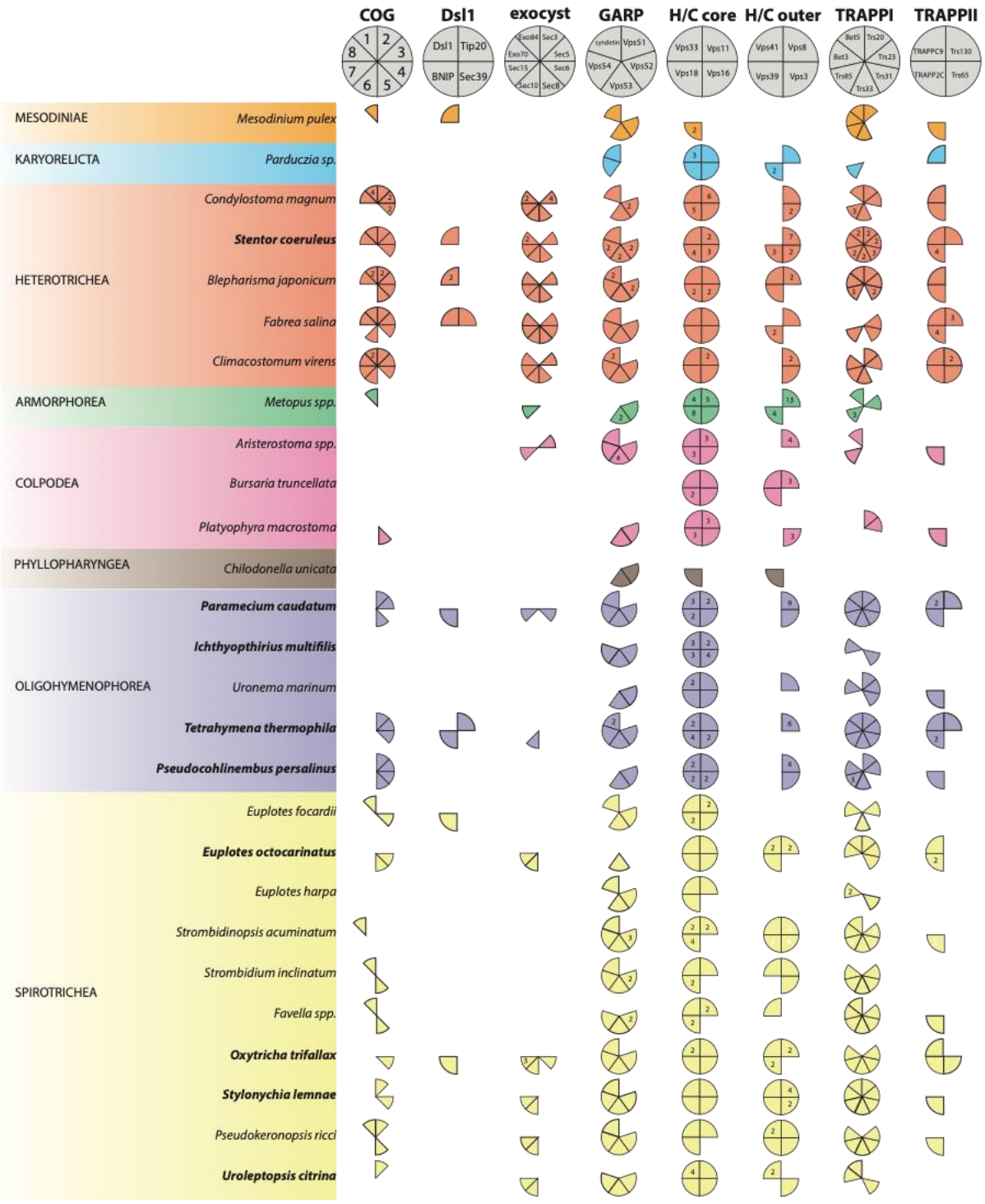


Figure 4.6 Coulson plot of multisubunit tethering complexes across ciliates. Coloured sections indicate presence of the protein, while numbers on the segment indicate how many paralogues of that protein were identified. Ciliate classes are separated by colour, and bolded taxa are genomes. Bolded pie segments indicate subunits which were only detected via further searching (by ciliate-specific BLAST in the transcriptomes and hidden Markov modelling in genomes).

search of each of the genomes and transcriptomes for other peroxin-associated proteins to determine whether peroxins appear to be present or absent across ciliate diversity.

The dot plot shown in Figure 4.7 shows that the majority of conserved peroxisomal proteins are detected across the sampled ciliate genomes. The machinery appears to be fairly complete and well conserved when compared to canonical peroxisomal proteins and *T. thermophila,* the model ciliate with characterised peroxisomes. This, along with the evidence from the literature, suggests that despite the near-total putative absence of trafficking machinery usually associated with cargo transport into and out of peroxisomes, peroxisomes are functional components of ciliates. An additional search for peroxin targeting sequence 1 (an SKL motif at the C-terminal end of the protein) in proteins from the sampled genomes identified in each genome peroxisome-associated genes with the SKL motif (Table 4.1). A complete list of proteins identified with SKL motifs and their BLAST results is in the Appendix for this chapter. This suggests that even in the ciliate species for which peroxisomes have not been identified via microscopy, they are likely present in the cell.

While Dsl1 is lost consistently in all ciliate lineages, components of the HOPS complex are lost in the Oligohymenophorea. HOPS/CORVET is a multifunctional MTC. The core subunits are conserved across both trafficking pathways with which they are associated, but the HOPS subunits are associated with late endosomes and lysosomal formation, and the CORVET subunits are associated with early endosomes, successive stages of the endosomal maturation pathway[332]. However, in the Oligohymenophorea, the HOPS subunits, *Vps39* and *Vps43*, are not detected. Conversely, the CORVET subunit *Vps8* appears to be expanded. *Vps8* is represented by a single gene in most organisms; in contrast, six *VPS8* paralogs exist in *T. thermophila*, and other ciliates show expansion in VPS8, notably *P. caudatum* and *S. coeruleus*. The question then arose of how trafficking to lysosomes was possible in ciliates without a HOPS complex, and whether there was any functional diversification in the expanded CORVET components. Due to the sophisticated genetic tools available in the Oligohymenophorean ciliate *T. thermophila*, it was feasible to test the effects of these proteins *in vivo.*

Our phylogenetic analysis showed that at least four paralogs were maintained since the ancestor of the Oligohymenophorean ciliate lineage (Figure 4.8A). The root of the tree was determined

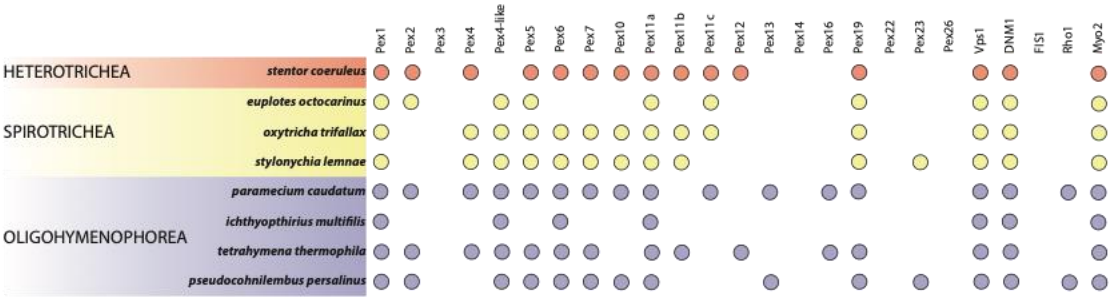


Figure 4.7: Dot plot of peroxisomal components in ciliate genomes with annotated coding sequences.

Table 4.1: Canonical PTS1 (SKL) motifs across the diversity of ciliate genomes.

Species	Class	PTS1 sequences
Bursaria truncellata	Colpodea	12
Euplotes octinarius	Spirotrichea	25
Ichthyopthirius multifilis	Oligohymenophorea	12
Oxytricha trifallax	Spirotrichea	23
Paramecium caudatum	Oligohymenophorea	14
Pseudocohlinembus persinalis	Oligohymenophorea	7
Stentor coeruleus	Heterotrichea	31
Stylonychia lemnae	Spirotrichea	24
Tetrahymena thermophila	Oligohymenophorea	24

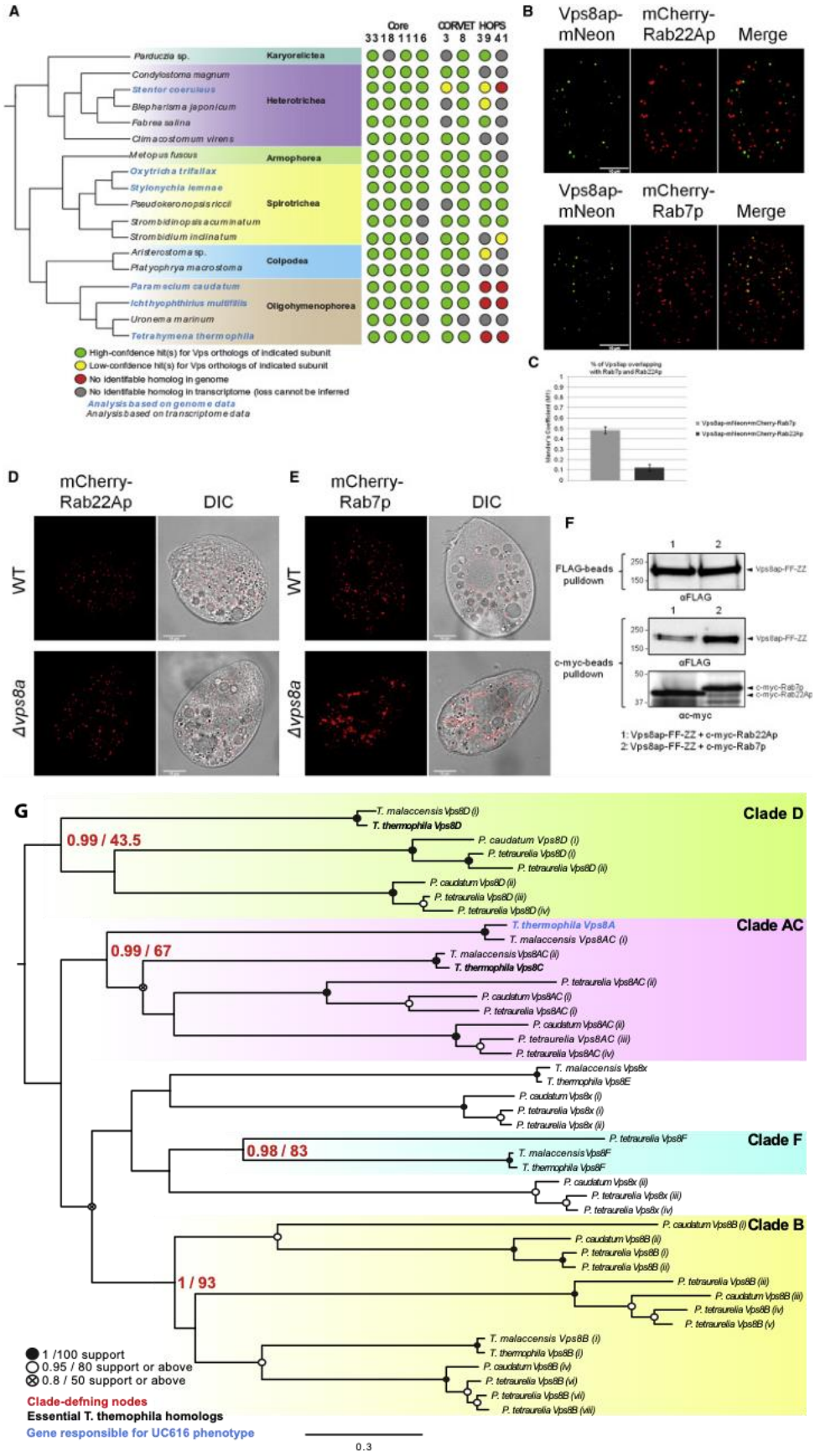

A
Core
CORVET
HOPS
33 18 11 16
3 8
39 41
Parduczia sp.
Karyorelictea
Condylostoma magnum
Stentor coeruleus
Blepharisma japonicum
Fabrea salina
Climacostomum virens
Heterotrichea
Metopus fuscus
Armophorea
Oxytricha trifallax
Stylonychia lemnae
Pseudokeronopsis riccii
Strombidinopsis acuminatum
Strombidium inclinatum
Spirotrichea
Aristerostoma sp.
Platyophrya macrostoma
Colpodea
Paramecium caudatum
Ichthyophthirius multifiliis
Uronema marinum
Tetrahymena thermophila
Oligohymenophorea
High-confidence hit(s) for Vps orthologs of indicated subunit
Low-confidence hit(s) for Vps orthologs of indicated subunit
No identifiable homolog in genome
No identifiable homolog in transcriptome (loss cannot be inferred
Analysis based on genome data
Analysis based on transcriptome data
B
Vps8ap-mNeon
mCherry-Rab22Ap
Merge
Vps8ap-mNeon
mCherry-Rab7p
Merge
C
D
mCherry-Rab22Ap
DIC
WT
Δvps8a
E
mCherry-Rab7p
DIC
WT
Δvps8a
F
FLAG-beads pulldown
Vps8ap-FF-ZZ
αFLAG
c-myc-beads pulldown
Vps8ap-FF-ZZ
αFLAG
c-myc-Rab7p
c-myc-Rab22Ap
αc-myc
1: Vps8ap-FF-ZZ + c-myc-Rab22Ap
2: Vps8ap-FF-ZZ + c-myc-Rab7p
G
Clade D
0.99 / 43.5
T. malaccensis Vps8D (i)
T. thermophila Vps8D
P. caudatum Vps8D (i)
P. tetraurelia Vps8D (i)
P. tetraurelia Vps8D (ii)
P. caudatum Vps8D (ii)
P. tetraurelia Vps8D (iii)
P. tetraurelia Vps8D (iv)
Clade AC
T. thermophila Vps8A
T. malaccensis Vps8AC (i)
0.99 / 67
T. malaccensis Vps8AC (ii)
T. thermophila Vps8C
P. tetraurelia Vps8AC (ii)
P. caudatum Vps8AC (i)
P. tetraurelia Vps8AC (i)
P. caudatum Vps8AC (ii)
P. tetraurelia Vps8AC (iii)
P. tetraurelia Vps8AC (iv)
T. malaccensis Vps8x
T. thermophila Vps8E
P. caudatum Vps8x (i)
P. tetraurelia Vps8x (i)
P. tetraurelia Vps8x (ii)
Clade F
P. tetraurelia Vps8F
0.98 / 83
T. malaccensis Vps8F
T. thermophila Vps8F
P. caudatum Vps8x (ii)
P. tetraurelia Vps8x (iii)
P. tetraurelia Vps8x (iv)
Clade B
P. caudatum Vps8B (i)
P. caudatum Vps8B (ii)
P. tetraurelia Vps8B (i)
P. tetraurelia Vps8B (ii)
1 / 93
P. tetraurelia Vps8B (iii)
P. caudatum Vps8B (iii)
P. tetraurelia Vps8B (iv)
P. tetraurelia Vps8B (v)
T. malaccensis Vps8B (i)
T. thermophila Vps8B (i)
P. caudatum Vps8B (iv)
P. tetraurelia Vps8B (vi)
P. tetraurelia Vps8B (vii)
P. tetraurelia Vps8B (viii)
1 /100 support
0.95 / 80 support or above
0.8 / 50 support or above
Clade-defning nodes
Essential T. themophila homologs
Gene responsible for UC616 phenotype
0.3

Figure 4.8: Functional and bioinformatic characterisation of Vps8 in *Tetrahymena thermophila* (Experiments B, C, D, E, F performed by Dr. Sparvoli).
A: Distribution of HOPS/CORVET subunits across ciliate diversity. Relationships between species are depicted by the evolutionary tree on the left (not to scale), with phylogenetic class indicated in bold. The dot plot on the right depicts the presence/absence of the indicated subunit within each species. **B:** Cells were transformed to coexpress mNeon-tagged *Vps8ap* at the endogenous locus together with either the Rab5 homolog mCherry-Rab22Ap (upper panel) or mCherry-Rab7p (lower panel), respectively. *Rab* transgene overexpression was induced with 1 μg/mL $CdCl_2$ for 2.5 hr in SPP. Scale bars, 10 μm. **C:** Vps8a-mNeon overlapped ∼50% with mCherry-Rab7 but only ∼12% with mCherry-Rab22A. The mean M1 values and SDs (error bars) were derived from 65 and 66 nonoverlapping images for Rab22Ap and Rab7p samples, respectively. **D and E:** Endosome distribution in wild type versus ***ΔVps8a*** cells. Shown are single frames from timelapse movies of wild type and ***ΔVps8a*** cells, overexpressing mCherry-Rab22Ap (D) and mCherry-Rab7p (E). Cells were transferred to S-medium and mCherry-Rab transgene overexpression was induced as in (B). Rab22Ap-positive endosomes have a similar distribution in wild type and ***ΔVps8a***, but Rab7p-positive endosomes form large clusters in ***ΔVps8a*** cells. Scale bars, 10 μm. **F:** Coimmunoisolation of Vps8ap and c-***myc***-Rabs. Cells were transformed to express Vps8ap-FF-ZZ (FLAG-ZZ domain) together with either c-***myc***-Rab22Ap or c-***myc***-Rab7p via gene replacement at the endogenous loci. Detergent lysates were incubated with bead-bound anti-c-***myc*** or anti-FLAG Abs, and bead eluates were analyzed by SDS-PAGE and western blotting using anti-myc or anti-FLAG antibodies. The top panel shows total Vps8a-FF-ZZ in each cell line, while the bottom panel shows total myc-Rab22a or myc-Rab7 in each cell line. The middle panel shows Vps8a-FF-ZZ that was immunoisolated via an interaction with Rab22A (left lane) or Rab7 (right lane). Vps8ap preferentially interacts with Rab7p. **G**: Phylogeny depicting the relationships between Paramecium and Tetrahymena *Vps8* homologs. Rooting is based on placement when Oxytricha trifallax sequences were included, resulting in the depicted monophyletic clade with a support value of 1 /100 (MrBayes support value/RAxML support value). Four well supported clades are observed, with support values bolded and in red. The four strongly supported clades are named based on the T. thermophila homologs present therein: Clade AC, Clade B, Clade D, and Clade F. T. thermophila gene *Vps8e* is not found within any well supported clade, while other genes not found within well supported clades are labelled Vps8x. Bolded T. thermophila homologs are essential for viability, and the T. thermophila homolog in blue is the gene responsible for the UC616 phenotype.

using *Oxytricha trifallax Vps8* sequences as an outgroup, and the monophyly of the Oligohymenophorea-containing clade to the exclusion of Oxytricha had complete support under both MrBayes and RAxML (1/100). This analysis indicates that the post-Spirotrichea expansions occurred independently of expansions in other lineages and led us to postulate distinct functional roles for individual Vps8 paralogs that arose in Oligomenophorea. Consistent with this idea, the six *T. thermophila* paralogs have nonidentical expression profiles (Figure 4.8B), and the *Vps8a* profile appears unique (the paralogues were named via their phylogenetic relationship to known *T. thermophila* proteins based on the nomenclature in ciliate.org). To confirm that the duplication was likely at the base of the Oligohymenophorean clade, we carried out an additional analysis of HOPS/CORVET components in two other Oligohymenophorean genomes, *P. tetraurelia* and *T. maccalensis*. Both also demonstrated a loss of HOPS and expansion of CORVET machinery (Appendix 4). *P. tetraurelia* had twice as many copies of *Vps8* as *P. caudatum*, likely due to species-specific whole-genome duplication.

To test whether functional diversification had occurred between paralogs, Dr. Daniela Sparvoli and Dr. Aaron Turkewitz, colleagues at the University of Chicago, targeted each of the additional five paralogs for knockout in the MAC. For two of the genes, they could not recover viable knockout cells, i.e., a drug resistance cassette integrated at the Macronuclear locus could not be driven to fixation (Figure 4.8C). The results indicate that two paralogs are essential under these culture conditions, unlike *Vps8A*. Targeted disruption of the remaining three paralogs revealed that complete loss of *Vps8B*, *E,* or *F* had no effect on mucocyst secretion (Figure 4.8D, E). Thus, *Vps8A* represents a CORVET subunit that appears specialized for a single pathway of endolysosomal trafficking and may have little functional overlap with other *Vps8* paralogs. Labeling of the Δ*Vps8b*, Δ*Vps8e* and Δ*Vps8f* cells with FM4-64 suggests that Vps8b and e may be involved in phagolysosome maturation (Figure 4.8F, G), a role which is usually associated with the HOPS complex.

4.5 Discussion

4.5.1 Genomes and transcriptomes across ciliate diversity

In this study, we used publicly available transcriptome and genome sequence data for the analysis. All of the transcriptome data has been gathered for the service of other projects; for example,

environmental surveys of marine microbes in the case of the MMETSP project[320], phylogenomic analysis of ciliate taxonomy in the case of Gentekaki et al. (2014)[321], or analysis of genes associated with ciliate sexuality in the case of Dunthorn et al. (2018)[322]. Reanalysis of publicly available datasets has been a fruitful line of research in many cases, and it is therefore of interest to determine how widely applicable this transcriptome dataset may be to answer other cell biological questions.

Overall, most of the transcriptomes had sufficient coverage of the genes of the membrane trafficking system to provide informative data in the comparative genomic analysis. It is not surprising, then, that we observe a relatively complete membrane trafficking system in most of the genes surveyed in the transcriptomes. Despite some unusual losses (discussed in more detail below), most of the components of each protein complex were detected in most of the genomes. The evidence of loss from transcriptomes is much less convincing than that of genomes. It is impossible to infer the loss of a protein from its absence in a transcriptome; the gene may not be expressed or only expressed under certain conditions. However, the transcriptomes provide useful background information on distribution of MTS components in less studied ciliate classes and generally show similar levels of completeness to the sampled genomes; in the case of *Blepharisma japonicum* and *Climacostomum virens*, the detection of MTS subunits appears equivalent to, if not superior to, that of *Stentor coeruleus*, the annotated Heterotrich genome. This transcriptomic dataset therefore offers an informative, complementary approach to comparative genomics in the ciliate phylum. The use of transcriptomic and metatranscriptomic techniques for studying genetic diversity and environmental assessments in protists has already been well established[51,296,333,334].

Only three transcriptomes *(Anophyroides haemophilia, Entodinium caudatum,* and *Polyplastron sp.)* were discarded due to limited gene coverage. All of these species are either parasitic or symbiotic: *A. haemophilia* causes bumper disease in Atlantic lobster[335], while *E. caudatum* and *Polyplaston sp.* are both rumen-dwelling ciliates in the stomachs of cows[282]. Assemblies of obligate endobionts is often difficult as DNA contamination from the host is inevitable, and this may be why transcriptomic information here is limited. The rumen-dwelling ciliate sequencing was also carried out through sequencing of expressed sequence tags rather than high-throughput sequencing, which may have also resulted in poor coverage[282].

Of the transcriptomes accepted for the analysis, the most populous ciliate classes are represented at least once. The only notable absence is the Litostomatea as all Litosomatean representatives (*Entodinium caudatum, Litonotus sp.* and *Polyplastron macrostoma*) were lost during the cleanup protocol. Due to their environmental importance, particularly in anoxic environments, obtaining genomic or transcriptomic data from Litostomatean ciliates is a priority for determination of the cell biological adaptations that may confer resistance in a reclamation context. However, some information can be gleaned from analysis of Litostomatean ciliates in other anoxic conditions: in the case of rumen-dwelling ciliates, there is evidence of extensive horizontal gene transfer of metabolic genes for cellulose degradation from bacteria, while analysis of ciliates associated with deep sea conditions shows increasing species richness correlates with available prey[282,336]. These dynamics provide important data on potential mechanisms of adaptation in hydrocarbon-contaminated environments and are likely relevant to Base Mine Lake; Litostomatean ciliates dominated the ciliate biodiversity in the summer of 2015 as described in Chapter 2.

4.5.2 The membrane trafficking system of the ciliate phylum

Ciliates are, generally, free-living heterotrophs and are noted for their expanded exocytosis and endocytosis machinery[265,301]. One would therefore expect to observe a relatively complete, if not expanded, membrane trafficking system, as all these traits are associated with a reliance on cell compartmentalisation and diversified membrane-bound compartments. The unique genetic traits of ciliates would also suggest an expanded genetic complement in ciliates as they have extensive whole-genome duplications in multiple lineages, and a two-nucleus system that has led to hundreds of copies of their genome being present in a single cell[269,296].

Most MTS expansions in ciliates appear to be lineage-specific; few components are reliably duplicated in a pattern that suggests ancestral duplications or expansions (with the notable exception of the CORVET subunits in Oligohymenophorea, discussed in more detail below). In general, the Heterotrichea demonstrate a more complete MTS complement than the other classes of ciliates for which genome representatives are available (Spirotrichea and Oligohymenophorea), while Oligohymenophorea appear to have the most reduced MTS complement. This is contrary to their phylogenetic position; while Heterotrichea is the most basal of these ciliate classes, Spirotrichea is generally considered the most derived of the ciliate classes both phylogenetically

and in its observed traits, such as extensive genome fragmentation[321]. Colpodea, the class most closely related to the Oligohymenphorea phylogenetically, also demonstrates a complete MTS in the sampled transcriptomes on some protein complexes that are not detected in Oligohymenophorea, like the HOPS complex. This suggests that the reduction of the MTS complement of Oligohymenophorea is not due to a gradual reduction of MTS components across the ciliate phylum but is specific to this ciliate class.

The lineage-specific diversification of ciliate MTS proteins could be influenced by multiple aspects of their biology. The effect of the unusual genome structure on ciliate genetics has been studied in detail in the model organisms *Paramecium urelia spp*[288]., *Tetrahymena thermophila*[289], as well as across ciliate diversity using a single-cell 'omics approach[296]. In the *P. aurelia* species complex, there are numerous whole-genome duplications, the origins and diversifications of which have been the object of extreme interest from researchers[288]. The potential for gene duplication in an organism which regularly duplicates its entire genetic complement is obvious and has resulted in a much-expanded MTS across ciliate diversity. Indeed, the reason the genome of *Paramecium caudatum* was selected as the *Paramecium* representative in this chapter was due to the difficulty of untangling whole-genome duplications from expansion of specific MTS machinery. Studies of mutation accumulation in *P. tetraurelia* and *T. thermophila* have also shown that, contrary to what may be expected from an organism that has to reassemble an expanded version of its nucleus after every cell division, accumulation of mutations in the germline micronuclear genome (MIC) is slow; this would be consistent with the conservation of most MTS complexes consistently across ciliate diversity[337,338]. Finally, the single cell 'omics studies across the diversity of eukaryotes show that some of the derived traits of ciliates that were thought to be restricted to specific lineages, such as gene scrambling in Spirotrichea, are actually found in multiple ciliate classes[296]. Studies also suggest a correlation between gene scrambling and genetic diversification, such as the use of alternate genetic codes[296]. While I did not observe any correlation between overall abundance of MTS components and genome scrambling in the overall comparative genomic dataset, this is a potential avenue for future research into environmental adaptation in the ciliate MTS.

4.5.3 Adaptins and altered selection on AP3

Adaptins are a homologous family of proteins which have diversified according to the organelle paralogy hypothesis to act at different regions of the cell[323]. Because these protein complexes are homologous, it means that phylogenetic analysis is an appropriate tool to determine the diversification of each subunit (mu, beta, sigma, and epsilon/gamma/alpha/delta/zeta) across ciliate diversity. As the divergence of each of the subunits of the HTACs into the adaptin complexes AP1, AP2, AP3, AP4, and AP5 occurred well before the diversification of the ciliates[324], we expect and observe separation of each of the subunit trees into their respective protein complexes with strong support (the exception to this is AP1/2B, which is an AP2 subunit that functions as the beta subunit in both AP1 and AP2 in many protist lineages including ciliates[324]).

Generally, proteins which have lost their function in an organism are subject to relaxed purifying selection as mutations in the sequences can accumulate without having a detrimental effect on the phenotype of the organism. The proxy for divergence in phylogenetic trees is branch length; the longer the branch, the less conserved the protein. The branch lengths we observe in the adaptin subunit trees for the ciliate genomes are consistent with less conservation of these subunits; in turn, this is consistent with loss or diversification of AP3 function across the ciliate phylum. AP3 is an adaptor protein involved in the recruitment of cargo for vesicle-mediated traffic between the Golgi body/tubular endosomes and late endosomes/lysosomes[339]. Tethering of AP3-coated vesicles is in most model systems mediated by the HOPS complex[339], which is notably lost in Oligohymenophorean ciliates as discussed in section 4.3.10. However, AP3 is still absent or divergent across all sampled ciliate classes, while the loss of HOPS and functional diversification of CORVET is only observed in Oligohymenophorea. It may be that mutation in AP3 was a driving force behind the functional diversification of CORVET in this lineage, though further study would be required to confirm this observation.

AP3 has also been lost in other organisms, including Haptophyta, an algal lineage basal to the SAR clade[340]. Multiple independent losses in adaptor complexes 2, 3, 4 and especially 5 is relatively common, so partial loss of the AP3 complex is not unusual from a eukaryotic diversity perspective. However, in *T. thermophila*, which retains a full AP3 complement, AP3 is associated with

biogenesis and maturation of their extrusomes, the mucocysts; these have also been shown to be derived from lysosome-related organelles and the related membrane-trafficking pathways[264]. Since extrusomes are found across ciliate diversity, it is surprising that the adaptor complex that is most closely associated with their function is also one of the less detected complexes. This may be due to functional diversification of the components causing them to be undetected by comparative genomic analysis; however, I used every possible method for detecting additional subunits, so if this is the case, the divergence from the canonical AP3 subunit gene sequences would have to be profound. It is also worth noting that each of the four components have long branches from the base of the AP3 grouping, which suggests that the subunits are not diversifying as a functional complex within ciliates as a whole; this would result in a long branch at the base of the AP3 clade and then shorter branches to each of the species (or a specific clade if the diversification was at the base of a class rather than at the phylum). The long branches for each subunit being individual to each species implies that either diversification is lineage-specific in every lineage, which would be statistically unlikely, or that AP3 is instead under relaxed selection, which implies loss of function. If the AP3 machinery or the function of AP3 is lost in ciliates, there must be another mechanism for membrane trafficking to extrusomes which is yet to be uncovered.

4.5.4 Multisubunit tethering complexes

The membrane trafficking complexes (MTCs) COG, exocyst, and TRAPPII were substantially reduced across the diversity of ciliates. COG and exocyst are two MTCs which were identified in the initial MTC survey as being present in most genome, but only specific subunits and, generally, not enough to form a functional complex. The other general observation of interest was that the complexes appeared to be more complete in the basal Heterotrichea than any of the other ciliate classes. In the initial analysis of genomes and transcriptomes described in 4.3.4, I used a multistep process to try to identify missing subunits from the ciliate genomes and transcriptomes; I used hidden Markov models for more sensitive detection of divergent components, and I searched into the scaffolds of the genomes to detect any subunits that may have gone unannotated. In the relatively complete COG and exocyst components of the Heterotrichea, many subunits were detected only by more sophisticated homology searching techniques instead of the initial BLAST searches. This suggests that many of the subunits are more divergent from the initial query dataset than the other complexes across the membrane trafficking system.

COG, or Conserved Oligomeric Golgi complex, is associated with the morphological phenotype of stacked versus unstacked Golgi[341]. Though in most model organisms such as the Metazoa, the Golgi body is stacked in a series of flattened membranous sacs; this varies widely across eukaryotic diversity and other organisms have Golgi bodies as a series of punctate vesicles surrounding the nucleus[341]. Early morphological studies have identified a noncanonical Golgi in many ciliates; in the model ciliates *T. thermophila* and *P. tetraurelia,* the Golgi is in the form of small, flattened discs, called dictysomes, with a structure described as "beads on a string"[342]. In the Spirotrich ciliate *Nyctotherus ovalis,* the Golgi takes the form of short, twisted tubules surrounding the nucleus[343]. There is clearly a diversity of forms to the ciliate Golgi body, and how this relates to the COG complex is unclear due to the lack of understanding both of Golgi trafficking in ciliates and how this complex relates to Golgi stacking in other model eukaryotes[344]. Due to the small number of identified components of COG in this dataset, and because the subunits are not homologous, there was not enough data to run any sort of selection analysis on the remaining components. However, it is notable that the COG complex was nearly complete in *Blepharisma japonicum* and *Climacostomum virens*, two Heterotrich ciliates. Unfortunately, there is currently no data on the Golgi structure in these organisms; this may be of interest to future researchers in determining the importance of COG to Golgi stacking.

The second MTC that appeared to be mostly missing in multiple ciliate lineages is exocyst. Exocyst is a complex that has various roles in membrane trafficking at the plasma membrane[345]. These roles include primary biogenesis of cilia and constitutive exocytosis, so it is surprising that a phylum named for its cilia and is known for its unique exocytosis organelles has a reduced exocyst complex[345]. The paucity of exocyst complex proteins in ciliates has been noted for at least a decade. This protein is implicated in the extensively studied process of constitutive exocytosis and also in the formation of the contractile vacuole in the amoebozoan *Dictystelium discoideum*—the contractile vacuole also essential for ciliates[328,346]. It was also noted in a recent study of the evolution of the exocyst complex across the diversity of eukaryotes that it is only partially complete in the model ciliates *P. tetraurelia* and *T. thermophila*[347]. This study suggests that, due to the general distribution of either total loss or total conservation across eukaryotic diversity, that all subunits are probably necessary for the protein complex to be functional[347]. However, it is worth noting that all previous work on exocyst uses only *T. thermophila* and *P. tetraurelia* as

sampling points in ciliates as, until relatively recently, these were the only genomes available. Both species are from the class Oligohymenophorea and by including genomes from additional classes of ciliates, this view of ciliate diversity is changed. In the Heterotrichea, more than half of the exocyst subunits are conserved and nearly all the subunits (missing only Exo84 and Sec3) are present in *Fabrea salina.* This result highlights the importance of incorporating more diverse ciliate genomes into assessments of ciliate gene complements; while the majority of trends are conserved across the ciliates, the reduction of MTS components in the Oligohymenophorea means that it may not be the most informative sampling point when compared to the Spirotrichea or the Heterotrichea, both of which now contain species with annotated genomes.

Dsl1 is an MTC complex that is involved in protein trafficking from the Golgi to the ER[348]. Dsl1 is also extremely important in peroxisomal biogenesis; peroxisomes bud from the ER in a mechanism mediated by Dsl1[329]. This complex is also the most visibly reduced across ciliate diversity; no species have more than two of the four subunits, and most have no detected Dsl1 components at all, including five of the nine genomes. Though, as previously noted, it is impossible to conclusively prove absence from a comparative genomic study, the evidence that a functional Dsl1 complex across ciliates has been lost is compelling.

Peroxisomes in ciliates, however, have been described in the model organism *T. thermophila* since the 1960s, and some of the earliest methods for staining peroxisomes were developed in this organism[349]; *T. thermophila* has also been included in an analysis of peroxisome proteins across the alveolate clade[350]. However, most research on the molecular mechanism peroxisomal biogenesis (the process in which Dsl1 is implicated) has been carried out in budding yeast[329]. This organism has an extremely reduced genome and has often been shown to be an extremely poor model of most cellular processes[351]. Since the morphology of other membrane-bound compartments, such as the Golgi, varies across the diversity of ciliates, the presence of peroxisomes in *T. thermophila* is not necessarily indicative of their presence in other species of ciliates. I therefore used the comparative genomic approach to survey the presence of peroxisomal genes in all the sampled ciliate genomes.

This experiment demonstrated the conserved presence of core peroxisomal proteins in all the sampled genomes. Additionally, I also surveyed the coding sequences of the genomes for the

presence of the first of the two peroxisomal targeting sequences, which consists of an SKL motif at the C terminus of the protein[352]. There was a substantial presence of peroxisomal targeting sequences on these genomes, many of which were determined to be peroxisomal-associated proteins based on a BLAST search, found across the three classes Heterotrichea, Spirotrichea and Oligohymenophorea. This, along with the characterised peroxisomes found in *T. thermophila*, strongly suggests that peroxisomes are present and functionally active in ciliates.

The total loss of detectable Dsl1 components does raise the question of how peroxisomal biogenesis occurs in ciliates. Characterising the peroxisomal biogenesis and trafficking mechanisms in ciliates may provide insight into how this process works outside of Opistokonta.

The final noted MTC absence in ciliates is the HOPS complex from Oligohymenophorea. These two complexes share a core of subunits (Vps11, 16, 1,8 and 33) while HOPS has the additional subunits Vps39 and 41, and CORVET has the additional subunits Vps3 and 8[332]. CORVET as well as HOPS genes are found throughout eukaryotes[328], implying that both complexes arose during pre-LECA evolution, with duplication and co-evolution of genes producing two complexes whose subtly altered specificities contributed to differentiation of endosomal trafficking pathways. The CORVET complex, which functions at the early endosomes, had only been functionally characterised in Opistokonts before its investigation in *T. thermophila*[264]. The HOPS complex, which functions at the late endosomes, has also been characterised in plants (lineage Archaeplastida), where it appears to function at late endosomes/vacuoles[353]. We previously noted the unusual absence of HOPS-specific subunits in genomes of two ciliates, *T. thermophila* and *Ichthyophthirius multifiliis*[354].

By analyzing additional ciliate genomes and transcriptomes, we can now infer that those losses occurred prior to the expansion of the Oligohymenophorea. Though loss of genes cannot be inferred from transcriptomic data, presence of HOPS components in both genomes and transcriptomes from outside the Oligohymenophorea implies that this is a group-specific adaption. These losses are unlikely to reflect an evolutionary reduction in the complexity of endosomal pathways: previous studies of *T. thermophila* expresses at least three Rab8 paralogs and six likely Rab11 paralogs, proteins which are essential for vesicle-mediated trafficking, and HOPS-specific losses are paralleled by proliferation of both core and CORVET-specific subunits[264]. Of the six

paralogues of the CORVET complex subunit Vps8, Dr. Sparvoli and Dr. Turkewitz's studies in the functional characterisation of HOPS and CORVET found that only the *Vps8A* paralog is essential for mucocyst formation, while no other nonessential paralogs showed mucocyst phenotypes[264]. Moreover, two of the *Vps8* paralogs are essential, while the others have no growth phenotype. This also suggests that the subunits have important but disparate functions, at least in *T. thermophila*[264]. These studies by Drs. Sparvoli and Turkewitz of the distribution of CORVET components in *T. thermophila* have shown that it forms at least six different complexes with distinct cell localisations, further supporting the evolutionary flexibility of membrane tethering complexes first noted in these results.

The expansion of CORVET is one of the only examples of the expansion of MTS proteins that does not appear to be species-specific based on the initial survey of the diversity of membrane trafficking. This underscores the importance of lineage-specific diversification in ciliates, and subsequently, the importance of sampling a diversity of species to gain a full understanding of the diversification of protein complexes across a phylum. This is also the first recorded example of this level of functional plasticity in membrane trafficking evolution. Though there are examples of sharing of subunits in multiple complexes, this expansion and functional compensation of one MTC for another has not previously been observed; whether this is a pan-eukaryotic trait or a specific innovation in ciliate membrane trafficking remains to be seen[264].

HOPS is the complex which is usually associated with trafficking to lysosome-related organelles (LROs). As discussed in Section 4.4.3, AP3, another protein complex of the MTS, is also associated with trafficking to LROs. These organelles, which in ciliates have been subject to specific innovations in membrane trafficking in both MTCs and adaptins, have also been implicated in susceptibility to polyaromatic hydrocarbons in the ciliate *Euplotes crassus*[311]. Euplotes, which is found within the Spirotrichs is a genus of ciliates which is well sampled, and has three representatives, including a genome, in this dataset. Like the other Spirotrichs, the *Euplotes octinarius* genome has both HOPS complex subunits (*Vps39* and *Vps41*), though the CORVET subunit Vps3 was not detected. These genetic diversifications across the ciliate phylum may explain why some ciliates appear to be extremely susceptible to hydrocarbon contamination while others are extremely resistant.

4.5.7 Conclusions

Ciliates are an extremely diverse phylum, morphologically, structurally, and genetically. In this chapter, I have used comparative genomics to survey the heterotetrameric adaptin complexes, endocytic machinery, and multisubunit tethering complexes across the diversity of ciliates, including nine annotated genomes from three different ciliate classes (Heterotrichea, Spirotrichea, and Oligohymenophorea) and a further twenty-one ciliate transcriptomes, also from publicly available datasets, spanning nine classes. These data demonstrate the utility of combining transcriptome and genome data to expand sampling points to obtain more information of the distribution of a trait across a clade, and, due to the explosion in transcriptomes and single celled 'omics research programmes currently underway, this methodology will only expand.

We also observed ciliate-specific adaptations to the membrane trafficking system, including a mechanism for trafficking plasticity that had not previously been observed and has not been noted outside of ciliates. These innovations, which include expansion and functional compensation of HOPS/CORVET, loss of Dsl1 but retention of peroxisomes, and patchy distribution of other MTS complexes including AP3, TRAPPII, exocyst and COG, are not distributed evenly across the ciliate classes. This underscores the importance of expanding beyond the two traditional ciliate models, *Tetrahymena thermophila* and *Paramecium tetraurelia*, and embracing the full potential of the ciliate phylum.

4.6 Afterword

The most notable developments in this field since the publication and presentation of the various components that make up this chapter are the continuing work on the HOPS/CORVET complex in the model ciliate *Tetrahymena thermophila* by Drs. Sparvoli and Turkewitz, the substantial interest in single-celled genomics of ciliates obtained from environmental samples, and the reclassification of *Mesodinium* as a within the Litostomatea.

In Dr. Sparvoli and Dr. Turkewitz's work, most recently published in Sparvoli et al. (2019), the authors characterised the various complexes associated with the expanded paralogs of *Vps8* in *T. thermophila.* They have identified six different CORVET-like complexes, which all have distinct subunits and subcellular localisations. This evidence further shows how flexible the ciliate

membrane trafficking system can be and provides another potential mechanism for formation of trafficking pathways in ciliates.

Single-celled genomics is a relatively new technology and allows the sequencing of entire genomes from single cells obtained from either culture or environmental samples. Single-celled genomics is particularly effective in ciliates as the cells are large and easily distinguishable, making obtaining a diverse sampling of ciliates from a single environmental sample much more feasible. As referenced in this chapter, Maurer-Alcala et al. (2018)[296] used single-cell sequencing to examine the diversity of ciliate genome architecture in a project that produced several new ciliate transcriptomes, including two Litostomatean ciliates. Similarly, Hu et al. (2019)[355] describe advances in ciliate systematics in the South China Sea based on eDNA studies, an area which has provided fruitful for culturing and sequencing of ciliate genomes. The number of available ciliate genomes and transcriptomes is going to increase substantially in the coming years, which would provide the possibility of expanding on this analysis both for understanding ciliate diversity, and for examining the biogeography of traits which we may have previously thought restricted to a single environment like Base Mine Lake.

Finally, a recent paper by Lasek-Nesselquist and Johnson (2019)[356] used phylogenomic analysis to place the previously incertae sedis lineage *Mesodiniae* within the Litostomatea. This ciliate genus (which, unlike most heterotrophic ciliates, has a kleptoplastic lifestyle) was included in this study and can therefore be considered a Litostomatean representative. There are some notable absences from the *Mesodinium pulex* membrane trafficking system, including COPI, most of the endocytic machinery, and most of the multisubunit tethering complexes. However, due to its unusual lifestyle, *Mesodinium pulex* may not accurately represent the likely anaerobic or low oxygen adapted ciliates found early in Base Mine Lake, and, since this the available genetic information on species is a transcriptome, little can be inferred from absences. There would still be enormous benefit in an assembled genome from an anaerobic ciliate, ideally obtained from an environmental sample, to shed light on the questions of ciliate adaptation to reclamation environments.

Chapter 5
GENERAL DISCUSSION

5.1: Exploratory research and technological advances as epistemological frameworks

It is impossible to write a general discussion summarising comparative genomics, phylogenetics, and eDNA analyses without a brief discussion of the effect these advances, both technological and theoretical, have had on the practice of scientific inquiry as a whole.

Discussions on how 'Big Data' has affected science have been ongoing since the advent of high-throughput sequencing technologies[357]. Early discussions in the field used the example of microRNAs[358], but literature now includes genomics and biodiversity assessments made using eDNA sequencing and analysis[359,360]. Many philosophers and sociologists of science have made the distinction between hypothetical-deductive methods of science (popularised by the falsifiability requirements of Karl Popper) and data-intensive or data-driven methods, which are often compared to the methods of 'natural science' and observation used by Newton, Darwin and other 'natural scientists' pre-1800s[361–363]. The hypothesis-driven model is approximately summed up as the 'scientific method', where a previously established theoretical framework is used to formulate a hypothesis which is then falsified or supported through empirical observation[364]. On the other hand, the data-driven model involves empirical observation without a specific hypothesis in mind. In this method, theoretical models are often established after experimentation and usually rely on statistical techniques for data analysis[365]. Whether this divide is a positive or negative thing for scientific progress, or indeed whether it is a purely artificial construction[358], is still a powerful point of discussion[366,367].

The importance of technology and its effects on science are also an important philosophical consideration relevant to the analyses presented in Chapters 2-4. The role of technology in science is controversial; some argue that an overreliance on technology is detrimental to scientific inquiry, though the prevailing view is that scientific progress is irreducibly enmeshed with technological progress[368,369]. This is explored in the context of Big Data questions as a proposed 'fourth paradigm' of enquiry, after experimentation, analysis, and computer modelling[370]. Like the overall importance of technology in science, this issue is still under discussion. There are multiple

proposed models that aim to explain the practice of science as it is done in modern research environments. These models, which include integration, iteration, and eliminative inference combine technological advances, acceleration of data acquisition, and exploratory research[358,362,371,372].

While a full discussion of this topic, which is an open and lively debate within the philosophical field, is beyond the scope of this thesis, it is generally accepted that exploratory or 'data-driven' research is its own method of epistemological inquiry that is complementary to hypothetical-deductive or 'hypothesis-driven' research. Accordingly, the overall conclusions I draw from the data analysed in this thesis benefit from being addressed in the context of both frameworks; while there are some explicit hypotheses that can be conclusively discussed here, many of the scientific questions, particularly in Chapters 2 and 3, are best evaluated in the context of their exploratory value as opposed to a particular supported or falsified hypothesis. I also address open questions and further research in the field via multiple epistemological frameworks: some are best addressed via potential technological advances, some can be framed as explicit hypotheses, and some are exploratory.

5.2: Oil sands reclamation and protistological research 2014-2019

The first analyses which specifically assessed protist diversity in bitumen extraction sites in the Athabasca Oil Sands began in 2013 and were mostly summarised in Aguilar et al. (2016)[233]. Though the discovery of eukaryotic microbial diversity in the tailings ponds of Northern Alberta was not particularly surprising—identifying novel eukaryotes in extreme environments is a common method of prospecting for novel taxonomic diversity or cellular machinery—this was the first study to firmly support its diversity. There has been no further research published on protist diversity in tailings ponds specifically. Instead, the research focus has shifted to the reclamation process of tailings ponds and the development of end-pit lakes, the proposed endpoint for reclamation of FFT and tailings ponds in general[93].

Much of the overall research programme in oil sands reclamation still focuses on biogeochemical processes and the geophysical aspects of FFT settling such as mass balance and separation of the tailings and water phases[117,204,229]. There are extremely sophisticated models for how these will

affect the reclamation potential of tailings ponds going forward, and the biogeochemical research in particular has resulted in some fascinating discoveries in the bacterial microbiome of BML such as novel enzyme types and even novel clades[120,230]. Some of the geophysical and biogeochemical discoveries and models can be trivially incorporated into our understanding of the protistology of BML: for example, the reduction in turbidity extensively detailed in Tedford et al. (2018)[204] directly results in the resurgence of photosynthetic organisms noted in Chapter 3. Other interactions are more complex. The biogeochemical cycling and transient anoxia at the tailings/water interface detailed in Risacher et al. (2019) is likely to have profound effects on the heterotrophic eukaryote communities in BML, but the relationships are less defined and will likely only be elucidated through careful experimentation. Studies of oil sands reclamation sites have also expanded during the period of the analyses described in this thesis to include new end-pit lakes. In particular, Demonstration Pit Lake (now called Lake Miwasin) has been established on the Suncor lake site using a tailings consolidation technology to accelerate the settling process[373]. This will be the second full-scale end-pit lake in the Athabasca Oil Sands mining area and will provide a valuable second data point in studies of the microbiology of these environments.

The research that comprises Chapters 2 and 3 remains, to date, the first and only analysis of protist diversity in an oil sands end-pit lake. Chapter 2 assesses the protist diversity of BML in the summer of 2015 while the lake is still in the process of construction. As an exploratory study, Chapter 2 has undisputed value: this is the first published assessment of protist diversity in an end-pit lake at an early stage of reclamation. Chapter 3 extends this exploratory analysis across four summers. Both studies generated a large amount of data; I identified around 700 OTUs at 97% diversity in the summer of 2015 and around 7000 in the dataset that contains all four summers. Some of these data have already been distributed within the wider research community. OTUs from the summer of 2015 have been included in the EukRef environmental database[21] and are available for other researchers looking for OTUs similar to their own choice of study environments.

As well as oil sands research, eukaryotic microbial analysis via eDNA study has also expanded rapidly in the period of these analyses. From 2013-2014, when examination of the eukaryotic complement of tailings ponds and their associated environments began, the widespread adoption of these techniques was still relatively new. The theoretical implications of the development of

these techniques in the last decade is discussed in Chapter 1, and the practical implications on the data presented here is assessed at length in Appendix 1. Briefly, the most relevant ongoing debates in the field are the increasing standardisation of methodology[35], the ongoing discussion on what exactly constitutes species-level diversity in DNA barcoding analyses[35], and whether DNA-only analyses can be used to identify new organisms without additional cell biological data[31].

All of the eDNA analyses in this thesis, and so far in the literature around protist diversity in oil sands, were carried out using standard biodiversity assessment protocols at the time. The analyses in Aguilar et al. (2016)[233] were carried out using the usearch algorithm implemented in mothur[55], Chapter 2 was carried out using the usearch algorithm implemented via a command-line script, and Chapter 3 was carried out using vsearch implemented in QIIME2[374]. The changes in methodology here represent changing standards in the field during the six years in which the analyses were completed: eukaryotic diversity assessments in mothur are computationally intensive and had to be abandoned for the larger-scale analyses in Chapter 2. QIIME2, released in beta in 2017 and officially released in 2018, quickly became a tool of choice for many microbial ecologists due to its simple interface and language and its modular structure. The other major technological change in this thesis (which I did not adopt in the course of this project) was the dada2 pipeline for producing ASVs[375]. Introduced in 2016, ASVs have been described as technologically and theoretically superior to OTUs since they are independently comparable between samples and more representative of species-level diversity[72]. Despite these potential advantages of ASVs, my reasons for not using them more extensively in the research in this thesis are discussed in detail in Appendix 1. It is also worth noting that ASVs have not yet completely replaced OTUs, and the debate over whether OTUs, ASVs (another potential standard), or no DNA sequence at all can accurately represent microbial diversity is not settled[35].

The final data chapter of this thesis, Chapter 4, discusses comparative genomics of membrane trafficking across ciliate diversity. The ‘canonical’ (i.e. metazoan and fungal) membrane trafficking system is well understood, but the details of how this cellular process has been adapted across the diversity of eukaryotes is still very much an open question[376]. There are two potential data inputs for a comparative genomic analysis: more sampling points from different organisms within the clade of study or a more complete or accurate genome assembly of known data points.

The comparative genomic analysis of the membrane trafficking system of ciliates has benefitted from both of these inputs during its development; continued interest in ciliate genomes has resulted in the publication of the *Stentor coeruleus*[290], *Uroleptopsis citrina*[266], and *Euplotes octocarinus*[267] genomes in 2017, 2018, and 2018, respectively. *S. coeruleus* is an important Heterotrichean representative in a class which did not previously have any sequenced genomes, while *E. octinarius* was a complete and annotated *Euplotes* genome, a genus which was previously only represented by two environmentally derived transcriptomes.

A notable trend which has emerged during the period over which this research occurred is an explosion in single-cell derived genomes, a technology which has also been applied to ciliates. This advance has resulted in additional genomic data that can potentially be added to these analyses[296]. Chapter 4 represents an initial, and hardly exhaustive, assessment of the MTS system across the diversity of eukaryotes. This analysis is also exploratory: while I identified clear trends in the distribution of membrane trafficking components, the results of this chapter provide a baseline of MTS diversity across the ciliates which can potentially inform future studies as well as an interesting insight into potential cellular idiosyncrasies. For example, the replacement of HOPS components with CORVET components in the membrane tethering system is a mechanism of cellular plasticity that has not previously been observed in model organisms, while the confirmed presence of peroxisomes in the absence of Dsl1 implies a method of peroxisomal trafficking different from any known eukaryotic microbes. The research in Chapter 4 provides support for the interpretations of the ciliate diversity from BML, which contributed considerably toward our understanding of heterotrophic diversity and its role in nutrient cycling in this environment.

5.3: The ecology of Base Mine Lake - Baselines, Mines and Lakes

The eDNA microbiological studies of the eukaryotic communities in BML began in the summer of 2015. Though the reclamation site was well established at this point, there were still some geophysical features—most notably the high turbidity—that were subject to industrial intervention through alum addition[204,224,229]. BML in 2015 had extremely low light penetration, an easily disturbed fluid/tailings interface and elevated concentrations of known freshwater contaminants such as heavy metals and naphthenic acids[118,204,230]. The major finding of Chapter 2 (that there were almost no photosynthetic species detected by eDNA sequencing) is explicable from the low-

light conditions. 2016 marked the return of photosynthetic taxa in BML. The photosynthetic complement of BML is currently the focus of further study, and though the results have not yet been published, they appear to have flourished since turbidity was substantially reduced[118].

Due to the abundance of understudied clades of heterotrophs identified in Chapter 2, I focused my approach in Chapter 3 on heterotrophic diversity and found that, overall, biodiversity increased after the reduction in turbidity. Three clades were identified as particularly abundant in 2015: the Litostomatea from the ciliate phylum, the Glissomonadida from the cercozoan phylum, and the Microsporidia from the fungal phylum. Across the time series between 2016 and 2018, other heterotroph taxa appeared to recover from being minimally present in 2015; most notable were the cercozoan Vampyrellida and ciliate Oligohymenophorea, with overall OTU numbers approximately equal to the previously noted Glissomonadida and Litostomatea, respectively. As an exploratory analysis, Chapter 3 was hindered by substantial data loss in the first half of 2016 due to the Horse River fire that destroyed the majority of Fort McMurray and effectively shut down oil sands production and monitoring for several months[146]. Exploratory science requires a comprehensive dataset for extrapolation of theoretical models from the data which meant that it was impossible to make strong inferences from the dataset we obtained. However, I did identify statistical support for a community relationship with time, especially when compared to the tailings pond Mildred Lake Settling Basin and the nearby reservoir Beaver Creek Reservoir. Reid (2019)[377] suggested in their analysis of the biogeochemistry of BML that, at least from a bacterial standpoint, local reservoirs on oil sands sites may be good comparison points for successful tailings pond reclamation. Based on the data from Chapter 2 and Chapter 3, the heterotrophic protist community appears to be less similar to the local reservoir over time. This does not necessarily conflict with the results of Reid (2019)[377] as the two studies take into account very different trophic groups, but it does suggest that a direct comparison to local reservoirs may be too simplistic.

Identifying potential future trends or hypotheses for BML's protist community is difficult because this is the first time an environment such as this has been studied. Potential comparisons to other anthropogenically influenced environments are difficult. End-pit lakes from other mines such as copper or diamonds have very different geophysics and much quicker tailings dewatering timelines[117]. Heavy hydrocarbon contamination from oil spills tends to be acute rather than

chronic[80]. Toxicology studies of the individual components of oil sands tailings often focus on single study species rather than communities[378,379]. The natural history of the end-pit lake must also be considered; Chapter 3 illustrates the effect of natural disturbances such as wildfires on a region.

Ultimately, the future of BML can only be determined through further monitoring. Studies are currently underway evaluating the algal communities between 2016 and 2019 to minimise the effects of the Horse River Fire on the time series. For future eDNA studies, there are also techniques which can limit the effect of relic DNA in time series, which would make any potential trends in time clearer[380]. As the current research suggests different trends in the ecology of organisms at different trophic levels, to understand this environment it will be necessary to ensure that these monitoring programmes are combined. It is also worth considering what the ideal endpoint of reclamation will look like. The conceptual ideal endpoints for reclamation in the context of oil sands mining have been discussed at length in the ecological literature, most notably in Johnson and Miyashi (2008)[381]. This paper introduced the concept of 'equivalent land capacity' in the Athabasca Oil Sands: as defined here, an equivalent land capacity does not necessarily resemble the environment that was present before the disturbance, but provides equal value, whether from industrial uses such as logging, from recreation, or as a conservation area. Recent toxicity studies of BML water cap on invertebrates have noted the elevated salinity of BML, suggesting it may end up a brackish rather than freshwater ecosystem[379]. Whether an equivalent land capacity is an acceptable endpoint for reclamation depends on the local communities as well as industrial considerations and is ultimately a political decision.

5.4: The heterotrophic flagellates of Base Mine Lake and the paradox of novelty

BML has been noted in multiple studies, and within this thesis, as a useful source for 'bioprospecting'. This phrase covers several activities broadly connected by their exploratory capacity and can refer to the search for novel organisms, gene families, or ecological processes. Over the four years of study, BML has proven its capacity in this regard as discussed in more detail in Chapter 2. From a bioprospecting standpoint, the main discovery detailed in this thesis is the two OTUs whose abundance appears to be closely associated with the wildfire in Chapter 3. Though the research here only details the 18S rRNA subunit of these organisms, this information

provides a starting point for detection and culturing of these organisms within a sample. Fluorescence in situ hybridisation (FISH) is a relatively straightforward way of detecting culturable cells from within a sample[382] and using this technique to identify these organisms within the BML water cap is certainly a possibility.

In the context of eDNA studies and exploratory data-driven research in general, the concept of novelty is potentially problematic. Some researchers describe data-driven research as a form of eliminative inference, famously described by Sir Arthur Conan Doyle's Sherlock Holmes as "once we have eliminated the impossible, whatever remains, however improbable, must be the truth"[362,372]. Rather than testing individual hypotheses, we can simply test every possible hypothesis and discard any that fail. There are two obvious issues with this approach: the statistical significance of one's results decrease the more hypotheses one tests, and that in order to successfully answer a question via this route of investigation one must be certain that every possible hypothesis has been tested. The first can already be accounted for in multivariate statistics; for example, the Maximal Information Coefficient tests used in Chapter 3 to detect significant relationships with time includes a step to correct for multiple testing[249]. The second problem is more insidious. If a novel organism is detected in BML, can we presuppose that it is specifically adapted to this environment without knowing what other environments in which it can possibly be found? There is an example of this problem in Chapter 2: when the analyses were initially completed in June 2017, one OTU consistently grouped with the Amoebozoan outgroup despite being classified as an Amoebozoa by BLAST. This remained a puzzle until a secondary search identified it as *Syssomonas multiformis*, an organism within a novel group of holozoans that were published and had their descriptions added to GenBank in July 2017[217]. Without this additional data, I may have concluded that this OTU represented novel diversity and therefore the organism was specifically adapted to end-pit lakes. In fact, *Syssomonas multiformis* has also been identified in watery buffalo excrement in Vietnam[217] and freshwater streams in Argentina[383], a fascinatingly diverse series of environments.

There are multiple approaches to solving this issue. The first is, simply, more data: there are many projects that aim to do exactly this, from GenBank itself[18] to the Earth Microbiome Project[384], EukRef[21], and others too numerous to name. Curating, collating, comparing and reanalysing data

from DNA databases has been a diverse and thriving field of its own for many years. However, it is impossible, at least with current technology, to sequence all diversity. This may change in the future, and a notable advance in this respect is the availability of affordable handheld DNA sequencers[65]. For now, any claims of novelty that are interpreted in respect to the environment in which they are detected must be cautious.

Another potential mechanism of detecting similarities in ecosystems when describing a new study site is focusing on the functionality, rather than the taxonomy, of the organisms under study. This is extremely effective in bacterial microbiome studies since the function and taxonomy of bacteria is often very closely linked and there is a much more comprehensive understanding of the metabolic pathways found within each genome[260]. These techniques already been successfully carried out in Base Mine Lake as discussed in Chapter 1[120,230]. The linkage of function and taxonomy in eukaryotes is much less apparent. Functional studies of eukaryotic eDNA are still in their infancy, though some extremely interesting results have already been obtained. The latest Adl et al. (2019)[28] classification of eukaryotes also includes functional classifications of the described taxa, though the categories used are necessarily broad. Functional traits of eukaryotes have also been described in greater detail in specific eukaryotic taxa, including Cercozoa and Microsporidia[385]. By focusing on commonalities in function rather than the individual taxonomy of each organism, it is possible to compare the ecosystems in BML to natural lakes without necessarily needing to know the exact species responsible for ecosystem functionality. These techniques could then be used to determine if the protist community in BML is, over time, becoming more functionally similar to the local lakes and reservoirs as described in the biogeochemical cycling. The taxonomic results in Chapter 3 when comparing the community before and after the Horse River Fire suggest that a comparison of functionality may indicate some level of resiliency in the Base Mine Lake protist community. Though the abundant species were almost entirely replaced in this period, the overall broad taxonomy of the heterotrophic community remained extremely similar; this may also be the case for ecosystem processes.

Chapter 4 of this thesis is also aimed at addressing potentially spurious claims of novelty. In extreme environments, it is possible that organisms have cellular adaptations that are associated with that environment in particular: a well-studied example is the prevalence of anerobic

respiration mechanisms obtained by horizontal gene transfer in protists that live in anoxic environments[386]. Ciliates are a group of organisms which have been previously noted for their resistance to anthropogenic contamination, as discussed in Chapter 4. It would be tempting to assume that any ciliate isolated from BML with novel cell biology may specifically be adapted toward hydrocarbon resistance. However, ciliates are also extremely diverse morphologically and genetically[269], and it would be important to determine whether any potential resistance is only novel due to a lack of baseline reference data. In this context, and as an exploratory analysis, Chapter 4 is a substantial advance on the previously known genetic diversity of ciliate membrane trafficking which was mostly limited to model organisms within the single class Oligohymenophorea[265]. Oligohymenophorea showed some cell biological adaptations which were unique to this class: most notably, the loss of HOPS and diversification of the CORVET components in regulated exocytosis. Based on the results in Chapter 3, ciliate biodiversity may be an interesting reclamation indicator in end-pit lakes in their early stages of reclamation. Any future research in this sphere would need to ensure that comparisons between ciliates isolated from BML and model organisms (or organisms isolated from undisturbed environments) takes the natural diversity of ciliates into account.

5.5: Further heterotrophic eukaryote research in reclamation environments - do we need more data or more hypotheses?

Base Mine Lake is a complex and emerging environment, and yet it is essential for the environmental health of the Athabasca Oil Sands region that research into end-pit lakes produces some real possibilities for reclamation of tailings ponds in the coming decades[93]. The future research which I believe can be built upon the work in this thesis includes both exploratory studies and some explicit hypotheses about what will be observed in Base Mine Lake in 2019 and beyond.

The first exploratory study that would complement this work is an eDNA assessment of the FFT layer in Base Mine Lake. This has been completed for years 2013 and 2014 for bacteria, as described in Chapter 1 and in Rudderham (2019)[121], and has shown stratification in biogeochemical processes in the layers beneath the tailings/water interface. Since Aguilar et al. (2016)[233] detected eukaryotes in active tailings ponds, it is reasonable to assume that there may be a eukaryotic component to this environment. Bacterial studies suggest that nutrient cycling and

particularly oxygen-consuming processes at the tailings/water interface may impact the success of reclaiming BML to a state similar to a natural lake[120,230], and understanding the heterotrophic protist contribution to these processes may assist in understanding and potentially controlling them.

An additional exploratory study that directly builds on Chapter 3 would be isolation and genome sequencing of the two organisms which responded particularly strongly to the Horse River Fire. Despite the caveats in assuming an organism found in an environment is specifically responding to that environment or that it cannot be found in other environments, the extreme abundance of these OTUs only following the wildfire makes them an interesting candidate for further research. This is listed as an exploratory study as there is no specific hypothesis associated with this genome sequencing. It would make phylogenomic taxonomy of these organisms possible, likely resulting in a more definitive classification than the preliminary pplacer classification obtained in Chapter 3. As some of the changes associated with lake exposure to wildfire (such as phosphorus and nitrogen deposition) are known[228], it may be possible to test the effects of nutrient addition to cultures in vitro.

It is also possible, due to technological advances in the last decade, to expand the eDNA monitoring programme of microbial eukaryotes in BML past taxonomic assessment of the 18S subunit. There are two particularly relevant recent innovations: long-read sequencing and single-cell genomics. Long-read sequencing, where reads of up to thousands of base pairs can be generated from a single strand, are now commonly available,[387] as are handheld versions of long-read sequencers[65]. A combination of long-read sequencing and improvement of assembly algorithms means that metagenomics, previously only the purview of prokaryotic research, is now possible in eukaryotes[388]. Indeed, metagenome-assembled genomes of eukaryotes have been successfully produced from environmental data, and tools are now available to assess their quality[389]. This would provide an avenue for the kinds of analyses of biogeochemical analyses that have resulted in such exciting discoveries from BML's bacterial microbiome. Eukaryotic metagenomes would also get around the issue of poorly understood or catalogued taxonomy in heterotrophic lineages outlined in Chapter 2 and 3; because the taxonomy of eukaryotes is not necessarily correlated with function, it is essential to have a very accurate classification of any

given eukaryotic OTU before one can make assumptions about its role in the ecosystem. A shift to analysing functional pathways rather than taxonomy would solve this issue, and, if taxonomy was required, many researchers are now showing that whole ribosomal subunits, or phylogenomic analyses, are more accurate methods of obtaining classifications than the V4 region of the 18S rRNA gene[321,390]. Single-cell genomics, on the other hand, also removes the requirements for culturing organisms before sequencing their genome, accelerating the timeline for obtaining whole genomes from environmental isolates and allowing researchers to sequence the genomes of organisms that may not be amenable to culture[391]. A single organism extracted from an extremophile environment can be studied in depth, or multiple organisms from a single sample can be analysed for common cell processes or biology.

Monitoring of the entire protist community of BML, as well as geophysical processes, biogeochemical processes, and toxicity assessments, is ongoing. In the heterotrophic community in particular, the results from this thesis suggest that overall biodiversity will continue to increase, including both species richness and diversity measures which take into account abundance. I also predict that, due to the elevated salinity, protists more commonly associated with marine environments will be found within BML. Though Alberta is landlocked for at least a thousand miles in every direction, the cosmopolitan distribution of microorganisms described in the Baas-Becking hypothesis and detected in the case *Syssomonas multiformis* will likely prevail in this case. Ultimately, future protist research in BML and end-pit lakes in general will be exploratory for at least a decade to come as the second end-pit lake in the Athabasca Oil Sands region has only recently been established.

Conceptually integrating data-driven, technological and hypothesis-driven research will be essential for the future of scientific enquiry, in the context of molecular ecology and in science as a whole and requires a cultural shift on the part of many academic scientists. Scientists often describe the "ultimate truth" that scientific frameworks aim to uncover as a Holy Grail, and this imagery is often repeated in analogies and dedications throughout the scientific literature[392,393]. The notion of "seeking the truth" in this manner implies that the scientist is a sort of Arthurian figure, nobly trying to discover that has been hidden from him by Nature for the greater good. Francis Bacon saw Nature as a female figure, an illogical, confusing being who would ultimately,

through the correct applications of the scientific method (which survives today in hypothesis-driven research) would be converted to the masculine ideals of logic, reason and rationality[394]. Leaving aside the extreme misogyny of this viewpoint, it once again places the scientist as the actor, who in the Baconian/Popperian scientific method possesses an ultimate tool (or Excalibur) to become the possessor of truth, which he can then disseminate to the public as an objective reality.

Though it is usually expressed in less literary terms, this viewpoint is still widespread in modern science. In his book "Theory and Reality" Peter Godfrey-Smith, in his discussion of the still-venerated scientific philosopher Karl Popper, says: "For Popper, a good scientist or a great scientist is someone who combines two features, one corresponding to each stage of [scientific inquiry]. The first feature is an ability to come up with imaginative, creative, and risky ideas. The second is a hard-headed willingness to subject those radical ideas to rigorous critical testing. A good scientist has a creative, almost artistic, streak and a tough-minded, no-nonsense streak. Imagine a hard-headed cowboy out on the range, with a Stradivarius violin in his saddlebags. (Perhaps at this point you can see some of the reasons for Popper's popularity among scientists)"[392]. It is easy to see why scientists might want to see themselves as a cowboy King Arthur. However, there is another archetype found in the Grail canon who I believe has striking relevance to scientific inquiry: The Fisher King.

Though the Grail legends are disparate and often contradictory, The Fisher King appears in many as the guardian of the Holy Grail. Once a mighty ruler who is still thought of fondly by his subjects, he is crippled by a wound and can no longer walk, fight, or effectively rule his kingdom. In some versions of the story, the wound was inflicted by God for the sin of pride[395]. When encountered by various Knights of the Grail, the Fisher King's lands are a barren wasteland, and the Fisher King himself ignores the suffering of his subjects and chooses to fish in the moat of his castle, paralyzed by his misfortune and sorrow. In the poem from which the epigraph of this thesis is taken, Parzival (anglicised as Percival) encounters the Fisher King, who shows him the wonders of the Grail at a great feast. It occurs to him to enquire about the Fisher King's injuries and the source of his power, but he chooses to keep to himself. In the morning, Parzival wakes and finds the castle and

surrounding lands empty. He learns that had he asked the Fisher King these questions, he could have healed the King, taken possession of the Grail, and ended the suffering of the people[396].

The Popperian view of the scientist places him as the protagonist of an Arthurian quest for hidden truth and emphasises the importance of the scientist's own personal convictions and motivations[392]. The Fisher King, on the other hand, is already in possession of the Grail, but is prevented by his own personal limitations from acting as an effective steward for it. I believe that this is a more accurate metaphor for the scientific endeavour, and one that does not privilege the hypotheses of the scientist as the most important aspect of scientific enquiry. In a world dominated by data, a view of science limited to hypotheses is paralyzing to innovative research and limits the collective scientific imagination. Our ability to understand scientific truths is limited by our methods of perception; increases in data and technology improve our collective perception of reality, an endeavour just as important as the hypothesis-tested theoretical frameworks that we use to explain it. As scientists, we are not seeking a truth that is hidden from us, we are trying to make sense of the truth of the physical world which surrounds us and in which we (and the scientific method) are embedded. We are all in the Fisher King's castle, trying to articulate how to ask the right questions.

5.6: General conclusions

Base Mine Lake is the first oil sands tailings pond in the Athabasca region and represents a new technique for environmental reclamation of fluid fine tailings. The preliminary and four-year studies of the eukaryote community in BML presented in this thesis reveal an environment initially dominated by heterotrophs which increases in heterotrophic biodiversity as reclamation continues. Though the confounding variables associated with the severe natural disturbance of the Horse River Fire prevent us from establishing any other significant trends, Base Mine Lake has proven to be a fascinating environment for bioprospecting as well as general environmental study over time. To ensure that any novel diversity or cellular adaptations ultimately detected in Base Mine Lake are actually a result of adaptation to this tailings environment rather than simply due to lack of sequencing diversity, this thesis also includes an analysis of the membrane trafficking system across the diversity of ciliates. This concludes that the standard model organisms for ciliates, *Tetrahymena thermophila* and *Paramecium tetraurelia*, are likely insufficient for identifying

environmental adaptations in this phylum and considering the natural diversity of ciliates is essential for understanding their biological interactions with anthropogenically influenced environments.

REFERENCES

1. Aristotle. *The History of Animals*. (Project Gutenberg).

2. Mayr, E. The Role of Systematics in Biology: The study of all aspects of the diversity of life is one of the most important concerns in biology. *Science* **159**, 595–599 (1968).

3. Walker, G., Dorrell, R. G., Schlacht, A. & Dacks, J. B. Eukaryotic systematics: a user's guide for cell biologists and parasitologists. *Parasitology* **138**, 1638–63 (2011).

4. Keeling, P. J. Combining morphology, behaviour and genomics to understand the evolution and ecology of microbial eukaryotes. *Philosophical Transactions of the Royal Society B: Biological Sciences* **374**, 20190085 (2019).

5. Heather, J. M. & Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* **107**, 1–8 (2016).

6. Zuckerkandl, E. & Pauling, L. Molecules as documents of evolutionary history. *Journal of Theoretical Biology* **8**, 357–366 (1965).

7. Brown, W. M., George, M. & Wilson, A. C. Rapid evolution of animal mitochondrial DNA. *Proc. Natl. Acad. Sci. U.S.A.* **76**, 1967–1971 (1979).

8. Schwartz, R. M. & Dayhoff, M. O. Origins of prokaryotes, eukaryotes, mitochondria, and chloroplasts. *Science* **199**, 395–403 (1978).

9. Bernhardt, H. S. The RNA world hypothesis: the worst theory of the early evolution of life (except for all the others)a. *Biology Direct* **7**, 23 (2012).

10. Ben-Shem, A., Jenner, L., Yusupova, G. & Yusupov, M. Crystal Structure of the Eukaryotic Ribosome. *Science* **330**, 1203–1209 (2010).

11. Noller, H. F. Structure of Ribosomal Rna. *Annual Review of Biochemistry* **53**, 119–162 (1984).

12. Olsen, G. J., Lane, D. J., Giovannoni, S. J., Pace, N. R. & Stahl, D. A. Microbial Ecology and Evolution: A Ribosomal RNA Approach. *Annual Review of Microbiology* **40**, 337–365 (1986).

13. Bik, H. M. *et al.* Sequencing our way towards understanding global eukaryotic biodiversity. *Trends in ecology & evolution* **27**, 233–43 (2012).

14. Demoulin, V. Protein and nucleic acid sequence data and phylogeny. *Science* **205**, 1036–1039 (1979).

15. Field, K. G. *et al.* Molecular Phylogeny of the Animal Kingdom. *Science* **239**, 748–753 (1988).

16. Tautz, D., Arctander, P., Minelli, A., Thomas, R. H. & Vogler, A. P. DNA points the way ahead in taxonomy. *Nature* **418**, 479–479 (2002).

17. Blaxter, M. L. The promise of a DNA taxonomy. *Philos Trans R Soc Lond B Biol Sci* **359**, 669–679 (2004).

18. Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. *Nucleic Acids Res* **44**, D67–D72 (2016).

19. Guillou, L. *et al.* The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic acids research* **41**, D597-604 (2013).

20. Pruesse, E. *et al.* SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research* **35**, 7188–7196 (2007).

21. Campo, J. del *et al.* EukRef: Phylogenetic curation of ribosomal RNA to enhance understanding of eukaryotic diversity and distribution. *PLOS Biology* **16**, e2005849 (2018).

22. Potvin, M. & Lovejoy, C. PCR-Based Diversity Estimates of Artificial and Environmental 18S rRNA Gene Libraries. *Journal of Eukaryotic Microbiology* **56**, 174–181 (2009).

23. Hadziavdic, K. *et al.* Characterization of the 18S rRNA Gene for Designing Universal Eukaryote Specific Primers. *PLoS One* **9**, (2014).

24. Decelle, J., Romac, S., Sasaki, E., Not, F. & Mahé, F. Intracellular Diversity of the V4 and V9 Regions of the 18S rRNA in Marine Protists (Radiolarians) Assessed by High-Throughput Sequencing. *PLOS ONE* **9**, e104297 (2014).

25. Luddington, I. A., Kaczmarska, I. & Lovejoy, C. Distance and Character-Based Evaluation of the V4 Region of the 18S rRNA Gene for the Identification of Diatoms (Bacillariophyceae). *PLOS ONE* **7**, e45664 (2012).

26. Fazekas, A. J. *et al.* Are plant species inherently harder to discriminate than animal species using DNA barcoding markers? *Molecular Ecology Resources* **9**, 130–139 (2009).

27. Suárez-Díaz, E. & Anaya-Muñoz, V. H. History, objectivity, and the construction of molecular phylogenies. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* **39**, 451–468 (2008).

28. Adl, S. M. *et al.* Revisions to the Classification, Nomenclature, and Diversity of Eukaryotes. *Journal of Eukaryotic Microbiology* **66**, jeu.12691-jeu.12691 (2018).

29. Will, K. W., Mishler, B. D. & Wheeler, Q. D. The Perils of DNA Barcoding and the Need for Integrative Taxonomy. *Syst Biol* **54**, 844–851 (2005).

30. Pennisi, E. DNA barcodes jump-start search for new species. *Science* **364**, 920–921 (2019).

31. Pinheiro, H. T., Moreau, C. S., Daly, M. & Rocha, L. A. Will DNA barcoding meet taxonomic needs? *Science* **365**, 873–874 (2019).

32. Howe, A. T., Bass, D., Chao, E. E. & Cavalier-Smith, T. New Genera, Species, and Improved Phylogeny of Glissomonadida (Cercozoa). *Protist* **162**, 710–722 (2011).

33. Bass, D. *et al.* Clarifying the Relationships between Microsporidia and Cryptomycota. *Journal of Eukaryotic Microbiology* (2018) doi:10.1111/jeu.12519.

34. Creer, S. *et al.* The ecologist's field guide to sequence-based identification of biodiversity. *Methods in Ecology and Evolution* (2016) doi:10.1111/2041-210X.12574.

35. Geisen, S. *et al.* A user guide to environmental protistology: primers, metabarcoding, sequencing, and analyses. *bioRxiv* 850610 (2019) doi:10.1101/850610.

36. Leray, M. & Knowlton, N. Censusing marine eukaryotic diversity in the twenty-first century. *Philosophical Transactions of the Royal Society B: Biological Sciences* **371**, 20150331 (2016).

37. Teletchea, F. After 7 years and 1000 citations: Comparative assessment of the DNA barcoding and the DNA taxonomy proposals for taxonomists and non-taxonomists. *Mitochondrial DNA* **21**, 206–226 (2010).

38. Berche, P. Louis Pasteur, from crystals of life to vaccination. *Clinical Microbiology and Infection* **18**, 1–6 (2012).

39. Mendelsohn, J. A. 'Like All That Lives': Biology, Medicine and Bacteria in the Age of Pasteur and Koch. *History and Philosophy of the Life Sciences* **24**, 3–36 (2002).

40. Caumette, P., Bertrand, J.-C. & Normand, P. Some Historical Elements of Microbial Ecology. in *Environmental Microbiology: Fundamentals and Applications: Microbial Ecology* (eds. Bertrand, J.-C. et al.) 9–24 (Springer Netherlands, 2015). doi:10.1007/978-94-017-9118-2_2.

41. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, (2016).

42. Djurhuus, A. *et al.* Evaluation of Filtration and DNA Extraction Methods for Environmental DNA Biodiversity Assessments across Multiple Trophic Levels. *Front. Mar. Sci.* **4**, (2017).

43. Hinlo, R., Gleeson, D., Lintermans, M. & Furlan, E. Methods to maximise recovery of environmental DNA from water samples. *PLOS ONE* **12**, e0179251 (2017).

44. Shahraki, A. H., Chaganti, S. R. & Heath, D. Assessing high-throughput environmental DNA extraction methods for meta-barcode characterization of aquatic microbial communities. *J Water Health* **17**, 37–49 (2019).

45. Hunter, M. E., Ferrante, J. A., Meigs-Friend, G. & Ulmer, A. Improving eDNA yield and inhibitor reduction through increased water volumes and multi-filter isolation techniques. *Sci Rep* **9**, 1–9 (2019).

46. Thomsen, P. F. & Willerslev, E. Environmental DNA – An emerging tool in conservation for monitoring past and present biodiversity. *Biological Conservation* **183**, 4–18 (2015).

47. Bik, H. M. *et al.* Metagenetic community analysis of microbial eukaryotes illuminates biogeographic patterns in deep-sea and shallow water sediments. *Molecular Ecology* **21**, 1048–1059 (2012).

48. Obbels, D. *et al.* Bacterial and eukaryotic biodiversity patterns in terrestrial and aquatic habitats in the Sør Rondane Mountains, Dronning Maud Land, East Antarctica. *FEMS microbiology ecology* fiw041–fiw041 (2016) doi:10.1093/femsec/fiw041.

49. Mesa, V. *et al.* Bacterial, Archaeal, and Eukaryotic Diversity across Distinct Microhabitats in an Acid Mine Drainage. *Frontiers in Microbiology* **8**, 1756–1756 (2017).

50. Bates, S. T. *et al.* Global biogeography of highly diverse protistan communities in soil. *The ISME journal* **7**, 652–9 (2013).

51. Geisen, S. *et al.* Metatranscriptomic census of active protists in soils. *The ISME journal* (2015) doi:10.1038/ismej.2015.30.

52. Planes, S. *et al.* The Tara Pacific expedition—A pan-ecosystemic approach of the "-omics" complexity of coral reef holobionts across the Pacific Ocean. *PLOS Biology* **17**, e3000483 (2019).

53. Flegontova, O. *et al.* Extreme Diversity of Diplonemid Eukaryotes in the Ocean. *Current Biology* **26**, 3060–3065 (2016).

54. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7**, 335–336 (2010).

55. Schloss, P. D. *et al.* Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Applied and Environmental Microbiology* **75**, 7537–7541 (2009).

56. Boscaro, V., Syberg-Olsen, M. J., Irwin, N. A. T., Campo, J. del & Keeling, P. J. What Can Environmental Sequences Tell Us About the Distribution of Low-Rank Taxa? The Case of Euplotes (Ciliophora, Spirotrichea), Including a Description of Euplotes enigma sp. nov. *Journal of Eukaryotic Microbiology* **66**, 281–293 (2019).

57. Flegontova, O. *et al.* Neobodonids are dominant kinetoplastids in the global ocean. *Environmental Microbiology* **20**, 878–889 (2018).

58. Tashyreva, D. *et al.* Phylogeny and Morphology of New Diplonemids from Japan. *Protist* **169**, 158–179 (2018).

59. Geisen, S. *et al.* A methodological framework to embrace soil biodiversity. *Soil Biology and Biochemistry* 107536 (2019) doi:10.1016/j.soilbio.2019.107536.

60. Stoeck, T. *et al.* Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Molecular Ecology* **19**, 21–31 (2010).

61. Tanabe, A. S. *et al.* Comparative study of the validity of three regions of the 18S-rRNA gene for massively parallel sequencing-based monitoring of the planktonic eukaryote community. *Molecular ecology resources* (2015) doi:10.1111/1755-0998.12459.

62. Hu, S. K. *et al.* Estimating Protistan Diversity using High-throughput Sequencing. *Journal of Eukaryotic Microbiology* n/a-n/a (2015) doi:10.1111/jeu.12217.

63. Wylezich, C., Herlemann, D. P. R. & Jürgens, K. Improved 18S rDNA amplification protocol for assessing protist diversity in oxygen-deficient marine systems. *Aquatic Microbial Ecology* **81**, 83–94 (2018).

64. Jamy, M. *et al.* Long metabarcoding of the eukaryotic rDNA operon to phylogenetically and taxonomically resolve environmental diversity. *bioRxiv* 627828 (2019) doi:10.1101/627828.

65. Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology* **17**, 239 (2016).

66. Schlee, D. Review of Numerical Taxonomy. The Principles and Practice of Numerical Classification. *Systematic Zoology* **24**, 263–268 (1975).

67. Mahé, F., Rognes, T., Quince, C., de Vargas, C. & Dunthorn, M. Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ* **3**, (2015).

68. Konstantinidis, K. T. & Tiedje, J. M. Genomic Insights That Advance the Species Definition for Prokaryotes. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 2567–2572 (2005).

69. Nguyen, N.-P., Warnow, T., Pop, M. & White, B. A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity. *npj Biofilms Microbiomes* **2**, 1–8 (2016).

70. Edgar, R. C. Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics* **34**, 2371–2375 (2018).

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

72. Callahan, B. J., McMurdie, P. J. & Holmes, S. P. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J* **11**, 2639–2643 (2017).

73. Forster, D. *et al.* Improving eDNA-based protist diversity assessments using networks of amplicon sequence variants. *Environmental Microbiology* **0**,.

74. OPEC. *OPEC : World Oil Outlook*. (2018).

75. Canadian Association of Petroleum Producers. *Crude Oil Forecast*. https://www.capp.ca/resources/crude-oil-forecast/.

76. Neilson, E. W. & Boutin, S. Human disturbance alters the predation rate of moose in the Athabasca oil sands. *Ecosphere* **8**, e01913 (2017).

77. Sogin, M. L. *et al.* Microbial diversity in the deep sea and the underexplored "rare biosphere". *PNAS* **103**, 12115–12120 (2006).

78. Bradley, M. & Neufeld, L. Climate and management interact to explain the decline of

woodland caribou (Rangifer tarandus caribou) in Jasper National Park. *Rangifer* 183–191 (2012) doi:10.7557/2.32.2.2268.

79. Rivera-Perez, J. I., Santiago-Rodriguez, T. M. & Toranzos, G. A. Paleomicrobiology: a Snapshot of Ancient Microbes and Approaches to Forensic Microbiology. *Microbiol Spectr* **4**, (2016).

80. Colwell, R. R. *et al.* Gulf of Mexico Research Initiative. (2010).

81. Bretherton, L. *et al.* Physiological response of 10 phytoplankton species exposed to macondo oil and the dispersant, Corexit. *Journal of Phycology* **54**, 317–328 (2018).

82. Ozhan, K., Parsons, M. L. & Bargu, S. How Were Phytoplankton Affected by the Deepwater Horizon Oil Spill? *BioScience* **64**, 829–836 (2014).

83. Lara, E., Berney, C., Harms, H. & Chatzinotas, A. Cultivation-independent analysis reveals a shift in ciliate 18S rRNA gene diversity in a polycyclic aromatic hydrocarbon-polluted soil. *FEMS microbiology ecology* **62**, 365–73 (2007).

84. Lanzén, A., Lekang, K., Jonassen, I., Thompson, E. M. & Troedsson, C. High-throughput metabarcoding of eukaryotic diversity for environmental monitoring of offshore oil-drilling activities. *Molecular Ecology* **25**, 4392–4406 (2016).

85. Rahsepar, S., Smit, M. P. J., Murk, A. J., Rijnaarts, H. H. M. & Langenhoff, A. A. M. Chemical dispersants: Oil biodegradation friend or foe? *Marine pollution bulletin* (2016) doi:10.1016/j.marpolbul.2016.04.044.

86. Snyder, R. A. *et al.* Polycyclic aromatic hydrocarbon concentrations across the Florida Panhandle continental shelf and slope after the BP MC 252 well failure. *Marine pollution bulletin* **89**, 201–8 (2014).

87. Rodriguez-R, L. M. *et al.* Microbial community successional patterns in beach sands impacted by the Deepwater Horizon oil spill. *The ISME journal* (2015) doi:10.1038/ismej.2015.5.

88. Alberta Energy Regulator. *Historical Overview of the Fort McMurray Area and Oil Sands Industry in Northeast Alberta (with expanded bibliographies on oil sands, surficial geology, hydrogeology, minerals and bedrock in Northeast Alberta)*. https://ags.aer.ca/publications/ESR_2000_05.html.

89. Brandt, J. P., Flannigan, M. D., Maynard, D. G., Thompson, I. D. & Volney, W. J. A. An introduction to Canada's boreal zone: ecosystem processes, health, sustainability, and environmental issues. *Environmental Reviews* **21**, 207–226 (2013).

90. Government of Alberta. *Oil sands facts and statistics*. https://www.alberta.ca/oil-sands-facts-and-statistics.aspx.

91. Government of Alberta. *Oil sands 101*. https://www.alberta.ca/oil-sands-101.aspx.

92. Jordaan, S. M., Keith, D. W. & Stelfox, B. Quantifying land use of oil sands production: a life cycle perspective. *Environ. Res. Lett.* **4**, 024004 (2009).

93. Government of Alberta. *End Pit Lakes Guidance Document*. http://library.cemaonline.ca/ckan/dataset/a5a7f266-44b4-44e2-babe-e6798ec612e2/resource/1632ce6e-d1a0-441a-a026-8a839f1d64bc/download/eplguidance2012jan23a.pdf (2012).

94. Foght, J. M., Gieg, L. M. & Siddique, T. The microbiology of oil sands tailings: past, present, future. *FEMS Microbiology Ecology* **93**, (2017).

95. Siddique, T., Stasik, S., Mohamad Shahimin, M. F. & Wendt-Potthoff, K. Microbial Communities in Oil Sands Tailings: Their Implications in Biogeochemical Processes and Tailings Management. in *Microbial Communities Utilizing Hydrocarbons and Lipids: Members,*

Metagenomics and Ecophysiology 1–33 (Springer International Publishing, 2018). doi:10.1007/978-3-319-60063-5_10-1.

96. Mahaffey, M. A. & Dube, Dr. M. Review of the composition and toxicity of oil sands process-affected water. *http://dx.doi.org/10.1139/er-2015-0060* (2016).

97. Alberta Energy Regulator. *Directive 085: Fluid Tailings Management for Oil Sands Mining Projects*.

98. Mossop, G. D. Geology of the Athabasca Oil Sands. *Science* **207**, 145–152 (1980).

99. Yergeau, E. *et al.* Next-generation sequencing of microbial communities in the Athabasca River and its tributaries in relation to oil sands mining activities. *Applied and environmental microbiology* **78**, 7626–37 (2012).

100. Reid, T., Chaganti, S. R., Droppo, I. G. & Weisener, C. G. Novel insights into freshwater hydrocarbon-rich sediments using metatranscriptomics: Opening the black box. *Water Research* **136**, 1–11 (2018).

101. Reid, T., Droppo, I. G., Chaganti, S. R. & Weisener, C. G. Microbial metabolic strategies for overcoming low-oxygen in naturalized freshwater reservoirs surrounding the Athabasca Oil Sands: A proxy for End-Pit Lakes? *Science of The Total Environment* **665**, 113–124 (2019).

102. Wong, M.-L. *et al.* Roles of Thermophiles and Fungi in Bitumen Degradation in Mostly Cold Oil Sands Outcrops. *Applied and Environmental Microbiology* **81**, 6825–6838 (2015).

103. Gieg, L. M. Microbial Communities in Oil Shales, Biodegraded and Heavy Oil Reservoirs, and Bitumen Deposits. in *Microbial Communities Utilizing Hydrocarbons and Lipids: Members, Metagenomics and Ecophysiology* 1–21 (Springer International Publishing, 2018). doi:10.1007/978-3-319-60063-5_4-1.

104. Ridley, C. M. & Voordouw, G. Aerobic microbial taxa dominate deep subsurface cores from the Alberta oil sands. *FEMS Microbiology Ecology* **94**, (2018).

105. An, D. *et al.* Metagenomics of Hydrocarbon Resource Environments Indicates Aerobic Taxa and Genes to be Unexpectedly Common. *Environmental Science & Technology* **47**, 10708–10717 (2013).

106. Yi, Z. *et al.* High-throughput sequencing of microbial eukaryotes in Lake Baikal reveals ecologically differentiated communities and novel evolutionary radiations. *FEMS Microbiology Ecology* **93**, (2017).

107. Bielewicz, S. *et al.* Protist diversity in a permanently ice-covered Antarctic lake during the polar night transition. *The ISME journal* **5**, 1559–64 (2011).

108. Amato, P. *et al.* Active microorganisms thrive among extremely diverse communities in cloud water. *PLOS ONE* **12**, e0182869–e0182869 (2017).

109. Saidi-Mehrabad, A. *et al.* Methanotrophic bacteria in oilsands tailings ponds of northern Alberta. *The ISME journal* **7**, 908–21 (2013).

110. Schuster, F. L. & Visvesvara, G. S. Free-living amoebae as opportunistic and non-opportunistic pathogens of humans and animals. *International Journal for Parasitology* **34**, 1001–1027 (2004).

111. Meyer, C. *et al.* Using testate amoeba as potential biointegrators of atmospheric deposition of phenanthrene (polycyclic aromatic hydrocarbon) on "moss/soil interface-testate amoeba community" microecosystems. *Ecotoxicology* **22**, 287–294 (2013).

112. Aguilar, M. *et al.* Next-generation Sequencing Assessment of Eukaryotic Diversity in Oil Sands Tailings Ponds Sediments and Surface Water. *The Journal of eukaryotic microbiology* (2016) doi:10.1111/jeu.12320.

113. Fung, M. Y. P. & Macyk, T. M. Reclamation of Oil Sands Mining Areas. *Reclamation of Drastically Disturbed Lands* **agronomymonogra**, 755–774 (2000).

114. Leung, S. S., MacKinnon, M. D. & Smith, R. E. H. The ecological effects of naphthenic acids and salts on phytoplankton from the Athabasca oil sands region. *Aquatic Toxicology* **62**, 11–26 (2003).

115. Woodward, F. I. Global primary production. *Current Biology* **17**, R269–R273 (2007).

116. Ruffell, S. E. *et al.* Assessing the bioremediation potential of algal species indigenous to oil sands process-affected waters on mixtures of oil sands acid extractable organics. *Ecotoxicology and Environmental Safety* **133**, 373–380 (2016).

117. Dompierre, K. A. & Barbour, S. L. Characterization of physical mass transport through oil sands fluid fine tailings in an end pit lake: a multi-tracer study. *Journal of Contaminant Hydrology* (2016) doi:10.1016/j.jconhyd.2016.03.006.

118. White, K. B. & Liber, K. Early chemical and toxicological risk characterization of inorganic constituents in surface water from the Canadian oil sands first large-scale end pit lake. *Chemosphere* **211**, 745–757 (2018).

119. Risacher. The interplay of methane and ammonia as key oxygen consuming constituents in early stage development of Base Mine Lake, the first demonstration oil sands pit lake. *Applied Geochemistry* **93**, 49–59 (2018).

120. Arriaga, D. *et al.* The co-importance of physical mixing and biogeochemical consumption in controlling water cap oxygen levels in Base Mine Lake. *Applied Geochemistry* 104442 (2019) doi:10.1016/j.apgeochem.2019.104442.

121. Rudderham, S. B. 1993-. Geomicrobiology and geochemistry of fluid fine tailings in an oil sands end pit lake. (University of Saskatchewan, 2019).

122. Rochman, F. F. *et al.* Novel copper-containing membrane monooxygenases (CuMMOs) encoded by alkane-utilizing Betaproteobacteria. *The ISME Journal* 1–13 (2019) doi:10.1038/s41396-019-0561-2.

123. Kong, J. D. *et al.* Second-generation stoichiometric mathematical model to predict methane emissions from oil sands tailings. *arXiv:1907.00247 [q-bio]* (2019).

124. Poon Ho Yin, Brandon Jordan T., Yu Xiaoxuan & Ulrich Ania C. Turbidity Mitigation in an Oil Sands Pit Lake through pH Reduction and Fresh Water Addition. *Journal of Environmental Engineering* **144**, 04018127 (2018).

125. Alharbi, H. A., Alcorn, J., Al-Mousa, A., Giesy, J. P. & Wiseman, S. B. Toxicokinetics and toxicodynamics of chlorpyrifos is altered in embryos of Japanese medaka exposed to oil sands process-affected water: evidence for inhibition of P-glycoprotein. *Journal of Applied Toxicology* **37**, 591–601 (2017).

126. Mueller, Z. The impacts of metal and salts similar in composition to Oil sands processes affected water (OSPW) on Rainbow trout respirometry, gill structure, and gill enzyme dynamics. *ERA* https://era.library.ualberta.ca/items/5cc343c8-f269-4f80-a375-0e88a0fc925c (2018) doi:10.7939/R39S1M21S.

127. Marentette, J. R. *et al.* Molecular responses of Walleye (Sander vitreus) embryos to naphthenic acid fraction components extracted from fresh oil sands process-affected water. *Aquatic Toxicology* **182**, 11–19 (2017).

128. Morandi, G. D. *et al.* Effect of Lipid Partitioning on Predictions of Acute Toxicity of Oil Sands Process Affected Water to Embryos of Fathead Minnow (Pimephales promelas). *Environ. Sci. Technol.* **50**, 8858–8866 (2016).

129. Sun, J. *et al.* Identification of Chemicals that Cause Oxidative Stress in Oil Sands Process-

Affected Water. *Environ. Sci. Technol.* **51**, 8773–8781 (2017).

130. Yu, X., Lee, K., Ma, B., Asiedu, E. & Ulrich, A. C. Indigenous microorganisms residing in oil sands tailings biodegrade residual bitumen. *Chemosphere* **209**, 551–559 (2018).

131. Yu, X., Lee, K. & Ulrich, A. C. Model naphthenic acids removal by microalgae and Base Mine Lake cap water microbial inoculum. *Chemosphere* (2019) doi:10.1016/j.chemosphere.2019.06.110.

132. Suncor. *Report on Sustainability*. https://sustainability.suncor.com/en (2019).

133. Peng, H. *et al.* Peroxisome Proliferator-Activated Receptor γ is a Sensitive Target for Oil Sands Process-affected Water: Effects on Adipogenesis and Identification of Ligands. *Environmental science & technology* (2016) doi:10.1021/acs.est.6b01890.

134. Pesce, S., Ghiglione, J.-F. & Martin-Laurent, F. Microbial Communities as Ecological Indicators of Ecosystem Recovery Following Chemical Pollution. in *Microbial Ecotoxicology* (eds. Cravo-Laureau, C., Cagnon, C., Lauga, B. & Duran, R.) 227–250 (Springer International Publishing, 2017). doi:10.1007/978-3-319-61795-4_10.

135. Rogerson, A. & Berger, J. Ultrastructural Modification of the Ciliate Protozoan, Colpidium colpoda, Following Chronic Exposure to Partially Degraded Crude Oil. *Transactions of the American Microscopical Society* **101**, 27–35 (1982).

136. Kota, S., Borden, R. C. & Barlaz, M. A. Influence of protozoan grazing on contaminant biodegradation. *FEMS Microbiology Ecology* **29**, 179–189 (1999).

137. Tso, S.-F. & Taghon, G. L. Protozoan grazing increases mineralization of naphthalene in marine sediment. *Microbial ecology* **51**, 460–9 (2006).

138. Gilbert, D., Jakobsen, H. H., Winding, A. & Mayer, P. Co-transport of polycyclic aromatic hydrocarbons by motile microorganisms leads to enhanced mass transfer under diffusive conditions. *Environmental science & technology* **48**, 4368–75 (2014).

139. Stoeck, T. & Edgcomb, V. Role of Protists in Microbial Interactions with Hydrocarbons. in *Handbook of Hydrocarbon and Lipid Microbiology* 2423–2434 (Springer Berlin Heidelberg, 2010). doi:10.1007/978-3-540-77587-4_178.

140. Caron, D. A. Mixotrophy stirs up our understanding of marine food webs. *Proceedings of the National Academy of Sciences of the United States of America* **113**, 2806–8 (2016).

141. Foissner, W. Protists as bioindicators in activated sludge: Identification, ecology and future needs. *European Journal of Protistology* **55**, 75–94 (2016).

142. Howe, A. T., Bass, D., Vickerman, K., Chao, E. E. & Cavalier-Smith, T. Phylogeny, Taxonomy, and Astounding Genetic Diversity of Glissomonadida ord. nov., The Dominant Gliding Zooflagellates in Soil (Protozoa: Cercozoa). *Protist* **160**, 159–189 (2009).

143. Energy Regulator, A. Directive 085: Fluid Tailings Management for Oil Sands Mining Projects.

144. Syncrude. The past, present and future of tailings at Syncrude. in (2008).

145. White, K. B. & Liber, K. Early chemical and toxicological risk characterization of inorganic constituents in surface water from the Canadian oil sands first large-scale end pit lake. *Chemosphere* **211**, 745–757 (2018).

146. Alberta Agriculture and Forestry. *A Review of the 2016 Horse River Wildfire Alberta Agriculture and Forestry Preparedness and Response*. https://wildfire.alberta.ca/resources/reviews/docum

ents/2016HorseRiverWildfireReview-Mar2017.pdf (2017).

147. Lax, G. *et al.* Hemimastigophora is a novel supra-kingdom-level lineage of eukaryotes. *Nature; London* **564**, 410 (2018).

148. Gawryluk, R. M. R. *et al.* Non-photosynthetic predators are sister to red algae. *Nature* **572**, 240–243 (2019).

149. Aguilar, M. *et al.* Next-generation Sequencing Assessment of Eukaryotic Diversity in Oil Sands Tailings Ponds Sediments and Surface Water. *The Journal of eukaryotic microbiology* (2016) doi:10.1111/jeu.12320.

150. de Vargas, C. *et al.* Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. *Science (New York, N.Y.)* **348**, 1261605 (2015).

151. Bates, S. T. *et al.* Global biogeography of highly diverse protistan communities in soil. *The ISME journal* **7**, 652–9 (2013).

152. Geisen, S. *et al.* Metatranscriptomic census of active protists in soils. *The ISME journal* (2015) doi:10.1038/ismej.2015.30.

153. Colwell, R. R. *et al.* Gulf of Mexico Research Initiative. (2010).

154. Özhan, K., Miles, S. M., Gao, H. & Bargu, S. Relative phytoplankton growth responses to physically and chemically dispersed South Louisiana sweet crude oil. *Environmental monitoring and assessment* **186**, 3941–56 (2014).

155. Gertler, C., Näther, D. J., Gerdts, G., Malpass, M. C. & Golyshin, P. N. A mesocosm study of the changes in marine flagellate and ciliate communities in a crude oil bioremediation trial. *Microbial ecology* **60**, 180–91 (2010).

156. Leung, S. S., MacKinnon, M. D. & Smith, R. E. H. The ecological effects of naphthenic acids and salts on phytoplankton from the Athabasca oil sands region. *Aquatic Toxicology* **62**, 11–26 (2003).

157. González, J. *et al.* Effect of a simulated oil spill on natural assemblages of marine phytoplankton enclosed in microcosms. *Estuarine, Coastal and Shelf Science* **83**, 265–276 (2009).

158. Sargian, P., Mas, S., Pelletier, É. & Demers, S. Multiple stressors on an Antarctic microplankton assemblage: water soluble crude oil and enhanced UVBR level at Ushuaia (Argentina). *Polar Biology* **30**, 829–841 (2007).

159. Mahdavi, H., Prasad, V., Liu, Y. & Ulrich, A. C. In situ biodegradation of naphthenic acids in oil sands tailings pond water using indigenous algae-bacteria consortium. *Bioresource Technology* (2015) doi:10.1016/j.biortech.2015.03.091.

160. Al-Hawash, A. B. *et al.* Isolation and characterization of two crude oil-degrading fungi strains from Rumaila oil field, Iraq. *Biotechnology Reports* **17**, 104–109 (2018).

161. Tso, S.-F. & Taghon, G. L. Protozoan grazing increases mineralization of naphthalene in marine sediment. *Microbial ecology* **51**, 460–9 (2006).

162. Gilbert, D., Jakobsen, H. H., Winding, A. & Mayer, P. Co-transport of polycyclic aromatic hydrocarbons by motile microorganisms leads to enhanced mass transfer under diffusive conditions. *Environmental science & technology* **48**, 4368–75 (2014).

163. Debastiani, C., Meira, B. R., Lansac-Tôha, F. M., Velho, L. F. M. & Lansac-Tôha, F. A. Protozoa ciliates community structure in urban streams and their environmental use as indicators. *Brazilian journal of biology = Revista brasleira de biologia* (2016) doi:10.1590/1519-6984.08615.

164. Xu, G. & Xu, H. Can annual cyclicity of protozoan communities reflect water quality status in coastal ecosystems? *Ecological Indicators* **67**, 730–734 (2016).

165. Fang, L. *et al.* De novo transcriptomic profiling of Dunaliella salina reveals concordant flows of glycerol metabolic pathways upon

reciprocal salinity changes. *Algal Research* **23**, 135–149 (2017).

166. Moreno-Sánchez, R., Rodríguez-Enríquez, S., Jasso-Chávez, R., Saavedra, E. & García-García, J. D. Biochemistry and Physiology of Heavy Metal Resistance and Accumulation in Euglena. in *Advances in experimental medicine and biology* vol. 979 91–121 (2017).

167. Aguilera, A. Eukaryotic organisms in extreme acidic environments, the río tinto case. *Life (Basel, Switzerland)* **3**, 363–74 (2013).

168. Foght, J. M., Gieg, L. M. & Siddique, T. The microbiology of oil sands tailings: past, present, future. *FEMS Microbiology Ecology* **93**, (2017).

169. Syncrude. *Alberta issues first-ever oil sands land reclamation certificate*. http://www.syncrude.ca/users/folder.asp?FolderID=5703 (2008).

170. Dompierre, K. A. & Barbour, S. L. Characterization of physical mass transport through oil sands fluid fine tailings in an end pit lake: a multi-tracer study. *Journal of Contaminant Hydrology* (2016) doi:10.1016/j.jconhyd.2016.03.006.

171. End Pit Lakes Guidance Document - 2012 - End Pit Lakes Guidance Document 2012 - CEMA. http://library.cemaonline.ca/ckan/dataset/2010-0016/resource/1632ce6e-d1a0-441a-a026-8a839f1d64bc.

172. Risacher, F. F. *et al.* The interplay of methane and ammonia as key oxygen consuming constituents in early stage development of Base Mine Lake, the first demonstration oil sands pit lake. *Applied Geochemistry* **93**, 49–59 (2018).

173. Hurley, D. L. Wind waves and internal waves in Base Mine Lake. (2017) doi:10.14288/1.0351993.

174. Tedford, E., Halferdahl, G., Pieters, R. & Lawrence, G. A. Temporal variations in turbidity in an oil sands pit lake. *Environmental Fluid Mechanics* 1–17 (2018) doi:10.1007/s10652-018-9632-6.

175. Ohiozebau, E. *et al.* Potential health risks posed by polycyclic aromatic hydrocarbons in muscle tissues of fishes from the Athabasca and Slave Rivers, Canada. *Environmental geochemistry and health* (2016) doi:10.1007/s10653-016-9815-3.

176. Korosi, J. B. *et al.* Examining spatial patterns in polycyclic aromatic compounds measured in stream macroinvertebrates near a small subarctic oil and gas operation. *Environmental monitoring and assessment* **188**, 189 (2016).

177. Mohseni, P. *et al.* Naphthenic Acid Mixtures from Oil Sands Process-affected Water Enhance Differentiation of Mouse Embryonic Stem Cells and Affect Development of the Heart. *Environmental science & technology* (2015) doi:10.1021/acs.est.5b02267.

178. Alharbi, H. A., Alcorn, J., Al-Mousa, A., Giesy, J. P. & Wiseman, S. B. Toxicokinetics and toxicodynamics of chlorpyrifos is altered in embryos of Japanese medaka exposed to oil sands process-affected water: evidence for inhibition of P-glycoprotein. *Journal of Applied Toxicology* **37**, 591–601 (2017).

179. Marentette, J. R. *et al.* Molecular responses of Walleye (Sander vitreus) embryos to naphthenic acid fraction components extracted from fresh oil sands process-affected water. *Aquatic Toxicology* **182**, 11–19 (2017).

180. Stoeck, T. *et al.* Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Molecular ecology* **19 Suppl 1**, 21–31 (2010).

181. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).

182. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data

processing and web-based tools. *Nucleic acids research* **41**, D590-6 (2013).

183. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

184. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics (Oxford, England)* **30**, 1312–3 (2014).

185. Ondov, B. D., Bergman, N. H. & Phillippy, A. M. Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics* **12**, 385 (2011).

186. Matsen, F. A., Kodner, R. B. & Armbrust, E. V. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC bioinformatics* **11**, 538 (2010).

187. Han, M. V & Zmasek, C. M. phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics* **10**, 356 (2009).

188. Katoh, K., Rozewicki, J. & Yamada, K. D. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics* (2017) doi:10.1093/bib/bbx108.

189. Miller, M. A., Miller, M. A., Pfeiffer, W. & Schwartz, T. Creating the CIPRES Science Gateway for Inference of Large Phylogenetic Trees.

190. Dixon, P. VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science* **14**, 927–930 (2003).

191. Wickham, Hadley. *Ggplot2 : elegant graphics for data analysis*. (Springer, 2009).

192. Slapeta, J., Moreira, D. & López-García, P. The extent of protist diversity: insights from molecular ecology of freshwater eukaryotes. *Proceedings. Biological sciences / The Royal Society* **272**, 2073–81 (2005).

193. del Campo, J. *et al.* Ecological and evolutionary significance of novel protist lineages. *European Journal of Protistology* (2016) doi:10.1016/j.ejop.2016.02.002.

194. Seeleuthner, Y. *et al.* Single-cell genomics of multiple uncultured stramenopiles reveals underestimated functional diversity across oceans. *Nature Communications* **9**, 310 (2018).

195. Reynolds, C. S., Watanabe, Yasunori. & International Congress in Ecology (5th : 1990 : Yokohama, J. *Vertical structure in aquatic environments and its impact on trophic linkages and nutrient fluxes*. (E. Schweizerbart'sche Verlagsbuchhandlung, 1992).

196. Goldman, C. R. *et al.* Thermal stratification, nutrient dynamics, and phytoplankton productivity during the onset of spring phytoplankton growth in Lake Baikal, Russia. *Hydrobiologia* **331**, 9–24 (1996).

197. *The Ciliated Protozoa*. (Springer Netherlands, 2010). doi:10.1007/978-1-4020-8239-9.

198. Richardson, E. *et al.* Next-Generation Sequencing of Protists as a Measure of Natural Soil Microbial Eukaryotic Community in the Oil Sands Region. (2014).

199. Bass, D. *et al.* Clarifying the Relationships between Microsporidia and Cryptomycota. *Journal of Eukaryotic Microbiology* (2018) doi:10.1111/jeu.12519.

200. Ortiz-Álvarez, R., Triadó-Margarit, X., Camarero, L., Casamayor, E. O. & Catalan, J. High planktonic diversity in mountain lakes contains similar contributions of autotrophic, heterotrophic and parasitic eukaryotic life forms. *Scientific Reports* **8**, 4457 (2018).

201. Giner, C. R. *et al.* Quantifying long-term recurrence in planktonic microbial eukaryotes. *Molecular Ecology* **28**, 923–935 (2019).

202. Mukherjee, I., Hodoki, Y. & Nakano, S.-I. Kinetoplastid flagellates overlooked by universal primers dominate in the oxygenated hypolimnion of Lake Biwa, Japan. *FEMS microbiology ecology* fiv083 (2015) doi:10.1093/femsec/fiv083.

203. Unrein, F., Gasol, J. M., Not, F., Forn, I. & Massana, R. Mixotrophic haptophytes are key bacterial grazers in oligotrophic coastal waters. *The ISME Journal* **8**, 164–176 (2014).

204. Tedford, E., Halferdahl, G., Pieters, R. & Lawrence, G. A. Temporal variations in turbidity in an oil sands pit lake. *Environmental Fluid Mechanics* 1–17 (2018) doi:10.1007/s10652-018-9632-6.

205. White, K. B. & Liber, K. Early chemical and toxicological risk characterization of inorganic constituents in surface water from the Canadian oil sands first large-scale end pit lake. *Chemosphere* **211**, 745–757 (2018).

206. Huang, J. B. *et al.* Further insights into the highly derived haptorids (Ciliophora, Litostomatea): Phylogeny based on multigene data. *Zoologica Scripta* (2018) doi:10.1111/zsc.12269.

207. Saidi-Mehrabad, A. *et al.* Methanotrophic bacteria in oilsands tailings ponds of northern Alberta. *The ISME journal* **7**, 908–21 (2013).

208. Howe, A. T., Bass, D., Vickerman, K., Chao, E. E. & Cavalier-Smith, T. Phylogeny, Taxonomy, and Astounding Genetic Diversity of Glissomonadida ord. nov., The Dominant Gliding Zooflagellates in Soil (Protozoa: Cercozoa). *Protist* **160**, 159–189 (2009).

209. Howe, A. T., Bass, D., Chao, E. E. & Cavalier-Smith, T. New Genera, Species, and Improved Phylogeny of Glissomonadida (Cercozoa). *Protist* **162**, 710–722 (2011).

210. Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K. & Schloss, P. D. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Applied and environmental microbiology* **79**, 5112–20 (2013).

211. Grossmann, L., Bock, C., Schweikert, M. & Boenigk, J. Small but Manifold - Hidden Diversity in " *Spumella* -like Flagellates". *Journal of Eukaryotic Microbiology* **63**, 419–439 (2016).

212. Graham, M. D., Vinebrooke, R. D. & Turner, M. Coupling of boreal forests and lakes: Effects of conifer pollen on littoral communities. *Limnology and Oceanography* **51**, 1524–1529 (2006).

213. Freeman, K. R. *et al.* Evidence that chytrids dominate fungal communities in high-elevation soils. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 18315–20 (2009).

214. Wong, M.-L. *et al.* Roles of Thermophiles and Fungi in Bitumen Degradation in Mostly Cold Oil Sands Outcrops. *Applied and Environmental Microbiology* **81**, 6825–6838 (2015).

215. Kachieng'a, L. & Momba, M. The synergistic effect of a consortium of protozoan isolates (Paramecium sp., Vorticella sp., Epistylis sp. and Opercularia sp.) on the biodegradation of petroleum hydrocarbons in wastewater. *Journal of Environmental Chemical Engineering* (2018) doi:10.1016/J.JECE.2018.07.005.

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

217. Hehenberger, E. *et al.* Novel Predators Reshape Holozoan Phylogeny and Reveal the Presence of a Two-Component Signaling System in the Ancestor of Animals. *Current biology : CB* **27**, 2043-2050.e6 (2017).

218. Chen, L.-X. *et al.* Wide Distribution of Phage That Infect Freshwater SAR11 Bacteria. *mSystems* **4**, (2019).

219. Mori, J. F. *et al.* Putative Mixotrophic Nitrifying-Denitrifying Gammaproteobacteria Implicated in Nitrogen Cycling Within the Ammonia/Oxygen Transition Zone of an Oil Sands Pit Lake. *Front. Microbiol.* **10**, (2019).

220. Dompierre, K. A., Lindsay, M. B. J., Cruz-Hernández, P. & Halferdahl, G. M. Initial geochemical characteristics of fluid fine tailings in an oil sands end pit lake. *Science of The Total Environment* **556**, 196–206 (2016).

221. Canadian Association of Petroleum Producers. *Canada's Oil Sands Fact Book*. https://www.capp.ca/publications-and-statistics/publications/316441 (2018).

222. Clark, M. G. The initial biometerology of the constructed Sandhill Fen Watershed in Alberta, Canada. (Carleton University, 2018).

223. Lawrence, G. A., Ward, P. R. B. & MacKinnon, M. D. Wind-wave-induced suspension of mine tailings in disposal ponds – a case study. *Can. J. Civ. Eng.* **18**, 1047–1053 (1991).

224. Syncrude. Base Mine Lake Monitoring and Research Summary Report. (2019).

225. Brandon, J. T. Turbidity Mitigation in an Oil Sands End Pit Lake through pH Reduction and Fresh Water Addition. *ERA* https://era.library.ualberta.ca/items/3822f324-710d-4d0b-be45-a105508e79fa (2016) doi:10.7939/R3ST7F72D.

226. Bladon, K. D., Emelko, M. B., Silins, U. & Stone, M. Wildfire and the Future of Water Supply. *Environ. Sci. Technol.* **48**, 8936–8943 (2014).

227. Crouch, R. L., Timmenga, H. J., Barber, T. R. & Fuchsman, P. C. Post-fire surface water quality: comparison of fire retardant versus wildfire-related effects. *Chemosphere* **62**, 874–889 (2006).

228. Emelko, M. B. *et al.* Sediment-phosphorus dynamics can shift aquatic ecology and cause downstream legacy effects after wildfire in large river systems. *Glob Chang Biol* **22**, 1168–1184 (2016).

229. Dompierre, K. A., Barbour, S. L., North, R. L., Carey, S. K. & Lindsay, M. B. J. Chemical mass transport between fluid fine tailings and the overlying water cover of an oil sands end pit lake. *Water Resources Research* **53**, 4725–4740 (2017).

230. Risacher, F. F. *et al.* The interplay of methane and ammonia as key oxygen consuming constituents in early stage development of Base Mine Lake, the first demonstration oil sands pit lake. *Applied Geochemistry* **93**, 49–59 (2018).

231. Abbasian, F., Lockington, R., Megharaj, M. & Naidu, R. The integration of sequencing and bioinformatics in metagenomics. *Reviews in Environmental Science and Bio/Technology* (2015) doi:10.1007/s11157-015-9365-7.

232. Mori, J. F. *et al.* Putative mixotrophic nitrifying-denitrifying Gammaproteobacteria implicated in nitrogen cycling within the ammonia/oxygen transition zone of an oil sands pit lake. *Front. Microbiol.* **10**, (2019).

233. Aguilar, M. *et al.* Next-Generation Sequencing Assessment of Eukaryotic Diversity in Oil Sands Tailings Ponds Sediments and Surface Water. *Journal of Eukaryotic Microbiology* **63**, 732–743 (2016).

234. Richardson, E. *et al.* Next-Generation Sequencing of Protists as a Measure of Natural Soil Microbial Eukaryotic Community in the Oil Sands Region. (2014).

235. Richardson, E. *et al.* Phylogenetic Estimation of Community Composition and Novel Eukaryotic Lineages in Base Mine Lake: An Oil Sands Tailings Reclamation Site in Northern Alberta. *Journal of Eukaryotic Microbiology* **n/a**, (2019).

236. Richardson, E. & Dacks, J. B. Microbial Eukaryotes in Oil Sands Environments: Heterotrophs in the Spotlight. *Microorganisms* **7**, 178 (2019).

237. Huang, J. B. *et al.* Further insights into the highly derived haptorids (Ciliophora, Litostomatea): Phylogeny based on multigene data. *Zoologica Scripta* (2018) doi:10.1111/zsc.12269.

238. Debastiani, C., Meira, B. R., Lansac-Tôha, F. M., Velho, L. F. M. & Lansac-Tôha, F. A. Protozoa ciliates community structure in urban streams and their environmental use as indicators. *Brazilian journal of biology = Revista brasleira de biologia* (2016) doi:10.1590/1519-6984.08615.

239. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic acids research* **41**, D590-6 (2013).

240. Ondov, B. D., Bergman, N. H. & Phillippy, A. M. Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics* **12**, 385–385 (2011).

241. Matsen, F. A., Kodner, R. B. & Armbrust, E. V. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC bioinformatics* **11**, 538–538 (2010).

242. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics (Oxford, England)* **30**, 1312–3 (2014).

243. Miller, M. A., Miller, M. A., Pfeiffer, W. & Schwartz, T. Creating the CIPRES Science Gateway for Inference of Large Phylogenetic Trees.

244. Katoh, K., Rozewicki, J. & Yamada, K. D. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics* (2017) doi:10.1093/bib/bbx108.

245. Dixon, P. VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science* **14**, 927–930 (2003).

246. Wickham, Hadley. *Ggplot2 : elegant graphics for data analysis*. (Springer, 2009).

247. Baksi, K. D., Kuntal, B. K. & Mande, S. S. 'TIME': A Web Application for Obtaining Insights into Microbial Ecology Using Longitudinal Microbiome Data. *Front Microbiol* **9**, (2018).

248. Caporaso, J. G. *et al.* Moving pictures of the human microbiome. *Genome Biol* **12**, R50 (2011).

249. Albanese, D., Riccadonna, S., Donati, C. & Franceschi, P. A practical tool for maximal information coefficient analysis. *Gigascience* **7**, 1–8 (2018).

250. Berney, C. *et al.* UniEuk : Time to Speak a Common Language in Protistology! *Journal of Eukaryotic Microbiology* **64**, 407–411 (2017).

251. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421–421 (2009).

252. Logares, R. *et al.* Patterns of rare and abundant marine microbial eukaryotes. *Current biology : CB* **24**, 813–21 (2014).

253. Gertler, C., Näther, D. J., Gerdts, G., Malpass, M. C. & Golyshin, P. N. A mesocosm study of the changes in marine flagellate and ciliate communities in a crude oil bioremediation trial. *Microbial ecology* **60**, 180–91 (2010).

254. Moss, J. A. *et al.* Ciliated protists from the nepheloid layer and water column of sites affected by the Deepwater Horizon oil spill in the Northeastern Gulf of Mexico. *Deep Sea Research Part I: Oceanographic Research Papers* **106**, 85–96 (2015).

255. Rekik, A., Denis, M., Dugene, M., Maalej, S. & Ayadi, H. Journal of Oceanography, Research and Data JORD. **8**, (2015).

256. Ruehle, M. D. *et al.* Tetrahymena as a Unicellular Model Eukaryote: Genetic and Genomic Tools. *Genetics* **203**, 649–65 (2016).

257. Reynolds, C. S., Watanabe, Yasunori. & International Congress in Ecology (5th : 1990 :

Yokohama, J. *Vertical structure in aquatic environments and its impact on trophic linkages and nutrient fluxes*. (E. Schweizerbart'sche Verlagsbuchhandlung, 1992).

258. Davies, T. J., Urban, M. C., Rayfield, B., Cadotte, M. W. & Peres-Neto, P. R. Deconstructing the relationships between phylogenetic diversity and ecology: a case study on ecosystem functioning. *Ecology* (2016) doi:10.1002/ecy.1507.

259. Louca, S., Parfrey, L. W. & Doebeli, M. Decoupling function and taxonomy in the global ocean microbiome. *Science* **353**, 1272–1277 (2016).

260. Louca, S. *et al.* Function and functional redundancy in microbial systems. *Nat Ecol Evol* **2**, 936–943 (2018).

261. Needham, D. M. & Fuhrman, J. A. Pronounced daily succession of phytoplankton, archaea and bacteria following a spring bloom. *Nature Microbiology* 16005–16005 (2016) doi:10.1038/nmicrobiol.2016.5.

262. Geisen, S., Wall, D. H. & van der Putten, W. H. Challenges and Opportunities for Soil Biodiversity in the Anthropocene. *Current Biology* **29**, R1036–R1044 (2019).

263. Saleem, M., Hu, J. & Jousset, A. More Than the Sum of Its Parts: Microbiome Biodiversity as a Driver of Plant Growth and Soil Health. *Annual Review of Ecology, Evolution, and Systematics* **50**, 145–168 (2019).

264. Sparvoli, D. *et al.* Remodeling the Specificity of an Endosomal CORVET Tether Underlies Formation of Regulated Secretory Vesicles in the Ciliate Tetrahymena thermophila. *Current Biology* (2018) doi:10.1016/j.cub.2018.01.047.

265. Guerrier, S., Plattner, H., Richardson, E., Dacks, J. B. & Turkewitz, A. P. An evolutionary balance: conservation vs innovation in ciliate membrane trafficking. *Traffic* **18**, 18–28 (2017).

266. Zheng, W. *et al.* Insights into an Extensively Fragmented Eukaryotic Genome: De Novo Genome Sequencing of the Multinuclear Ciliate Uroleptopsis citrina. *Genome Biol Evol* **10**, 883–894 (2018).

267. Wang, R., Miao, W., Wang, W., Xiong, J. & Liang, A. EOGD: the Euplotes octocarinatus genome database. *BMC Genomics* **19**, (2018).

268. Abraham, J. S. *et al.* Soil ciliates of the Indian Delhi Region: Their community characteristics with emphasis on their ecological implications as sensitive bio-indicators for soil quality. *Saudi Journal of Biological Sciences* **26**, 1305–1313 (2019).

269. Lynn, D. H. *The ciliated protozoa : characterization, classification, and guide to the literature*. (Springer, 2008).

270. Foissner, W., Chao, A. & Katz, L. A. Diversity and geographic distribution of ciliates (Protista: Ciliophora). in 111–129 (Springer Netherlands, 2007). doi:10.1007/978-90-481-2801-3_9.

271. Arrigo, K. R. Marine microorganisms and global nutrient cycles. *Nature* **437**, 349 (2005).

272. Gimmler, A., Korn, R., de Vargas, C., Audic, S. & Stoeck, T. The Tara Oceans voyage reveals global diversity and distribution patterns of marine planktonic ciliates. *Scientific Reports* **6**, 1–13 (2016).

273. Posch, T. *et al.* Network of Interactions Between Ciliates and Phytoplankton During Spring. *Front Microbiol* **6**, (2015).

274. Rosetta, C. H. & McManus, G. B. Feeding by ciliates on two harmful algal bloom species, Prymnesium parvum and Prorocentrum minimum. *Harmful Algae* **2**, 109–126 (2003).

275. Arregui, L., Pérez-Uz, B., Salvadó, H. & Serrano, S. C. Progresses on the knowledge about the ecological function and structure of the protists

community in activated sludge wastewater treatment plants. (2010).

276. Puckett, G. L. & Carman, K. R. Ciliate epibiont effects on feeding, energy reserves, and sensitivity to hydrocarbon contaminants in an estuarine harpacticoid copepod. *Estuaries* **25**, 372–381 (2002).

277. Zaila, K. E., Cho, D. & Chang, W.-J. Interactions Between Parasitic Ciliates and Their Hosts: Ichthyophthirius multifiliis and Cryptocaryon irritans as Examples. in *Biocommunication of Ciliates* (eds. Witzany, G. & Nowacki, M.) 327–350 (Springer International Publishing, 2016). doi:10.1007/978-3-319-32211-7_18.

278. Lafferty, K. D. *et al.* Infectious diseases affect marine fisheries and aquaculture economics. *Ann Rev Mar Sci* **7**, 471–496 (2015).

279. Dziallas, C., Allgaier, M., Monaghan, M. T. & Grossart, H.-P. Act together—implications of symbioses in aquatic ciliates. *Front Microbiol* **3**, (2012).

280. Caron, D. A. & Sieburth, J. McN. Response of Peritrichous Ciliates in Fouling Communities to Seawater-Accommodated Hydrocarbons. *Transactions of the American Microscopical Society* **100**, 183–203 (1981).

281. Lara, E. & Acosta-Mercado, D. A molecular perspective on ciliates as soil bioindicators. *European Journal of Soil Biology* **49**, 107–111 (2012).

282. Ricard, G. *et al.* Horizontal gene transfer from Bacteria to rumen Ciliates indicates adaptation to their anaerobic, carbohydrates-rich environment. *BMC Genomics* **7**, 22 (2006).

283. Rogerson, A., Wan Ying Shiu, Guo Lan Huang, Mackay, D. & Berger, J. Determination and interpretation of hydrocarbon toxicity to ciliate protozoa. *Aquatic Toxicology* **3**, 215–228 (1983).

284. Somasundaram, S. *et al.* Cellular and molecular basis of heavy metal-induced stress in ciliates. *CURRENT SCIENCE* **114**, 8 (2018).

285. Martínez-López, A., Pérez-Morales, A., Ayala-Rodríguez, G. A., Escobedo-Urías, D. & Hakspiel-Segura, C. Abundance and seasonal variability of aloricate ciliates and tintinnids in a eutrophic coastal lagoon system of the Gulf of California, Mexico. *Regional Studies in Marine Science* 100814 (2019) doi:10.1016/j.rsma.2019.100814.

286. Bamdad, M., Reader, S., Grolière, C. A., Bohatier, J. & Denizeau, F. Uptake and efflux of polycyclic aromatic hydrocarbons by Tetrahymena pyriformis: Evidence for a resistance mechanism. *Cytometry* **28**, 170–175 (1997).

287. Al-Yamani, F., Madhusoodhanan, R., Skryabin, V. & Al-Said, T. The response of microzooplankton (tintinnid) community to salinity related environmental changes in a hypersaline marine system in the northwestern Arabian Gulf. *Deep Sea Research Part II: Topical Studies in Oceanography* (2019) doi:10.1016/J.DSR2.2019.02.005.

288. Aury, J.-M. *et al.* Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. *Nature* **444**, 171–178 (2006).

289. Eisen, J. A. *et al.* Macronuclear Genome Sequence of the Ciliate Tetrahymena thermophila, a Model Eukaryote. *PLOS Biology* **4**, e286 (2006).

290. Slabodnick, M. M. *et al.* The macronuclear genome of Stentor coeruleus reveals tiny introns in a giant cell. *Curr Biol* **27**, 569–575 (2017).

291. Aeschlimann, S. H. *et al.* The Draft Assembly of the Radically Organized Stylonychia lemnae Macronuclear Genome. *Genome Biol Evol* **6**, 1707–1723 (2014).

292. Swart, E. C. *et al.* The Oxytricha trifallax Macronuclear Genome: A Complex Eukaryotic

Genome with 16,000 Tiny Chromosomes. *PLOS Biology* **11**, e1001473 (2013).

293. Coyne, R. S. *et al.* Comparative genomics of the pathogenic ciliate Ichthyophthirius multifiliis, its free-living relatives and a host species provide insights into adoption of a parasitic lifestyle and prospects for disease control. *Genome Biology* **12**, R100 (2011).

294. Xiong, J. *et al.* Genome of the facultative scuticociliatosis pathogen Pseudocohnilembus persalinus provides insight into its virulence through horizontal gene transfer. *Scientific Reports* **5**, 1–12 (2015).

295. Yan, Y., Maurer-Alcalá, X. X., Knight, R., Pond, S. L. K. & Katz, L. A. Single-Cell Transcriptomics Reveal a Correlation between Genome Architecture and Gene Family Evolution in Ciliates. *mBio* **10**, (2019).

296. Maurer-Alcalá, X. X., Yan, Y., Pilling, O. A., Knight, R. & Katz, L. A. Twisted Tales: Insights into Genome Diversity of Ciliates Using Single-Cell 'Omics. *Genome Biol Evol* **10**, 1927–1938 (2018).

297. Blackburn, E. H., Greider, C. W. & Szostak, J. W. Telomeres and telomerase: the path from maize, Tetrahymena and yeast to human cancer and aging. *Nat. Med.* **12**, 1133–1138 (2006).

298. Boscaro, V. *et al.* Strengths and Biases of High-Throughput Sequencing Data in the Characterization of Freshwater Ciliate Microbiomes. *Microbial Ecology* 1–11 (2016) doi:10.1007/s00248-016-0912-8.

299. Chen, X. *et al.* The Architecture of a Scrambled Genome Reveals Massive Levels of Genomic Rearrangement during Development. *Cell* **158**, 1187–1198 (2014).

300. Hamilton, E. P. *et al.* Structure of the germline genome of Tetrahymena thermophila and relationship to the massively rearranged somatic genome. *eLife* **5**, (2016).

301. Briguglio, J. S. & Turkewitz, A. P. Tetrahymena thermophila: a divergent perspective on membrane traffic. *Journal of experimental zoology. Part B, Molecular and developmental evolution* **322**, 500–16 (2014).

302. Dacks, J. B. & Field, M. C. Evolution of the eukaryotic membrane-trafficking system: origin, tempo and mode. *Journal of cell science* **120**, 2977–85 (2007).

303. Bonifacino, J. S. & Glick, B. S. The Mechanisms of Vesicle Budding and Fusion. *Cell* **116**, 153–166 (2004).

304. Dacks, J. B. & Field, M. C. Eukaryotic Cell Evolution from a Comparative Genomic Perspective: The Endomembrane System.

305. Kugrens, P., Lee, R. E. & Corliss, J. O. Ultrastructure, biogenesis, and functions of extrusive organelles in selected non-ciliate protists. *Protoplasma* **181**, 164–190 (1994).

306. Rosati, G. & Modeo, L. Extrusomes in ciliates: diversification, distribution, and phylogenetic implications. *J. Eukaryot. Microbiol.* **50**, 383–402 (2003).

307. Hutton, J. C. Tetrahymena: The key to the genetic analysis of the regulated pathway of polypeptide secretion? *PNAS* **94**, 10490–10492 (1997).

308. Plattner, H. Trichocysts-Paramecium's Projectile-like Secretory Organelles: Reappraisal of their Biogenesis, Composition, Intracellular Transport, and Possible Functions. *J. Eukaryot. Microbiol.* **64**, 106–133 (2017).

309. Briguglio, J. S., Kumar, S. & Turkewitz, A. P. Lysosomal sorting receptors are essential for secretory granule biogenesis in Tetrahymena. *The Journal of cell biology* **203**, 537–50 (2013).

310. Rogerson, A. & Berger, J. Enhancement of the microbial degradation of crude oil by the ciliate Colpidium Colpoda. *The Journal of General and Applied Microbiology* **29**, 41–50 (1983).

311. Gomiero, A., Dagnino, A., Nasci, C. & Viarengo, A. The use of protozoa in ecotoxicology: Application of multiple endpoint tests of the ciliate E. crassus for the evaluation of sediment quality in coastal marine ecosystems. *Science of The Total Environment* **442**, 534–544 (2013).

312. Bamdad, M., Brousseau, P. & Denizeau, F. Identification of a multidrug resistance-like system in Tetrahymena pyriformis: evidence for a new detoxication mechanism in freshwater ciliates. *FEBS Letters* **456**, 389–393 (1999).

313. Ozhan, K. & Bargu, S. Distinct responses of Gulf of Mexico phytoplankton communities to crude oil and the dispersant corexit(®) Ec9500A under different nutrient regimes. *Ecotoxicology (London, England)* **23**, 370–84 (2014).

314. Yoon, B.-J. Hidden Markov Models and their Applications in Biological Sequence Analysis. *Curr Genomics* **10**, 402–415 (2009).

315. Ronquist, F. *et al.* MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. *Syst Biol* **61**, 539–542 (2012).

316. Miller, M. A., Pfeiffer, W. & Schwartz, T. The CIPRES Science Gateway: A Community Resource for Phylogenetic Analyses. in *Proceedings of the 2011 TeraGrid Conference: Extreme Digital Discovery* 41:1–41:8 (ACM, 2011). doi:10.1145/2016741.2016785.

317. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **32**, 1792–7 (2004).

318. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol Biol Evol* **32**, 268–274 (2015).

319. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Res* **42**, D222–D230 (2014).

320. Keeling, P. J. *et al.* The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLoS biology* **12**, e1001889–e1001889 (2014).

321. Gentekaki, E. *et al.* Large-scale phylogenomic analysis reveals the phylogenetic position of the problematic taxon Protocruzia and unravels the deep phylogenetic affinities of the ciliate lineages. *Molecular Phylogenetics and Evolution* **78**, 36–42 (2014).

322. Dunthorn, M. *et al.* Meiotic Genes In Colpodean Ciliates Support Secretive Sexuality. *bioRxiv* (2017).

323. Dacks, J. B. & Robinson, M. S. Outerwear through the ages: evolutionary cell biology of vesicle coats. *Curr. Opin. Cell Biol.* **47**, 108–116 (2017).

324. Boehm, M. & Bonifacino, J. S. Adaptins. The Final Recount. *Molecular Biology of the Cell* **12**, 2907–2920 (2001).

325. Sorkin, A. The endocytosis machinery. *Journal of Cell Science* **113**, 4375–4376 (2000).

326. van Bergen en Henegouwen, P. M. Eps15: a multifunctional adaptor protein regulating intracellular trafficking. *Cell Commun Signal* **7**, 24 (2009).

327. Bröcker, C., Engelbrecht-Vandré, S. & Ungermann, C. Multisubunit tethering complexes and their role in membrane fusion. *Curr. Biol.* **20**, R943-952 (2010).

328. Klinger, C. M., Klute, M. J. & Dacks, J. B. Comparative Genomic Analysis of Multi-Subunit Tethering Complexes Demonstrates an Ancient Pan-Eukaryotic Complement and Sculpting in Apicomplexa. *PLoS ONE* **8**, e76278–e76278 (2013).

329. Perry, R. J., Mast, F. D. & Rachubinski, R. A. Endoplasmic Reticulum-Associated Secretory

Proteins Sec20p, Sec39p, and Dsl1p Are Involved in Peroxisome Biogenesis. *Eukaryot Cell* **8**, 830–843 (2009).

330. Gabaldón, T. Evolution of the Peroxisomal Proteome. *Subcell. Biochem.* **89**, 221–233 (2018).

331. Gabaldón, T., Ginger, M. L. & Michels, P. A. M. Peroxisomes in parasitic protists. *Molecular and Biochemical Parasitology* **209**, 35–45 (2016).

332. Balderhaar, H. J. kleine & Ungermann, C. CORVET and HOPS tethering complexes – coordinators of endosome and lysosome fusion. *J Cell Sci* **126**, 1307–1316 (2013).

333. Voss, C., Fiore-Donno, A. M., Guerreiro, M. A., Peršoh, D. & Bonkowski, M. Metatranscriptomics reveals unsuspected protistan diversity in leaf litter across temperate beech forests, with Amoebozoa the dominating lineage. *FEMS Microbiol Ecol* doi:10.1093/femsec/fiz142.

334. Mukherjee, A. & Reddy, M. S. Metatranscriptomics: an approach for retrieving novel eukaryotic genes from polluted and related environments. *3 Biotech* **10**, 71 (2020).

335. Clark, K. F., Acorn, A. R. & Greenwood, S. J. A transcriptomic analysis of American lobster (Homarus americanus) immune response during infection with the bumper car parasite Anophryoides haemophila. *Dev. Comp. Immunol.* **40**, 112–122 (2013).

336. Schoenle, A., Nitsche, F., Werner, J. & Arndt, H. Deep-sea ciliates: Recorded diversity and experimental studies on pressure tolerance. *Deep Sea Research Part I: Oceanographic Research Papers* **128**, 55–66 (2017).

337. Dimond, K. L. & Zufall, R. A. Hidden genetic variation in the germline genome of Tetrahymena thermophila. *Journal of Evolutionary Biology* **29**, 1284–1292 (2016).

338. Long, H. & Zufall, R. A. Mutational Robustness of Morphological Traits in the Ciliate Tetrahymena thermophila. *Journal of Eukaryotic Microbiology* **62**, 249–254 (2015).

339. Kaur, H. *et al.* An endosomal syntaxin and the AP-3 complex are required for formation and maturation of candidate lysosome-related secretory organelles (mucocysts) in Tetrahymena thermophila. *Mol Biol Cell* **28**, 1551–1564 (2017).

340. Lee, L. J. Y., Klute, M. J., Herman, E. K., Read, B. & Dacks, J. B. Losses, Expansions, and Novel Subunit Discovery of Adaptor Protein Complexes in Haptophyte Algae. *Protist* **166**, 585–597 (2015).

341. Mowbrey, K. & Dacks, J. B. Evolution and diversity of the Golgi body. *FEBS Letters* **583**, 3738–3745 (2009).

342. Kurz, S. & Tiedtke, A. The Golgi Apparatus of Tetrahymena Thermophila. *Journal of Eukaryotic Microbiology* **40**, 10–13 (1993).

343. Hall, R. P. Vacuome and Golgi apparatus in the ciliate, stylonychia. *Z.Zellforsch* **13**, 770–782 (1931).

344. Barlow, L. D., Nývltová, E., Aguilar, M., Tachezy, J. & Dacks, J. B. A sophisticated, differentiated Golgi in the ancestor of eukaryotes. *BMC Biology* **16**, 27 (2018).

345. Wu, B. & Guo, W. The Exocyst at a Glance. *J Cell Sci* **128**, 2957–2964 (2015).

346. Plattner, H. The contractile vacuole complex of protists – New cues to function and biogenesis. *Critical Reviews in Microbiology* (2015).

347. Boehm, C. & Field, M. Evolution of late steps in exocytosis: conservation and specialization of the exocyst complex. *Wellcome Open Research* **4**, (2019).

348. Tagaya, M., Arasaki, K., Inoue, H. & Kimura, H. Moonlighting functions of the NRZ (mammalian Dsl1) complex. *Front. Cell Dev. Biol.* **2**, (2014).

349. Blum, J. J. The metabolic role of peroxisomes in Tetrahymena. *Ann. N. Y. Acad. Sci.* **386**, 217–227 (1982).

350. Gabaldón, T. Peroxisome diversity and evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences* **365**, 765–773 (2010).

351. Mohammadi, S., Saberidokht, B., Subramaniam, S. & Grama, A. Scope and limitations of yeast as a model organism for studying human tissue-specific pathways. *BMC Systems Biology* **9**, 96 (2015).

352. Mohan, K. V. K., Som, I. & Atreya, C. D. Identification of a Type 1 Peroxisomal Targeting Signal in a Viral Protein and Demonstration of Its Targeting to the Organelle. *Journal of Virology* **76**, 2543–2547 (2002).

353. Vukašinović, N. & Žárský, V. Tethering Complexes in the Arabidopsis Endomembrane System. *Front Cell Dev Biol* **4**, (2016).

354. Woo, Y. H. *et al.* Chromerid genomes reveal the evolutionary path from photosynthetic algae to obligate intracellular parasites. *eLife* **4**, (2015).

355. Hu, X., Lin, X. & Song, W. *Ciliate Atlas: Species Found in the South China Sea*. (Springer, 2019).

356. Lasek-Nesselquist, E. & Johnson, M. D. A Phylogenomic Approach to Clarifying the Relationship of Mesodinium within the Ciliophora: A Case Study in the Complexity of Mixed-Species Transcriptome Analyses. *Genome Biol Evol* **11**, 3218–3232 (2019).

357. Leonelli, S. What Difference Does Quantity Make? On the Epistemology of Big Data in Biology. *Big Data Soc* **1**, (2014).

358. O'Malley, M. A., Elliott, K. C. & Burian, R. M. From genetic to genomic regulation: iterativity in microRNA research. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* **41**, 407–417 (2010).

359. Klinger, C. M. & Richardson, E. Small Genomes and Big Data: Adaptation of Plastid Genomics to the High-Throughput Era. *Biomolecules* **9**, 299 (2019).

360. Kraus, W. L. Editorial: Would You Like A Hypothesis With Those Data? Omics and the Age of Discovery Science. *Mol Endocrinol* **29**, 1531–1534 (2015).

361. Kell, D. B. & Oliver, S. G. Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. *BioEssays* **26**, 99–105 (2004).

362. Ratti, E. Big Data Biology: Between Eliminative Inferences and Exploratory Experiments. *Philosophy of Science* **82**, 198–218 (2015).

363. Franklin, L. R. Exploratory Experiments. *Philosophy of Science* **72**, 888–899 (2005).

364. Glass, D. J. & Hall, N. A Brief History of the Hypothesis. *Cell* **134**, 378–381 (2008).

365. Kelling, S. *et al.* Data-intensive Science: A New Paradigm for Biodiversity Studies. *BioScience* **59**, 613–620 (2009).

366. Callebaut, W. Scientific perspectivism: A philosopher of science's response to the challenge of big data biology. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* **43**, 69–80 (2012).

367. Pietsch, W. The Causal Nature of Modeling with Big Data. *Philosophy & Technology* **29**, 137–171 (2016).

368. Raoult, D. Technology-driven research will dominate hypothesis-driven research: the future of microbiology. *Future Microbiology* **5**, 135–137 (2010).

369. Krohs, U. Convenience experimentation. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* **43**, 52–57 (2012).

370. Hey, T., Gannon, D. & Pinkelman, J. The Future of Data-Intensive Science. *Computer* **45**, 81–82 (2012).

371. Haufe, C., Elliott, K. C., Burian, R. M. & O'Malley, M. A. Machine science: what's missing. *Science* **330**, 317–318; author reply 318-320 (2010).

372. Forber, P. Reconceiving Eliminative Inference. *Philosophy of Science* **78**, 185–208 (2011).

373. Long, M. Suncor Energy Inc. 2018 Base Plant Fluid Tailings Management Report April 30, 2019. 156 (2018).

374. Bolyen, E. *et al.* Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology* **37**, 852–857 (2019).

375. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods* **13**, 581–583 (2016).

376. Dacks, J. B. & Field, M. C. Evolutionary origins and specialisation of membrane transport. *Current Opinion in Cell Biology* **53**, 70–76 (2018).

377. Reid, T. Evaluating the biogeochemistry and microbial function in the Athabasca Oil Sands region: Understanding natural baselines for reclamation end-points. *Electronic Theses and Dissertations* (2019).

378. Scott, A. C., Zubot, W., Davis, C. W. & Brogly, J. Bioaccumulation potential of naphthenic acids and other ionizable dissolved organics in oil sands process water (OSPW) – a review. *Science of The Total Environment* 134558 (2019) doi:10.1016/j.scitotenv.2019.134558.

379. White, K. B. & Liber, K. Chronic Toxicity of Surface Water from a Canadian Oil Sands End Pit Lake to the Freshwater Invertebrates Chironomus dilutus and Ceriodaphnia dubia. *Arch Environ Contam Toxicol* (2020) doi:10.1007/s00244-020-00720-3.

380. Carini, P. *et al.* Effects of Spatial Variability and Relic DNA Removal on the Detection of Temporal Dynamics in Soil Microbial Communities. *mBio* **11**, (2020).

381. Johnson, E. A. & Miyanishi, K. Creating new landscapes and ecosystems: the Alberta Oil Sands. *Annals of the New York Academy of Sciences* **1134**, 120–45 (2008).

382. Zwirglmaier, K. Fluorescence in situ hybridisation (FISH) – the next generation. *FEMS Microbiol Lett* **246**, 151–158 (2005).

383. Arroyo, A. S., López-Escardó, D., Kim, E., Ruiz-Trillo, I. & Najle, S. R. Novel Diversity of Deeply Branching Holomycota and Unicellular Holozoans Revealed by Metabarcoding in Middle Paraná River, Argentina. *Frontiers in Ecology and Evolution* **6**, 99–99 (2018).

384. Gilbert, J. A., Jansson, J. K. & Knight, R. The Earth Microbiome project: successes and aspirations. *BMC Biology* **12**, 69–69 (2014).

385. Dumack, K., Fiore-Donno, A. M., Bass, D. & Bonkowski, M. Making sense of environmental sequencing data: ecologically important functional traits of the protistan groups Cercozoa and Endomyxa (Rhizaria). *Molecular Ecology Resources* **n/a**,.

386. Tsaousis, A. D., Leger, M. M., Stairs, C. A. W. & Roger, A. J. The Biochemical Adaptations of Mitochondrion-Related Organelles of Parasitic and Free-Living Microbial Eukaryotes to Low Oxygen Environments. in *Anoxia: Evidence for Eukaryote Survival and Paleontological Strategies* (eds. Altenbach, A. V., Bernhard, J. M. & Seckbach, J.) 51–81 (Springer Netherlands, 2012). doi:10.1007/978-94-007-1896-8_4.

387. Karst, S. M. *et al.* Enabling high-accuracy long-read amplicon sequences using unique

molecular identifiers with Nanopore or PacBio sequencing. *bioRxiv* 645903 (2020) doi:10.1101/645903.

388. Obiol, A. *et al.* A metagenomic assessment of microbial eukaryotic diversity in the global ocean. *Molecular Ecology Resources* **n/a**,.

389. Saary, P., Mitchell, A. L. & Finn, R. D. Estimating the quality of eukaryotic genomes recovered from metagenomic analysis. *bioRxiv* 2019.12.19.882753 (2020) doi:10.1101/2019.12.19.882753.

390. Jamy, M. *et al.* Long-read metabarcoding of the eukaryotic rDNA operon to phylogenetically and taxonomically resolve environmental diversity. *Molecular Ecology Resources* **n/a**,.

391. Richards, T. A., Massana, R., Pagliara, S. & Hall, N. Single cell ecology. *Philosophical Transactions of the Royal Society B: Biological Sciences* **374**, 20190076 (2019).

392. Godfrey-Smith, P. *Theory and Reality*.

393. Christer Brönmark & Hansson, L.-A. *The Biology of Lakes and Ponds*. (Oxford University Press, 2017).

394. Lloyd, G. *The Man of Reason: 'Male' and 'Female' in Western Philosophy*. (University of Minnesota Press, 1993).

395. de Troyes, C. *Four Arthurian Romances*. (Project Gutenberg).

396. von Eschenbach, W. *Parzival*. (Project Gutenberg).

397. Zachos, F. E. A Brief History of Species Concepts and the Species Problem. in *Species Concepts in Biology: Historical Development, Theoretical Foundations and Practical Relevance* (ed. Zachos, F. E.) 17–44 (Springer International Publishing, 2016). doi:10.1007/978-3-319-44966-1_2.

398. Hegarty, M. J. & Hiscock, S. J. Hybrid speciation in plants: new insights from molecular studies. *New Phytologist* **165**, 411–423 (2005).

399. Boto, L. Horizontal gene transfer in evolution: facts and challenges. *Proceedings of the Royal Society B: Biological Sciences* **277**, 819–827 (2010).

400. Xu, J. The prevalence and evolution of sex in microorganisms. *Genome* **47**, 775–780 (2004).

401. Callahan, B. J., McMurdie, P. J. & Holmes, S. P. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal* **11**, 2639–2643 (2017).

402. Báldi, A. Using higher taxa as surrogates of species richness: a study based on 3700 Coleoptera, Diptera, and Acari species in Central-Hungarian reserves. *Basic and Applied Ecology* **4**, 589–593 (2003).

403. Puente, A. & Juanes, J. A. Testing taxonomic resolution, data transformation and selection of species for monitoring macroalgae communities. *Estuarine, Coastal and Shelf Science* **78**, 327–340 (2008).

404. Quijón, P. A. & Snelgrove, P. V. R. The use of coarser taxonomic resolution in studies of predation on marine sedimentary fauna. *Journal of Experimental Marine Biology and Ecology* **330**, 159–168 (2006).

405. Bett, B. J. & Narayanaswamy, B. E. Genera as proxies for species α- and β-diversity: tested across a deep-water Atlantic–Arctic boundary. *Marine Ecology* **35**, 436–444 (2014).

406. Warwick, R. M. The level of taxonomic discrimination required to detect pollution effects on marine benthic communities. *Marine Pollution Bulletin* **19**, 259–268 (1988).

407. Bertasi, F. *et al.* Comparing efficacy of different taxonomic resolutions and surrogates in detecting changes in soft bottom assemblages due

to coastal defence structures. *Marine Pollution Bulletin* **58**, 686–694 (2009).

408. Gomez Gesteira, J. L., Dauvin, J. C. & Salvande Fraga, M. Taxonomic level for assessing oil spill effects on soft-bottom sublittoral benthic communities. *Marine Pollution Bulletin* **46**, 562–572 (2003).

409. Bevilacqua, S., Fraschetti, S., Musco, L. & Terlizzi, A. Taxonomic sufficiency in the detection of natural and human-induced changes in marine assemblages: A comparison of habitats and taxonomic groups. *Marine Pollution Bulletin* **58**, 1850–1859 (2009).

Appendix i

SPECIES CONCEPTS AND THEIR APPLICABILITY TO CLUSTERING THRESHOLDS.

i.1 Species concepts

The concept of a 'species' is ancient and is the fundamental unit of the science of classification and taxonomy[1]. However, what actually constitutes a species is still, after thousands of years, an open question. A general working definition of species, and the one generally taught in schools, is that of reproductive isolation. A species is a group of organisms which cannot mate with other species and produce viable, fertile offspring[397]. This definition is useful for organisms in the Animalia, and mechanisms for understanding speciation have been well established in ecology and agriculture. However, this species definition does not hold up consistently even across multicellular organisms. Plants are notorious for their polyploidal genomes, which allow the creation of a new species in a single generation; though reproductive isolation still exists in this phylum, it is not the only method of speciation, and hybrid plants are often completely capable of producing viable offspring[398].

At the microbial level, the species concept becomes even more complex. Unlike multicellular organisms, where only a small portion of cells, the germline, will pass on their genetic information, single-celled organisms do not have this additional barrier to genetic diversity. Horizontal gene transfer, for example, is a mechanism of genetic reshuffling that is responsible for much of the functional diversity in various protist lineages[399]. It is in the microbial world that the concept of reproductive isolation completely breaks down; indeed, most microbial lineages do not engage in 'reproduction' as it is traditionally understood[400].

It is therefore clear that microbial species concepts require additional thought. Early studies of eukaryotic microbes used morphology and ultrastructure as the basis for taxonomy, and the morphological species concept is still incredibly valuable in understanding protist diversity[3,28]. However, DNA-based species concepts are increasingly used to define the scope of the tree of life, particularly in the microbial world[13]. The value of, and tension between, morphological and DNA species concepts is discussed in detail in Chapter 1, Section 1 of this thesis. This section also touches upon the difficulty of defining species concepts, particularly within a group of organisms

with as much genetic diversity and historical conflict between disciplines as eukaryotes. In this appendix, I do not attempt to suggest a definitive species concept or debate the merits of using DNA-based classification over morphological classification but explain why I have principally used a 97% clustering threshold to define the operational taxonomic units (OTUs) used in the downstream analyses throughout the research I have performed that uses eDNA methodology.

i.2 OTU to species: clustering thresholds across eukaryotic diversity

The 'clustering threshold' is a term that describes the level of similarity of sequences used to split OTUs in clustering analyses. When amplicons are created, the signal from a single DNA region from an eDNA sample is amplified using primers and PCR[35]. In eukaryotic community analyses, this is usually one of the nine variable regions from the 18S rRNA subunit. Clustering thresholds are based on the percentage similarity of two sequences to each other; for example, a 97% clustering threshold means that two amplicons that differ in three bases at every hundred bases or more will be split into separate OTUs. Most of the hypervariable regions of the 18S rRNA subunit are 200-500bp long, so a 97% clustering threshold corresponds (in most cases) to less than ten base pair discrepancies across the whole amplicon.

The origin of a 97% clustering threshold as a proxy for species-level diversity is unclear. However, it definitely originated from eDNA studies on bacteria, and it is still often used in bacteria today[70]. Bacterial horizontal gene transfer and rapid genome evolution has already resulted in a fuzzier species concept in bacteria as opposed to that of eukaryotes, but early studies of eDNA in eukaryotes have also used this clustering threshold. As species concepts (particularly in multicellular eukaryotes) were more strongly delineated, various studies examined the relationship between a 97% OTU clustering threshold and previously defined species boundaries. It quickly became apparent that it did not hold true across all eukaryotes. Boengik et al (2012)[71] noted that actually no single clustering threshold is likely to be relevant across the diversity of eukaryotes, particularly in single-celled lineages. Different eukaryotic clades have different rates of evolution, and some phyla demonstrate variability even in the extremely conserved 18S rRNA subunit; for example, it has been demonstrated that the pan-eukaryotic Stoeck et al (2010)[60] primers used to amplify the V4 region do a relatively poor job on Amoebozoa because their V4 region is substantially longer than the expected 400bp[35].

There are differing schools of thought regarding how this variable species-level diversity issue can be addressed. It is generally agreed that a systematic approach to analysing community diversity is important to ensure reproducibility between studies and to maximise the useful information that can be distributed publicly; biodiversity curation efforts such as the Earth Microbiome Project have extensive regulations for data collection and curation[384], and efforts within the protist community to standardise eDNA protocols are also underway[35]. However, there is still disagreement over the optimal clustering threshold in microbial community studies in both bacteria and protists. A recent study by Edgar et al (2018)[70] determined an optimal clustering threshold for species-level diversity in bacteria was 99.5%, much higher than the commonly used 97%. Results in eukaryotes are more mixed; some researchers suggest 99%, while others have discarded OTU clustering entirely and instead count every sequence as an individual species after data cleanup. These sequences, known as autosomal sequence variants (ASVs), were introduced in 2016 and are gaining in popularity[375,401].

Two chapters of my thesis, Chapter 2 and Chapter 3, concern eDNA studies with OTU clustering. Because OTU-clustered datasets are not comparable between analyses, these two studies, despite containing a subset of the same data, were clustered separately: the 2015 samples as a single entity, and then the entire dataset spanning 2015-2018. In both cases, the majority of the analyses were carried out at 97% identity. I initially chose this clustering threshold for the 2015 data from Chapter 2 because it was consistent with other studies at the time of analysis but chose to use the same clustering threshold for the time series analysis in Chapter 3. I chose a clustering threshold of 97% for these analyses for the following reasons:

- Consistency between the 2015 data and the 2015-2018 data.

- More scope to detect novel diversity: broader clustering threshold means that any detected novel diversity is more likely to be genuine biological diversity and not a result of splitting known species over multiple OTUs.

- Consistency with phylogenetic classification: due to the size of the dataset, which contains over 300 diverse samples from three different lakes, clustering at a 99% threshold or ASV threshold produces too many OTUs to be analysed phylogenetically within a reasonable

time frame. Since the 2015 analysis showed that the phylogenetic analysis was extremely informative for detecting novel diversity and classifying species, I wanted to be able to include this in the time series analysis.

In Chapter 2, I compared ordination of ASVs across the months of 2015 to the ordination of OTUs at 97% and found similar results (Appendix 2, Figure S2.6). In Chapter 3, I compared the ordination of the dataset 99% and 97% clustering thresholds, and also found similar results (Appendix 3, Figure S3.8).

In this appendix, I compare BLAST-based classifications and ordinations for a sample dataset (taken from July 2015) to determine how the dataset is affected by differing clustering thresholds or methods for generating taxonomic units.

i.3 OTU production at 97%, 99%, and ASV-level diversity

I ran an identical dataset (FASTQ files from July 2015) through a QIIME pipeline[54] (Figure S1.1) at 97% and 99% diversity, then through the dada2 pipeline[375] to produce ASVs. A summary of the statistics for these runs is found in Table S1.1 (ASV pipeline statistics), Table S1.2 (OTU pipeline statistics), and Table S1.3 (OTU cluster statistics). The two OTU pipelines had very similar results; the OTU read statistics were identical, but the 97% clustering threshold had a larger average cluster size (124 versus 24.9) and a smaller number of clusters, or OTUs (2545 versus 450). This is consistent with the methodology used as the two pipelines are identical during the cleanup, dereplication, and chimera removal stages.

The ASV pipeline is not directly comparable to the OTU pipelines because the samples are analysed separately and reads are not joined until the final stage of the analysis process. However, on average, 15% of the ASV dataset was found to be chimeric versus 5% of the OTU dataset. The average ASV length was also slightly shorter than the OTUs (388 versus 407), but with a smaller

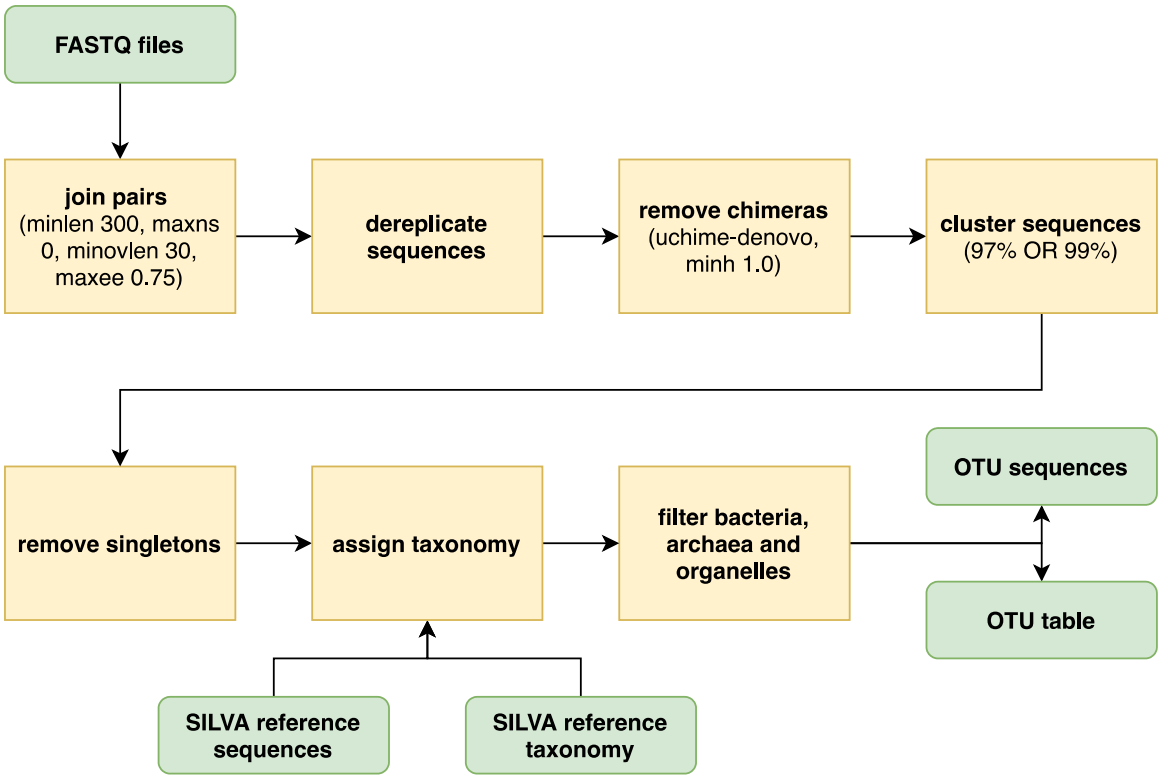


Figure S1.1: FASTQ processing pipeline implemented in QIIME2.
Green boxes are information inputs and outputs from QIIME2, yellow boxes are processing steps. Parameters for each processing step are outlined in round brackets.

Table S1.1: Read loss along ASV processing pipeline.

	input	filtered	denoisedF	denoisedR	merged	nonchim	classified
C10A	54353	4468	4329	4354	2287	1882	1013
C10B	136063	13524	13341	13370	4513	3179	3162
C10C	139733	10531	10334	10497	6032	5035	4971
C10D	69098	17648	17513	17614	11554	10165	8726
C6F	208640	87905	87562	87618	33204	27702	11962
C6G	82027	46216	46128	46143	17128	14617	14599
C6H	70104	43830	43561	43565	1736	1310	843
C7A	90343	63955	63452	63601	21086	17650	15881
C7B	200547	79362	78865	78927	22889	17075	16386
C7C	160014	64697	64573	64618	26960	23400	22653
C7D	209547	46155	46064	46117	12843	10743	10391
C7E	475284	73454	73167	73223	23584	20120	19935
C7F	287140	103082	102806	102974	38906	36170	35674
C7G	137608	67923	67512	67741	8801	8257	7972
C7H	172884	30960	30755	30901	6834	6085	5697
C8A	61882	37072	36960	36964	21829	17780	14034
C8B	64276	46844	46697	46730	28951	25886	18200
C8C	93966	74160	74061	74110	50018	38613	4704
C8D	56926	33174	33105	33144	24306	20934	20855
C8E	134728	75121	75042	75085	57913	53029	33180
C8F	165975	41875	41671	41796	13733	10852	10631
C8G	139614	51712	51419	51672	40342	32073	22284
C8H	69856	48397	48201	48337	35526	33307	15894
C9A	18887	10766	10607	10677	7889	6872	6852
C9B	26459	19552	19349	19495	16479	14030	13718
C9D	28659	12738	12618	12721	8756	6769	4978
C9E	29717	1812	1712	1779	1351	1332	1332
C9F	29885	22018	21921	21989	2438	2175	2175
C9G	45941	17938	17906	17913	11399	8737	8698
C9H	60236	21396	21239	21360	9106	7809	3397
average	117346.40	42276.17	42082.33	42167.83	18946.43	16119.60	12026.57

Table S1.2: Read loss along OTU clustering pipelines.

	merge pairs	uniques	non-chimeras	OTUs	OTUs without singletons	Classified OTUs
otu 99	980269	139931	966015	5321	2545	2445
otu 97	980269	139931	966015	1070	450	357

Table S1.3: Average cluster statistics for 97% versus 99% clustered OTUs.

	OTU length			Cluster size		
	min	**max**	**average**	**min**	**max**	**average**
otu 99	301	564	407	1	6618	24.9
otu 97	301	564	407	1	21176	124

range (148 versus 263). The sequences produced by ASVs appear to be more consistent than those of OTUs.

i.4 OTU classification at 97%, 99% and ASV-level diversity

After producing the taxonomic units (OTUs and ASVs), I classified them using a BLAST search[251] against the PR2 database[19], taking only the top hit. Though this classification protocol was less comprehensive than the one I used in Chapter 2 and Chapter 3 to classify the OTUs clustered at 97%, this allowed me to compare a consistent classification between all three clustering methods. The results of these classifications, which take into account the abundance of each OTU after clustering, are shown in Figure S1.2.

The first result from classification is the amount of OTUs that were not classified by a BLAST search against the PR2 database. In the case of the OTUs, there was a substantial difference in the percentage of OTUs that were unclassified in the two datasets (20.6% at 97% clustering versus 4.1% at 99% clustering). In fact, the absolute number of unclassified OTUs is very similar in both samples (93 at 97% clustering, 100 at 99% clustering). If the reason the BLAST search was unable to produce a hit is because the sequences are too dissimilar to known 18S diversity (which may be a result of true biological divergence or sequencing error), then it makes sense that the numbers of OTUs that are unclassified are similar in both, since sequences more than 3% different from the rest of the dataset would result in the same OTU number in 97% and 99% clustering. However, when the abundance of the unclassified OTUs is taken into account, both the number of sequences and the percentage of the overall abundance is very similar (1802 unclassified sequences at 97% clustering, 1789 at 99% clustering, both representing 0.2% of the overall abundance). In the case of the ASVs, the average percentage of unclassified sequences is only moderately higher than the highest OTU dataset (25.4% unclassified), but these ASVs account for 25% of the overall abundance of the dataset. ASVs appear to have a much higher rate of sequences not classified by BLAST.

The classified sequences are where the differences between the OTU and ASV datasets are most prominent. The ASV sequences appear to be extremely homogenous, with only two fungi and one ciliate making up the vast majority of the diversity. This is unlikely to be an accurate representation

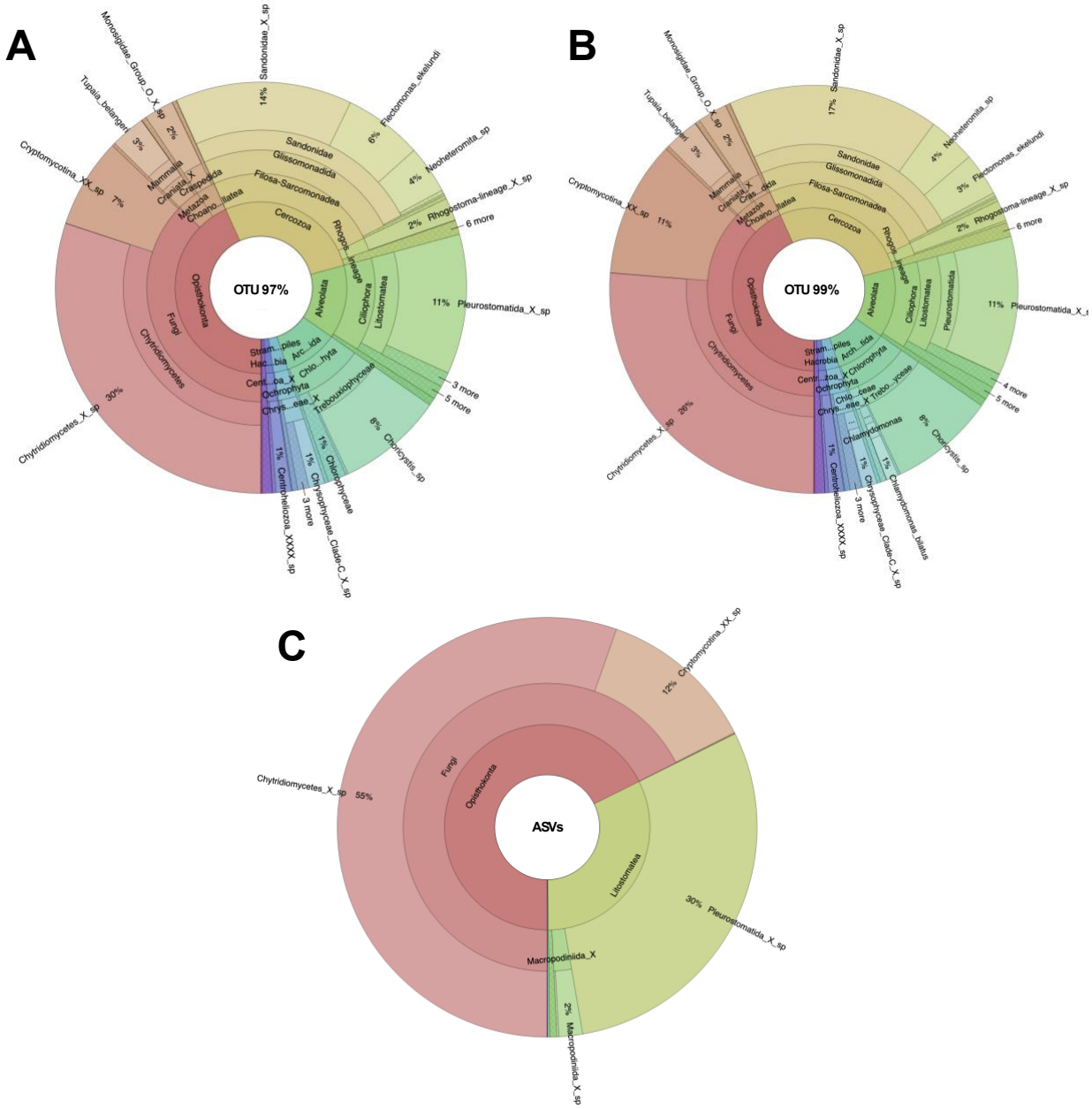


Figure S1.2: KronaPlots of OTUs clustered at 97% and 99% diversity, and ASVs.
A: OTUs clustered at 97% diversity. B: OTUs clustered at 99% diversity. C: ASVs. Interactive versions of these KronaPlots are available in Online Supplementary Information, Appendix I.

of the community based on the additional analyses done using this dataset in Chapter 2. It is unclear exactly what caused this overrepresentation of these two clades and loss of all other diversity.

On the contrary, the two OTU datasets show an extremely similar distribution of classifications. In each KronaPlot, the same phyla and classes make up the same proportion of the abundance. This suggests that though more OTUs are found in the OTU99 dataset, the classification methodology is robust enough that the same trends can be detected in both samples even though one is more approximate to species-level diversity and one is more approximate to genus-level diversity.

i.5 Ecological relevance of genus-level classifications

A 97% clustering threshold is, in most eukaryote clades, thought to represent genus-level diversity. One of the aims of bioprospecting in Base Mine Lake is detection of bioindicator species, which can then be used as proxies to assess the effectiveness of reclamation. Genus-level classification, as opposed to individual species classification, is well established in known bioindicator species, and is often used due to the difficulty of distinguishing between many of the smaller indicator species (usually insects, crustaceans, or mesofauna)[402–405]. This is used often in the context of reclamation ecology[406–409]. If any of the OTUs clustered at 97% appear to be useful as bioindicators, it would be reasonable to assume that the genus as a whole would be useful either as a bioindicator or as a source of potential bioindicator species.

i.6 Conclusions

Though a 97% clustering threshold is too high to be a reasonable approximation of species-level diversity, I have chosen to use this threshold in Chapter 2 and Chapter 3 for consistency with my published work on Base Mine Lake because using a more stringent clustering threshold would preclude the use of techniques such as phylogenetics for taxonomic analysis, and because a higher clustering threshold makes novel diversity easier to detect. I have tested both 97% and 99% clustering thresholds for a test dataset and found that the overall abundance, distribution, and classification of the OTUs at the two thresholds is very similar. I also tested ASVs but found that this resulted in a much higher abundance of unclassified sequences after BLAST classification and what appears to be either misclassification or data loss within this pipeline when used on the Base

Mine Lake dataset. Using genus-level rather than species-level diversity also does not necessarily affect the ecological inferences made in Chapter 2 and Chapter 3; any OTUs identified as potential bioindicators at the 97% threshold are likely to be useful in a reclamation context.

Appendix ii
SUPPLEMENTARY DATA FOR CHAPTER 2

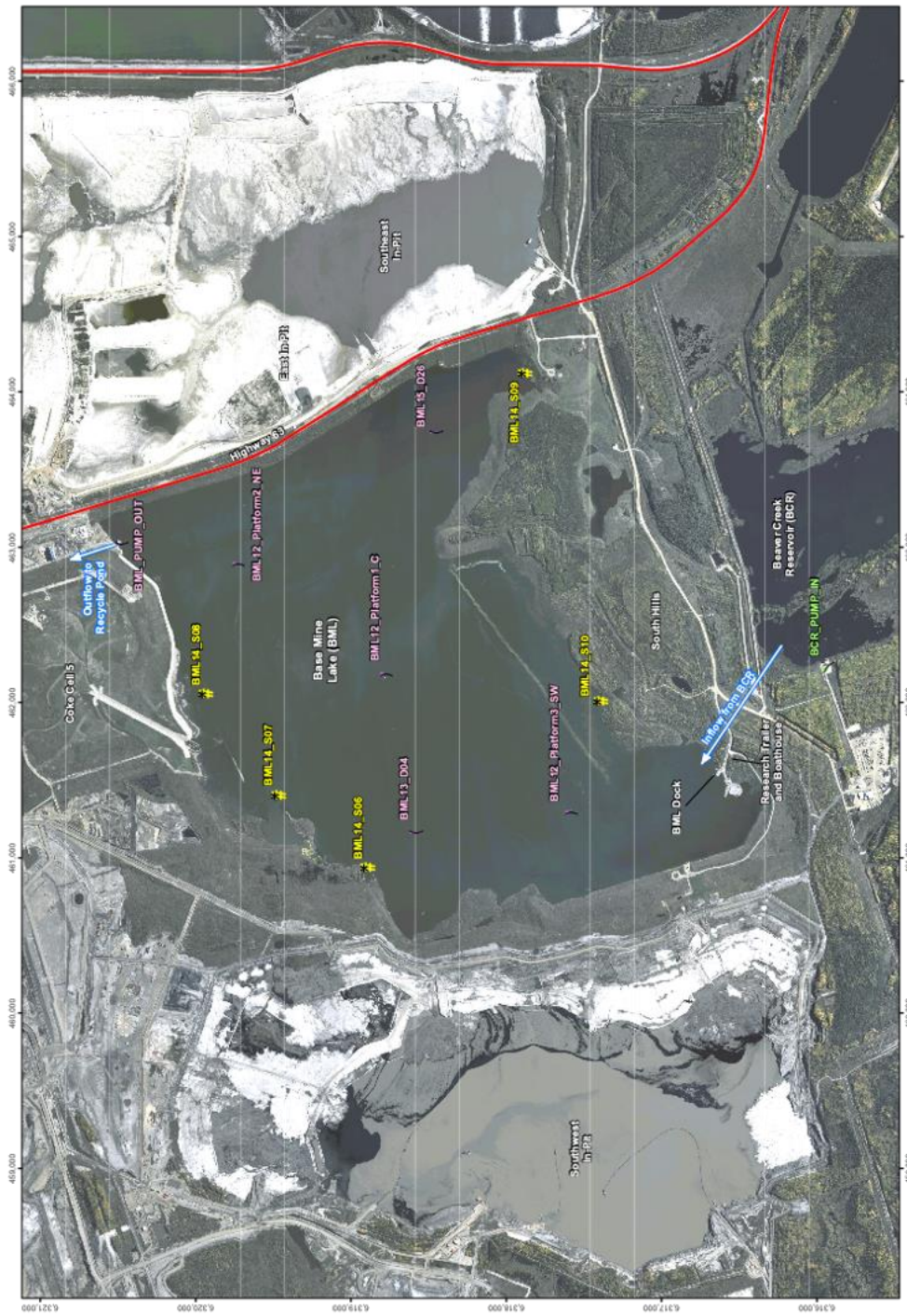


Figure S2.1: Map of Base Mine Lake and its associated sampling sites. This map was adapted from a figure from an internal report created by Hatfield Consultants with data from Syncude Ltd., as part of the company's environmental monitoring programme.

Supplementary Figure 2

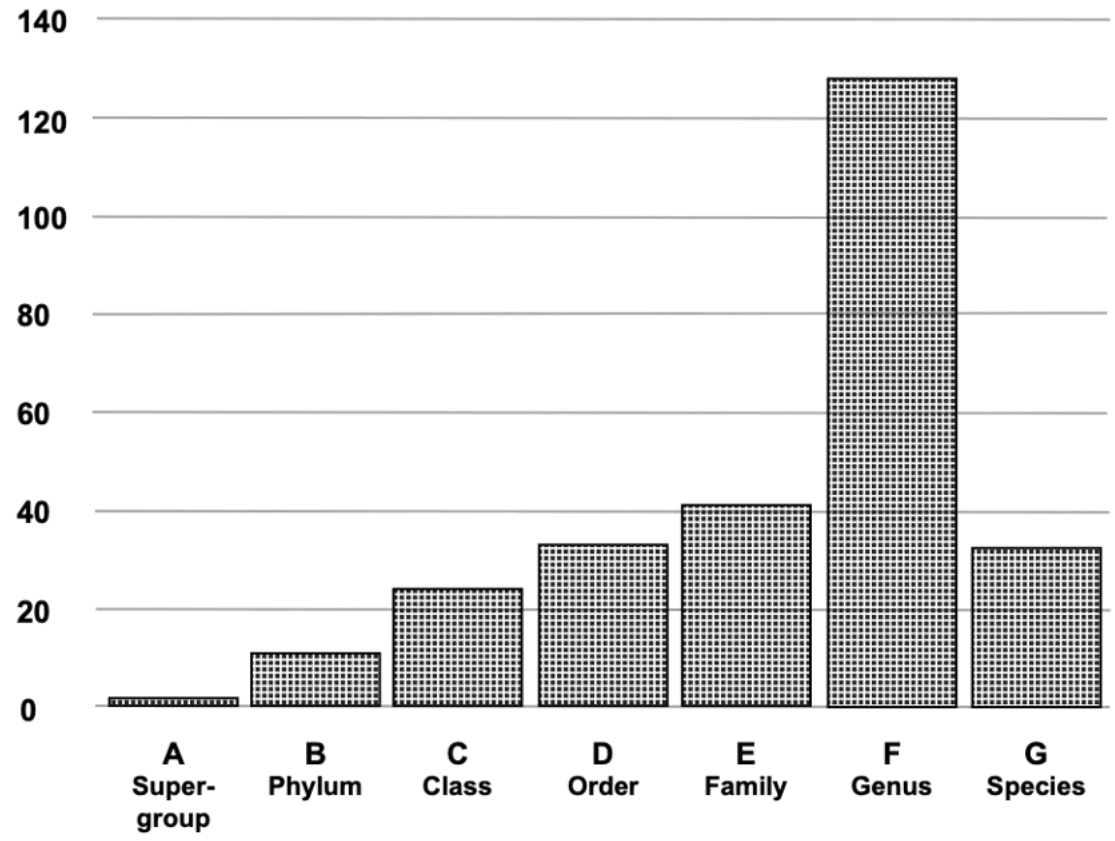


Figure S2.2: Distribution of classification levels in the classified OTUs from Base Mine Lake.
The Y axis shows number of OTUs classified to the indicated level, and the x axis shows the classification level (based on PR2 classification level designations).

Supplementary Figure 3

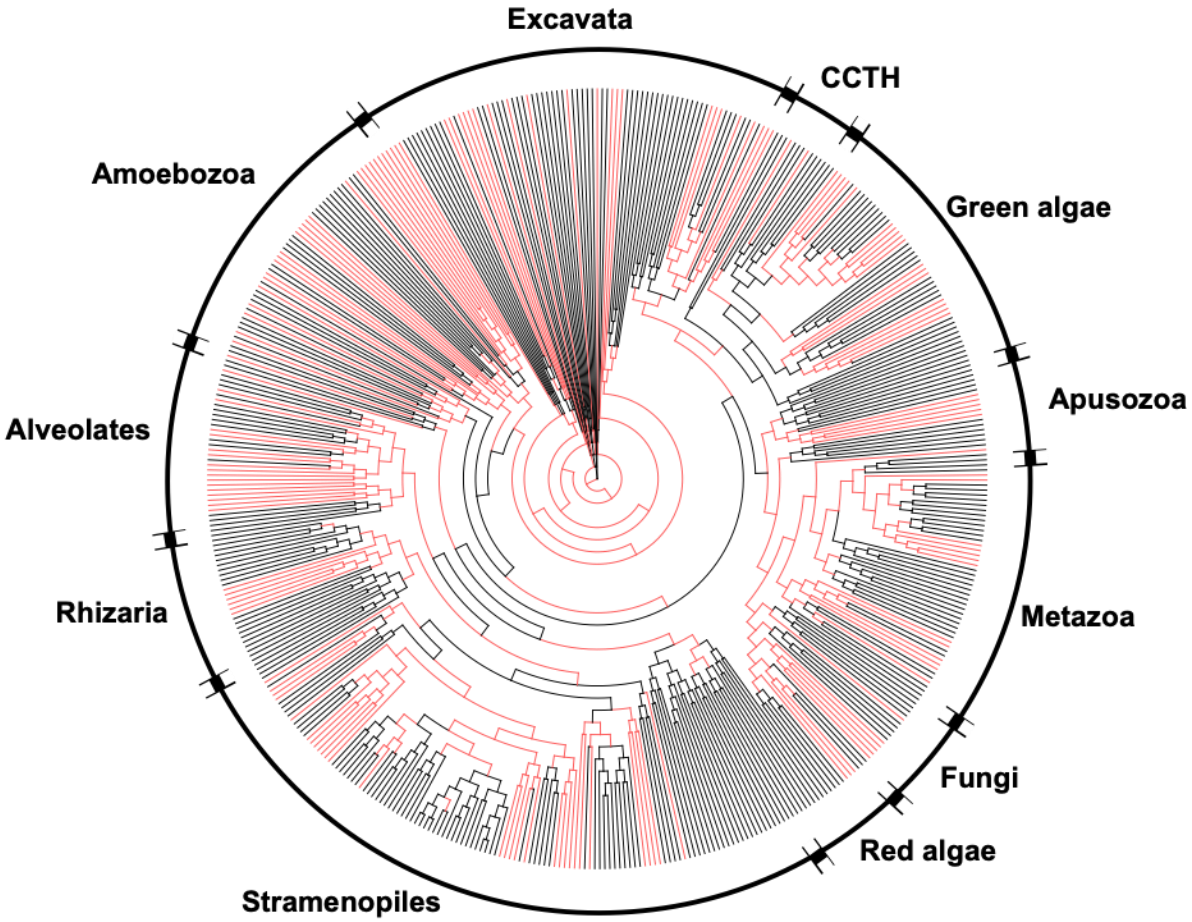


Figure S2.3: A pplacer plot of eukaryotic diversity in the Base Mine Lake samples.
Based on the phylogenetic position of each OTU, based on the pan-eukaryotic backbone tree used in Aguilar *et al* (2916). Red tree branches in this unrooted pan-eukaryotic tree indicate presence of at least one OTU grouping most strongly with that taxon.

Supplementary Figure 4

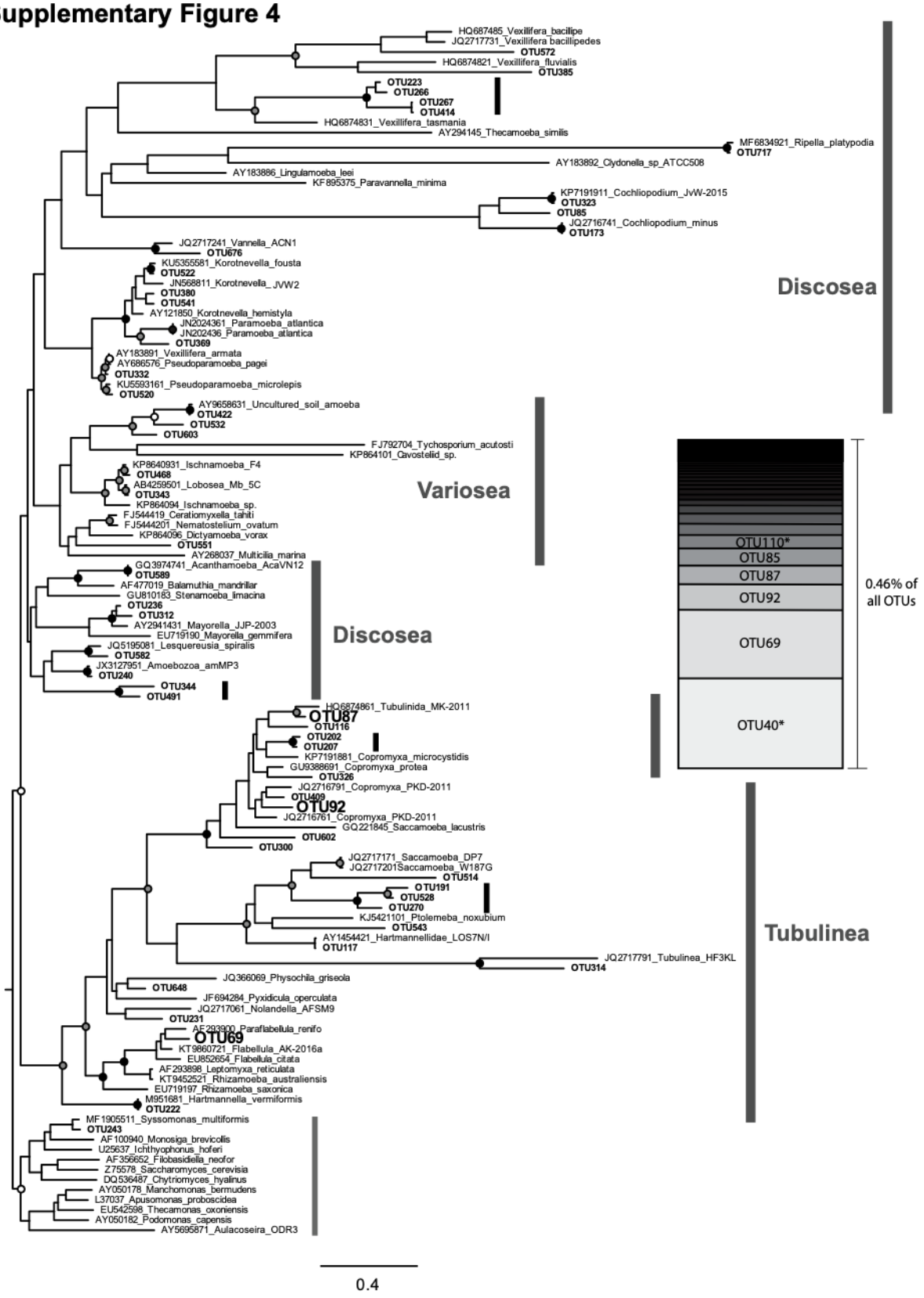


Figure S2.4: Phylogeny of Amoebozoa.
Including reference sequences, OTUs and the closest BLAST hit in GenBank to each reference sequence. Phylogeny is based on MrBayes and RAxML trees, mapped onto the MrBayes topology. Node support is indicated by the circles on each node: black indicates 1 / 100 MrBayes / RAxML support, grey indicates 0.9 / 90% or higher, and white indicates 0.75 / 75 support. High-level taxonomic groupings are indicated with grey bars. Particularly abundant OTUs are indicated in a larger font on the trees, and in the bar graph to the right of the phylogeny as a proportion of total OTUs. This tree also shows the phylogenetic positioning of *Syssomonas* (OTU243) in the Metazoan outgroup to Amoebozoa.

Supplementary Figure 5

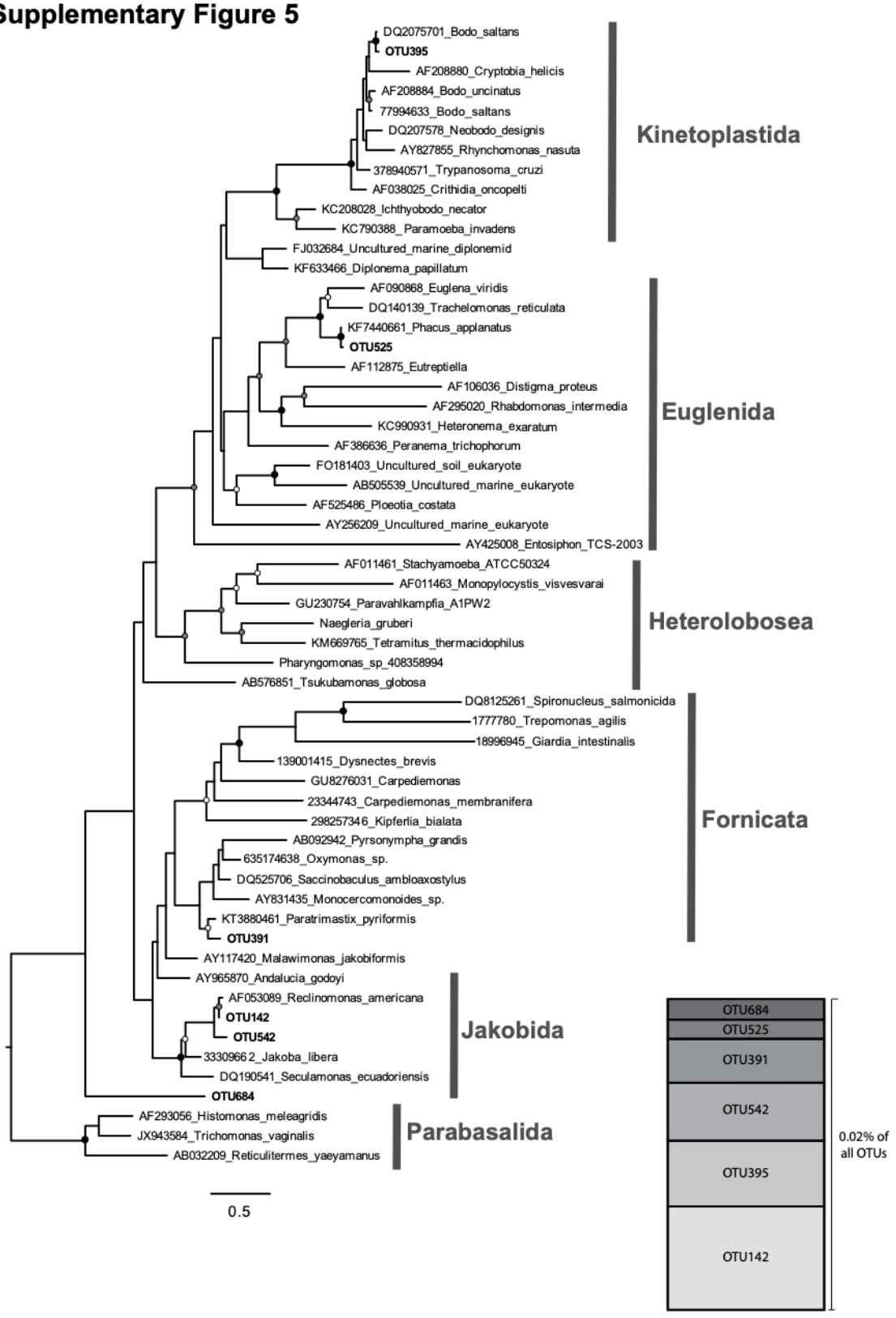


Figure S2.5: Phylogeny of Excavata.
Phylogeny of Excavata, using parameters described in Figure S2.4. The only Euglena sequence identified is OTU525, grouped in the Euglenida. There is also an extremely deep-branching OTU that is definitively grouped within the Excavata, but not with any given excavate clade (OTU684).

Supplementary Figure 6

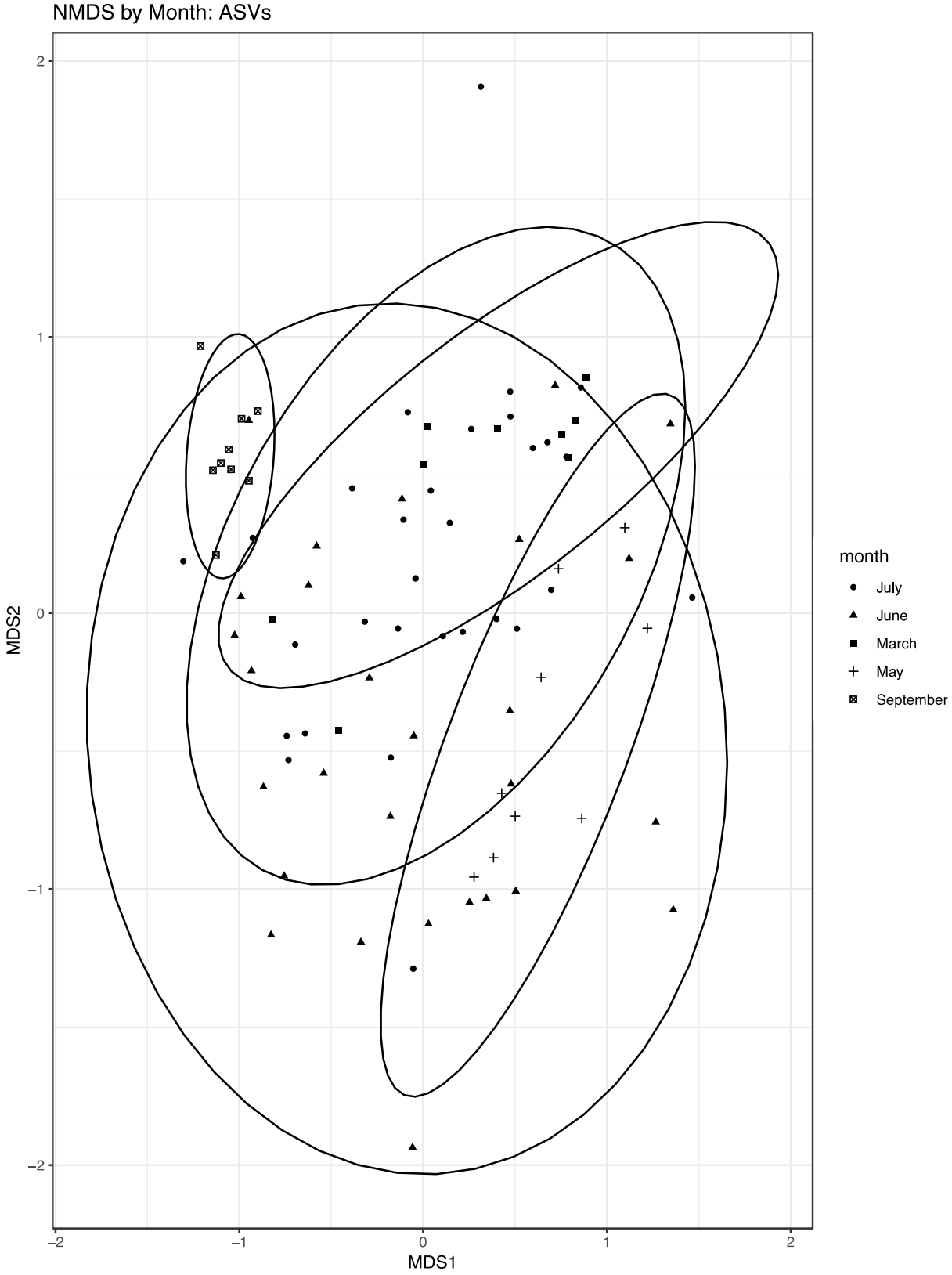


Figure S2.6: Ordination of ASVs from Base Mine Lake 2015.
NMDS analysis of Base Mine Lake ASV abundances using Bray-Curtis distance and default metaMDS analysis in R(vegan) of the total ASV abundance table by sample for all samples of each species taken from Base Mine Lake. Month is indicated by the shape of the data point for each sample, and ellipses show the clustering of the points for each month.

Table S2.1: Metadata associated with each sample used to generate the amplicon data used in this study.

	Platform	Depth	Month		Platform	Depth	Month
SMPL0	1	0m	June	SMPL47	1	6m	July
SMPL1	1	1m	June	SMPL48	1	7m	July
SMPL2	1	2m	June	SMPL49	1	8m	July
SMPL3	1	3m	June	SMPL50	1	9m	July
SMPL4	1	4m	June	SMPL51	2	0m	July
SMPL5	1	5m	June	SMPL52	2	1m	July
SMPL6	1	6m	June	SMPL53	2	2m	July
SMPL7	1	7m	June	SMPL54	2	3m	July
SMPL8	1	8m	June	SMPL55	2	4m	July
SMPL9	2	0m	June	SMPL56	2	5m	July
SMPL10	2	1m	June	SMPL57	2	6m	July
SMPL11	2	2m	June	SMPL58	2	7m	July
SMPL12	2	3m	June	SMPL59	2	8m	July
SMPL13	2	4m	June	SMPL60	2	9m	July
SMPL14	2	5m	June	SMPL61	2	10m	July
SMPL15	2	6m	June	SMPL62	3	0m	July
SMPL16	2	7m	June	SMPL63	3	1m	July
SMPL17	2	8m	June	SMPL64	3	2m	July
SMPL18	2	10m	June	SMPL65	3	3m	July
SMPL19	3	0m	June	SMPL66	3	4m	July
SMPL20	3	1m	June	SMPL67	3	5m	July
SMPL21	3	2m	June	SMPL68	3	6m	July
SMPL22	3	3m	June	SMPL69	3	7m	July
SMPL23	3	4m	June	SMPL70	3	8m	July
SMPL24	3	5m	June	SMPL74	1	0m	May
SMPL25	3	6m	June	SMPL75	1	4m	May
SMPL26	3	7m	June	SMPL76	1	8m	May
SMPL27	3	8m	June	SMPL77	2	0m	May
SMPL32	1	0m	March	SMPL78	2	4m	May
SMPL33	1	4m	March	SMPL79	2	8m	May
SMPL34	1	8m	March	SMPL80	3	0m	May
SMPL35	2	0m	March	SMPL81	3	4m	May
SMPL36	2	4m	March	SMPL82	3	8m	May
SMPL37	2	8m	March	SMPL83	1	0m	September
SMPL38	3	0m	March	SMPL84	1	4m	September
SMPL39	3	4m	March	SMPL85	1	8m	September
SMPL40	3	8m	March	SMPL86	2	0m	September
SMPL41	1	0m	July	SMPL87	2	4m	September
SMPL42	1	1m	July	SMPL88	2	8m	September
SMPL43	1	2m	July	SMPL89	3	0m	September
SMPL44	1	3m	July	SMPL90	3	4m	September
SMPL45	1	4m	July	SMPL91	3	8m	September

Table S2.2: Quality-controlled OTUs with their cross-referenced classifications, and their distributions based on OTU counts within each sample in the total dataset.

This table can be found as an .xlsx file in Online Supplementary Data, Appendix II.

Appendix iii
SUPPLEMENTARY DATA FOR CHAPTER 3

Figure S3.1: Pan-eukaryotic phylogeny of OTUs clustered at 97%.
The tree is assembled using the RAxML BlackBox algorithm, with the entire clustered OTU dataset and a pan-eukaryotic backbone adapted from Aguilar et al. (2016). The tree is found as a .nwk tree file in Online Supplementary Data, Appendix III.

Figure S3.2: Full-length phylogeny of OTUs preliminarily classified as Cercozoa.
The tree is assembled using the RAxML BlackBox algorithm, with the OTUs that were classed as Cercozoa from the pan-eukaryotic phylogeny in Figure S3.1. The tree is found as a .nwk tree file in Online Supplementary Data, Appendix III.

Figure S3.3: Full-length phylogeny of OTUs preliminarily classified as Ciliophora.
The tree is assembled using the RAxML BlackBox algorithm, with the OTUs that were classed as Ciliophora from the pan-eukaryotic phylogeny in Figure S3.1. The tree is found as a .nwk tree file in Online Supplementary Data, Appendix III.

Figure S3.4: Full-length phylogeny of OTUs preliminarily classified as Fungi.
The tree is assembled using the RAxML BlackBox algorithm, with the OTUs that were classed as Fungi from the pan-eukaryotic phylogeny in Figure S3.4. The tree is found as a .nwk tree file in Online Supplementary Data, Appendix III.

Figure S3.5: Full-length pplacer tree of core and persistent microbiome taxonomy.
The tree is assembled using pplacer, using the pan-eukaryotic backbone from Aguilar et al. (2016) and the OTUs identified as core and persistent microbiome by Temporal Insights in Microbial Ecology. The tree is found as a .nwk tree file in Online Supplementary Data, Appendix III.

Table S3.1: Metadata associated with each sample used to generate the amplicon data used in this study.
This table can be found as a .xslx file in Online Supplementary Data, Appendix III.

Table S3.2: Quality-controlled OTUs with their cross-references classifications, and their distributions based on OTU counts within each sample in the total dataset.
This table can be found as an .xlsx file in Online Supplementary Data, Appendix III.

Table S3.3: Classification outcome of OTUs

Classification type	# of OTUs	% of OTUs	Additional classification	# of OTUs	% of OTUs
Total	7900	100	(searched into GenBank using BLAST)	1610	20.4
UNKNOWN	766	9.7			
None	131	1.6			
Verified classification	6864	86.9	(verified after BLAST)	1340	19.5
Bacteria	7	0.1			
Archaea	10	0.1			
Chimeras	122	1.5			

Table S3.4: OTU sequences with statistically significant relationships with time, as indicated by MICtools analysis.
This table can be found as an .xlsx file in Online Supplementary Data, Appendix III.

Appendix iv
SUPPLEMENTARY DATA FOR CHAPTER 4

Figure S4.1: Phylogenetic tree of the beta subunit of the adaptin protein family, including the divergent AP5 subunits.
This tree was assembled using RAxML on the CIPRES web server and is available as a .nwk file in Online Supplementary Data, Appendix IV.

Figure S4.2: Phylogenetic tree of the sigma subunit of the adaptin protein family, including the divergent AP5 subunits.
This tree was assembled using RAxML on the CIPRES web server and is available as a .nwk file in Online Supplementary Data, Appendix IV.

Table S4.1: Information on ciliate genomes and transcriptomes, their provenance, and completeness
This file is available as an .xslx file in Online Supplementary Data, Appendix IV.

Table S4.2: Membrane trafficking components across the diversity of ciliates.
This file is available as an .xslx file in Online Supplementary Data, Appendix IV.