



UPPSALA  
UNIVERSITET

*Digital Comprehensive Summaries of Uppsala Dissertations  
from the Faculty of Science and Technology 2054*

# Environmental sequencing to infer patterns of eukaryotic evolution

*Combining long-read and short-read metabarcoding*

MAHWASH JAMY



ACTA  
UNIVERSITATIS  
UPSALIENSIS  
UPPSALA  
2021

ISSN 1651-6214  
ISBN 978-91-513-1242-2  
URN urn:nbn:se:uu:diva-446935

Dissertation presented at Uppsala University to be publicly examined in Ekmansalen, Evolutionary Biology Centre (EBC), Norbyvägen 14, Uppsala, Friday, 10 September 2021 at 13:00 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Professor Micah Dunthorn (Natural History Museum, University of Oslo).

**Online defence:** <https://uu-se.zoom.us/j/69496814472>

### Abstract

Jamy, M. 2021. Environmental sequencing to infer patterns of eukaryotic evolution. Combining long-read and short-read metabarcoding. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 2054. 70 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-513-1242-2.

Our view of eukaryotes is biased towards plants, animals, and fungi. But the vast majority of eukaryotic diversity is microbial in nature. These microbial eukaryotes are key players in all ecosystems on earth and are collectively known as protists. Over the past decade we have gathered a better understanding of environmental protist diversity and ecology through metabarcoding studies, which routinely generate millions of reads corresponding to short fragments (< 500 bp) of the 18S gene. However, the limited phylogenetic signal of these short reads hinders their use in investigating questions of an evolutionary nature.

To overcome this limitation, we introduced a method for long-read metabarcoding in Paper I of this thesis. We validated this method by amplifying DNA from three soil samples and sequencing with PacBio to obtain a ca. 4500 bp region of the ribosomal DNA operon spanning the 18S and 28S genes. The long-reads were taxonomically annotated using a phylogeny-aware approach, and were used to infer robust 18S-28S phylogenies of the environmental diversity.

In Paper II, we investigated habitat evolution across the eukaryotic tree of life, using a unique combination of long-read and short-read metabarcoding data in a phylogenetic framework. We showed that transitions across the marine-terrestrial habitat boundary are more frequent than previously assumed, and that eukaryotic groups vary in their ability to cross this habitat boundary. We inferred that the last eukaryotic common ancestor inhabited non-marine environments, and that subsequent transitions across the marine-terrestrial boundary likely played a key role in eukaryotic evolution by opening new niches to fill.

Paper III focused on determining the effects of habitat and latitude on the rates of molecular evolution of protists. Analyses on phylogenies inferred from long-read metabarcoding data found no systematic differences in the evolutionary rates of marine and terrestrial species. Additionally, contrary to expectations, not all eukaryotic groups showed an increase in evolutionary rates towards the equator, with some groups displaying the opposite trend.

Finally Paper IV isolates the parasite of the endangered freshwater pearl mussel in Sweden, and phylogenetic analyses including long-read metabarcoding data identifies it as a gregarine belonging to the genus *Nematopsis*.

In summary, this thesis introduces a new method for environmental sequencing of protists, and urges future studies to use both long-read and short-read metabarcoding data to study outstanding questions in eukaryotic evolution and ecology.

**Keywords:** Protists, eukaryotes, environmental sequencing, ribosomal DNA, PacBio, long-read metabarcoding, phylogenetics, salt barrier, habitat evolution, eukaryotes evolution, evolutionary rate, Apicomplexa, pathogen

*Mahwash Jamy, Department of Organismal Biology, Systematic Biology, Norbyv. 18 D, Uppsala University, SE-75236 Uppsala, Sweden.*

© Mahwash Jamy 2021

ISSN 1651-6214

ISBN 978-91-513-1242-2

URN urn:nbn:se:uu:diva-446935 (<http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-446935>)

*To my parents*



# List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

- I     **Jamy, M.**, Foster, B., Barbera, P., Czech, L., Kozlov, A., Stamatakis, A., Bending, G., Hilton, S., Bass, D., Burki, F. (2020) Long-read metabarcoding of the eukaryotic rDNA operon to phylogenetically and taxonomically resolve environmental diversity. *Molecular Ecology Resources*, 20(2):429–443
- II    **Jamy, M.**, Biwer, C., Vaultot, D., Obiol, A., Jing, H., Peura, S., Massana, R., Burki, F. (2021) Global patterns and rates of habitat transitions across the eukaryotic tree of life. *Manuscript*
- III   **Jamy, M.**, Vaultot, D., Burki, F. (2021) Habitat, latitude, and the rate of molecular evolution in eukaryotes. *Manuscript*
- IV    Alfjorden, A., Brännström, I. O., Wengström, N., **Jamy, M.**, Kristmundson, A., Jansson, E., Burki, F. (2021) Identification of a new gregarine parasite [Apicomplexa, Alveolata] in mass mortality events of freshwater pearl mussels (*Margaritifera margaritifera*). *Manuscript*

Reprints were made with permission from the respective publishers.

The following papers were published or submitted during the course of my doctoral studies, but are not part of this thesis.

Burki, F., Sandin, M. M., **Jamy, M.** (-) Diversity and ecology of protists revealed by metabarcoding. *Current Biology*. Submitted (invited review)

Strassert, J. F. H., **Jamy, M.**, Mylnikov, A. P., Tikhonenkov, D. V., Burki, F. (2019) New phylogenomic analysis of the enigmatic phylum Telonemia further resolves the eukaryotic tree of life. *Molecular Biology and Evolution*. 36(4):757-765

Avia, K., Lipinska, A., Mignerot, L., Montecinos, A., **Jamy, M.**, Ahmed, S., Valero, M., Peters, A., Cock, J., Roze, D., Coelho, S. (2018) Genetic diversity in the UV sex chromosomes of the brown alga *Ectocarpus*. *Genes*. 9(6):286

Onsbring, H., **Jamy, M.**, Ettema, T. J. G. E. (2018) RNA sequencing of *Stentor* cell fragments reveals transcriptional changes during cellular regeneration. *Current Biology*. 28(8):1281-1288.e3

# Contents

Introduction .....	11
Note .....	13
A brief overview of protist diversity and classification .....	14
Metabarcoding to reveal environmental protist diversity .....	17
The general metabarcoding workflow .....	17
What does a metabarcoding dataset represent? .....	20
A shift away from phylogeny-based analyses (and back again) .....	21
Long-read metabarcoding .....	24
Which sequencing platform should be used? .....	25
Insights into protist diversity from metabarcoding .....	28
How has evolution shaped protist diversity? Integrating ecology and evolution .....	31
What can phylogenies tell us? .....	33
Phylogenetic placement .....	33
Ancestral State Reconstruction .....	35
Inferring rates of evolution .....	37
Integrating metabarcoding data in a phylogenetic framework .....	40
Research Aims .....	41
Paper Summaries .....	42
Paper I. Long-read metabarcoding of eukaryotes .....	42
Paper II. Habitat evolution across eukaryotic tree of life .....	42
Paper III. Rate of molecular evolution in protists .....	43
Paper IV. A gregarine parasite infecting freshwater pearl mussels .....	44
Conclusions and Future Perspectives .....	45
Popular Science Summary .....	47
Svensk sammanfattning .....	49
خلاصہ برائے پاپولر سائنس .....	51

Acknowledgements .....53

References .....58

# Abbreviations

ASR	Ancestral state reconstruction
bp	Base pairs
DNA	Deoxyribonucleic acid
EPA	Evolutionary placement algorithm
GTR	General time reversible model, DNA substitution model
ITS	Internal transcribed spacer of the ribosomal operon
LDG	Latitudinal diversity gradient
LSU	Large subunit (28S) ribosomal gene
MCMC	Markov chain Monte Carlo
ML	Maximum likelihood
PacBio	Pacific Biosciences
PCR	Polymerase chain reaction
PR <sup>2</sup>	Protist Ribosomal Reference database
rDNA	Ribosomal DNA
RNA	Ribonucleic acid
SMRT	Single-molecule real-time sequencing
SSU	Small subunit (18S) ribosomal gene



# Introduction

*“These animalcules had divers colours, some being whitish and transparent; others with green and very glittering little scales; others again were green in the middle, and before and behind white; others yet w’ere ashen grey. And the motion of most of these animalcules in the water was so swift, and so various, upwards, downwards, and round about, that ’twas wonderful to see”*

[Leeuwenhoek’s discovery of protists and bacteria in Sep, 1674 (Dobell and Van Leeuwenhoek 1932)]

There are an estimated 8.7 million species of eukaryotes on Earth (Mora et al. 2011). Much of the work on characterizing this impressive diversity has been done on more conspicuous eukaryotes: animals, plants and fungi. However, these familiar groups are only a few branches on the eukaryotic tree of life (Martin 2015). It is in fact Leeuwenhoek’s “animalcules”, or more accurately, microbes, ranging in size from less than a micrometre to several millimetres, that comprise the vast bulk of eukaryotic diversity (Baldauf et al. 2000; Burki et al. 2020). Collectively known as protists, these microbes come in a breathtaking array of forms: ciliates with complex feeding structures that can be re-grown if lost, radiolarians with intricate mineral skeletons, symmetrical colonies of green algae, tiny armour-plated haptophytes, or glass-encased diatoms, to name only a few relatively well-studied examples. Protists have much to offer to those interested in understanding how life works in all its glorious facets. They are inherently interesting because of their peculiar cell biology, one example being the ciliates with their micro- and macronucleus, that differs from the standard textbook picture. So studying them can reveal the full breadth of biology in all its complexity, which otherwise would not be possible by looking at animals or plants alone. They hold clues important for solving the puzzle of eukaryote evolution and eukaryogenesis (Koonin 2010). Furthermore, they are key members of microbial communities that drive all ecosystems on earth. They are important primary producers, generating 40% of the world’s oxygen (Corliss 2004), and also perform critical roles in our ecosystems as decomposers, predators and parasites.

Although the small size of protists and our inability to culture most species in the lab has presented significant challenges historically, our knowledge about protist diversity and ecology has rapidly accumulated in recent years. This is

largely owing to recent technological advances that allow the extraction and sequencing of genetic material directly from environmental samples. Environmental sequencing, especially metabarcoding, has proven particularly useful for revealing unprecedented protist diversity and characterizing how microbial eukaryotic communities vary in space and time. However, current methodologies are limited by short sequencing length, often less than 500 base pairs (bp). Consequently, the data generated have limited phylogenetic signal, impairing our ability to investigate ecological and evolutionary questions that rely on phylogenetic analyses. Furthermore, it can be difficult to analyse novel lineages and characterise their position in the tree of life, especially if the novel lineage is very divergent from known reference sequences.

In this thesis, I introduce high-throughput long-read metabarcoding of natural eukaryotic communities whereby we sequence a ~4500 bp fragment of the ribosomal DNA operon, spanning the 18S and 28S genes (**Paper I**). The increased phylogenetic signal of the long-reads enables phylogeny-aware taxonomic annotation and the inference of more robust phylogenies of environmental diversity, thereby providing an evolutionary perspective of natural communities. Importantly, these taxon-rich phylogenies also allow us to incorporate the vast amount of existing short-read metabarcoding. With the resulting comprehensive phylogenies, we can investigate questions of an evolutionary nature using the full-scale of environmental diversity. In **Paper II**, we use these combined data to answer the following questions: How often do eukaryotes transition between marine and non-marine habitats? Are certain groups more adept at crossing this habitat boundary? Which environment did major eukaryotic clades originate in? In **Paper III** we investigate the effect of habitat and latitude/climate on the rate of molecular evolution. Do marine taxa evolve faster than non-marine taxa? Do species near the equator have faster evolutionary rates? Finally, in **Paper IV**, we use environmental sequencing, histology and imaging techniques to isolate and describe an apicomplexan parasite from the endangered freshwater pearl mussel. We use our long-read dataset to phylogenetically characterise the parasite and examine its prevalence in natural communities.

In the following summary text, I provide an overview of standard metabarcoding techniques, long-read metabarcoding, and insights into protist diversity and ecology provided by metabarcoding. The patterns of protist diversity and ecology raise several fundamental questions at the crossroads of ecology and evolution, and to investigate them, I advocate for the use of long-read and short-read metabarcoding. Answering these questions requires the use of several phylogeny-based methods, and I introduce these to the reader. Finally, I finish with the most important findings from this thesis.

# Note

Due to format constraints, some of the supplementary material from this thesis (alignments, and large phylogenies) is available on a Box folder, and will be maintained until 1<sup>st</sup> November 2021:

<https://uppsala.box.com/s/v2ulcxxkvzjoyhard2frq4sct26yqjn>

# A brief overview of protist diversity and classification

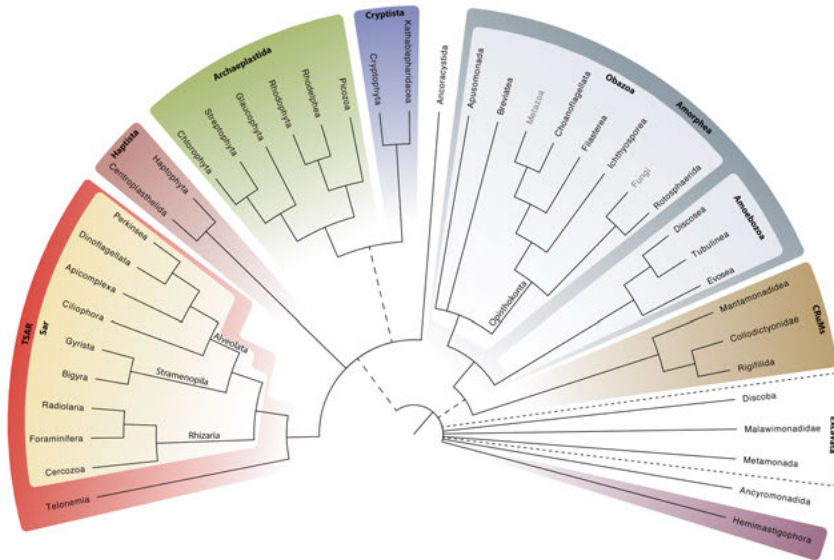
How are all eukaryotes on earth related to each other? While this is still an active area of research, significant progress has been made, particularly in the last 20 years (Baldauf et al. 2000; Burki et al. 2007, 2020; Brown et al. 2018; Strassert et al. 2019). Our current view of the eukaryotic tree of life (Figure 1) has largely been informed by phylogenomics, i.e. phylogenetic inferences using datasets involving several hundred genes, which is especially necessary in order to resolve deep-branching nodes (Burki 2014; Burki et al. 2020). The eukaryotic tree of life is typically divided into several “supergroups”—large groups for which we have reasonable evidence of their monophyly. These supergroups can include several traditional “kingdom”-level clades (e.g. the animal and fungal kingdoms are included in the supergroup Obazoa) (Figure 1). It should be noted that the root of the eukaryotic tree remains elusive, and several studies have proposed different roots over the years (Derelle and Lang 2012; He et al. 2014; Derelle et al. 2015; Jewari and Baldauf 2021).

Below, I briefly introduce the main eukaryotic lineages.

## **TSAR**

This group is composed of **T**elonemia, **S**tramenopila, **A**lveolata, and **R**hizaria, of which the latter three form the group Sar. Altogether, Sar is thought to contain nearly half of all protist diversity (Del Campo et al. 2014). Rhizaria includes groups such as radiolarians and foraminifera which form intricate skeletons and shells made of silica and calcium carbonate respectively. Also included are the cercozoans, a diverse group covering parasites of animals, plants, and algae (Sierra et al. 2016; Bass et al. 2019), heterotrophic flagellates dominating soils, voracious predators sucking out their prey’s cell contents, and many others. Stramenopila comprise diatoms, golden algae, the plant pathogen group oomycetes (Gyrista), as well as the network producing labyrinthulomycetes, **M**Arine **S**Tramenopiles and other flagellates (Bigyra). Finally, the Alveolata include ciliates, dinoflagellates, and the parasitic apicomplexa and perkinsids.

Haptista is made up of the photosynthetic haptophytes, and the heterotrophic sun-animalcules aka centrohelids.



**Figure 1.** Current view of the eukaryotic tree of life. Supergroups are indicated in different colours. Branches for which there are more uncertainties are shown as dashed lines. Lineages shown correspond to the fourth rank in the PR<sup>2</sup> database (*PR2\_transitions*; e-Supp Material), and black text indicates lineages composed (at least partially) of protist diversity. Figure adapted from Burki et al. 2020.

This group contains lineages with primary plastids (chloroplasts) such as land plants, green algae, red algae, and glaucophytes. Recently, it was found to also include non-photosynthetic lineages such as rhodelphids, and the picozoa (Gawryluk et al. 2019; Schön et al. 2021).

This group is home to the cryptophytes which are photosynthetic algae common in freshwater, and the heterotrophic kathablepharids.

Amorphea is composed of two large groups: Obazoa and Amoebozoa. The former includes animals, fungi, and their single-celled relatives (such as

choanoflagellates, breviate and apusomonads). Amoebozoa on the other hand contain a large diversity of amoeboid protists and slime moulds.

### **CRuMs**

This grouping puts together several free-living protist lineages: namely the collodictyons, rigifilids, and the mantamonads (Brown et al. 2018).

### **Excavata**

This is the only group that is supported by morphological evidence (a feeding groove “excavated” on one side of the cell), but not well supported by molecular phylogenies (Simpson and Patterson 1999; Derelle et al. 2015). It tentatively contains three main lineages: metamonads, discobids, and malawimonads. While most metamonads are symbionts or parasites with highly reduced mitochondria and live in low-oxygen environments, discobids mostly have non-reduced mitochondria and are either free-living or parasitic.

### **Hemimastigophora**

This is the most recently proposed supergroup, and is perhaps the most depauperate, containing only the hemimastigotes, flagellated cells occurring in soils and freshwater in low abundance (Lax et al. 2018).

# Metabarcoding to reveal environmental protist diversity

A small sample of soil taken from a football field will contain hundreds of thousands of eukaryotic cells. The same holds true for a sample taken from a pond in Uppsala, or from the Pacific Ocean—in fact from anywhere on earth. Which species are present in these samples? Do we find different protists in each of these samples? There are two main ways to tackle these questions. One could for instance, examine these samples under a microscope and try to identify species using morphology. However, while valuable, this method is slow, low-throughput, requires expertise, and it is difficult to morphologically distinguish protists smaller than 5  $\mu\text{m}$  or so (Massana 2015). Furthermore, some organisms might look morphological identical, but in fact represent different species (aka cryptic species) (Sarno et al. 2005; Gaonkar et al. 2017; De Luca et al. 2021). The other option is to extract DNA directly from these samples and sequence it (environmental sequencing). Environmental sequencing comes in two main flavours. The first is shotgun metagenomics and meta-transcriptomics which involve sequencing the entire DNA (or cDNA) content in a sample, and then assembling into genomes/transcriptomes. Metagenomic assembly is extremely challenging, even more so for eukaryotes due to the large size and complexity of their genomes. The problem is further exacerbated by the lack of reference genomes for much of eukaryotic diversity. For this reason, most environmental sequencing studies on protists have not used metagenomics (though there have been several recent exciting developments in this field, for example West et al. 2018; Richter et al. 2019; Delmont et al. 2020; Obiol et al. 2020). The other method of environmental sequencing, and the one that I will focus on in this thesis, is metabarcoding. Metabarcoding involves amplifying and sequencing a specific molecular marker, an approach that has proved to be highly valuable for studying environmental protist diversity.

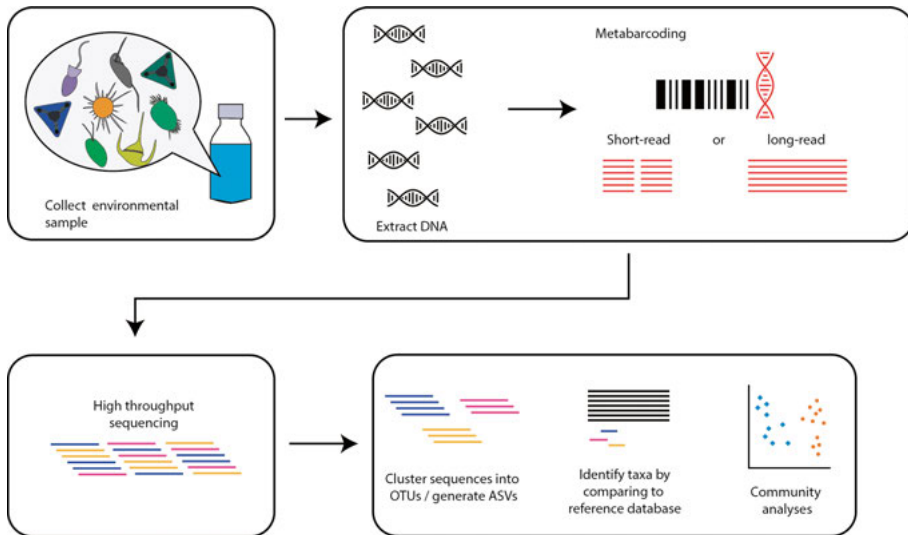
## The general metabarcoding workflow

The general metabarcoding workflow consists of a series of steps in the field, the lab, and then bioinformatics (illustrated in Figure 2; see Creer et al. 2016;

Taberlet et al. 2018; Santoferrara 2019) for a more detailed overview of the topic).

### Sample collection

Once a study has been designed, the first step is to collect environmental samples. The type, location, timing, volume, and number of samples depends on the scientific question under consideration. Additionally, samples are also handled differently depending on the source material—for instance, protists in aquatic samples are often separated into distinct size fractions in order to study the different size groups independently.

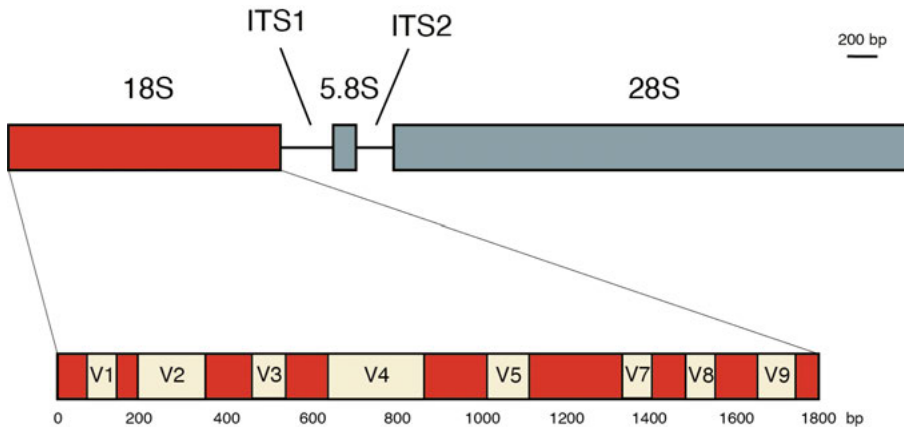


**Figure 2.** A simplified workflow of metabarcoding starting from sample collection to wet lab work, sequencing and ecological analyses.

### Lab work

Next, total DNA and/or RNA is extracted from the environmental samples, which involves lysing cells to release their genetic material. A targeted DNA region (or barcode) is then amplified with PCR and subsequently sequenced on a high-throughput sequencing platform. The choice of this barcode depends on the taxonomic group under consideration: for fungi, it is the internal transcribed spacer (ITS) in the ribosomal operon that is most widely used (Schoch et al. 2012); the cytochrome oxidase c subunit (COI) marker is commonly used for animals (Hebert et al. 2003); and plant researchers favour the ribulose-1,5 biphosphate carboxylase-oxygenase large subunit (rbcL) gene (Newmaster et al. 2006). But when studying broad eukaryotic diversity, it is the 18S (SSU) gene that is most commonly used due to its universality, ease

of amplification, and high taxonomic information (Figure 3) (Hillis and Dixon 1991). In particular, large-scale metabarcoding studies in the past decade have focused on the short, hypervariable V4 and V9 regions of the 18S gene (de Vargas et al. 2015; Duarte 2015; Kopf et al. 2015; Mahé et al. 2017; Santoferrara et al. 2020). While metabarcoding is often associated with amplifying and sequencing short (< 500 bp) genetic markers, in **Paper I** of this thesis, we present a method for long-read metabarcoding (see “Long-read metabarcoding”). Unless targeting a specific protist group, general eukaryotic primers are used to recover the entire eukaryotic community (but see the section “What does a metabarcoding dataset represent?”).



**Figure 3.** The top panel shows a general overview of the rDNA operon showing the 18S, 5.8S and 28S genes (5S is encoded elsewhere). Interstitial spacers, ITS1 and ITS2 separate the 18S and 5.8S genes and 5.8S and 28S genes respectively. The lower panel shows a more detailed schematic of the 18S gene. Darker regions represent conserved regions while lighter regions represent variable regions and are labelled 1-9.

## Bioinformatics

Raw sequencing data needs to be processed before downstream analyses. These filtering steps often include culling poor-quality reads, detecting and removing chimeras and other artefacts, handling sequencing errors, and finally, clustering sequences into operational taxonomic units (OTUs) or amplicon sequencing variants (ASVs). It should be noted that OTUs rarely reflect a direct correspondence to species, but instead are a practical solution for handling large data, and dealing with sequencing errors. Over the years, a number of popular tools and pipelines have been established for processing and clustering short-read metabarcoding data such as qiime, mothur, vsearch, dada2, swarm, and lulu (Schloss et al. 2009; Caporaso et al. 2010; Mahé et al. 2014; Rognes et al. 2016; Frøslev et al. 2017). On the other hand, fewer pipelines

and dedicated tools exist for long-read metabarcoding which is a much newer technology. In this thesis, we present a new pipeline to process long-read metabarcoding data (**Papers I and II**).

An important step after read clustering is assigning taxonomy to sequences. Again, a number of tools exist to tackle this, with the common principle to perform sequence similarity searches against reference databases such as PR<sup>2</sup> and SILVA (Guillou et al. 2012; Quast et al. 2013). These methods are rapid and well suited for large datasets, however, they do not perform well for sequences that are too divergent from labelled, reference sequences (Berger et al. 2011). Indeed, sequences less than 80% similar to known references are often discarded in metabarcoding studies (de Vargas et al. 2015). One method to identify such divergent sequences is to phylogenetically “place” them on to reference phylogenies inferred from known reference sequences (Berger et al. 2011; Barbera et al. 2019). However, this too requires labelled, reference sequences which are often generated through Sanger sequencing, a low-throughput and labour-intensive process. Fortunately, the development of long-read metabarcoding (**Papers I and II**) presents a high-throughput way of populating these reference databases.

Following sequence annotation, analyses to investigate ecological questions are carried out, ranging from the calculation of alpha and beta diversity, network analyses, ordination methods, and phylogeny-based analyses.

## What does a metabarcoding dataset represent?

Metabarcoding studies routinely generate millions of barcodes. But how well do these datasets correspond to the actual community present in the source samples? Numerous studies over the years have highlighted biases that can arise during different metabarcoding steps, and how to best alleviate them (Taberlet et al. 2018; Harrison et al. 2019; Santoferrara 2019). First, although we use general eukaryotic primers, not all members of the community are amplified. No primer pair is truly universal, and therefore certain taxa might be missed in biodiversity surveys. For example, animal parasites belonging to the fast-evolving rhizarian group, Ascetosporea are often missed by general primers, and the marine-dominant diplomonad clade is completely absent in V4 datasets (Mukherjee et al. 2015; Bochdansky et al. 2017; Bass et al. 2019). Furthermore, obligate intracellular parasites in general may be difficult to pick up in biodiversity surveys unless their hosts are specifically targeted (Hartikainen et al. 2014; Ward et al. 2018; Bass et al. 2019). Additionally, environmental sequencing studies rarely capture all of the components of the rare biosphere, though high-throughput sequencing does alleviate this problem to a certain extent. On the flip side, not all sequences in metabarcoding

datasets represent biologically active residents; some sequences may be sourced from dead or lysed cells, other sequences may represent cells dispersed from other habitats (e.g. surface runoff into coastal waters), and still others may simply be chimeras or other sequence artefacts (Taberlet et al. 2018; Harrison et al. 2019; Santoferrara 2019; Gottschling et al. 2020).

Comparisons with mock communities and metagenomic datasets have also revealed that relative abundances of taxa can be skewed due to preferential amplification during PCR, or due to high 18S copy number in cells (Elbrecht and Leese 2015; Krehenwinkel et al. 2017; Pitsch et al. 2019; Obiol et al. 2020). This may for example be the case for diplomonads which have multiple rDNA copies, and their dominance in the oceans may therefore have been overestimated (Mukherjee et al. 2020).

Finally, it is worthwhile to consider what our diversity estimates are actually measuring. As previously mentioned, OTUs (and ASVs) do not necessarily correspond to species. Indeed, some OTUs may represent multiple species with highly similar 18S sequences (diversity underestimation). On the other hand, some OTUs may be derived from the same organisms and reflect intra-genomic variability (diversity overestimation) (Caron and Hu 2019).

Despite all of these biases and limitations, metabarcoding is currently the best and most cost-effective tool for assessing the diversity of natural microbial communities in detail. Comparisons with metagenomic approaches (which should be unbiased by the PCR step), microscopy and mock communities have revealed that overall, metabarcoding provides a robust, useful, and qualitative or semi-quantitative view of eukaryotic communities (Santoferrara 2019; Obiol et al. 2020).

## A shift away from phylogeny-based analyses (and back again)

The different sequencing technologies available over the years have heavily shaped the possibilities of research on microbial communities. Initial studies in the early 2000s used Sanger sequencing which was costly and involved the laborious and time-intensive step of cloning amplicons before sequencing (López-García et al. 2001; Moon-Van Der Staay et al. 2001). These studies regularly obtained 18S sequences of around 1000 bp or more which were analysed in a phylogenetic context to assess novel diversity. It was by inferring phylogenies with environmental and reference sequences that entirely novel clades were revealed, such as the Marine Stramenopiles (MASTs) and Marine

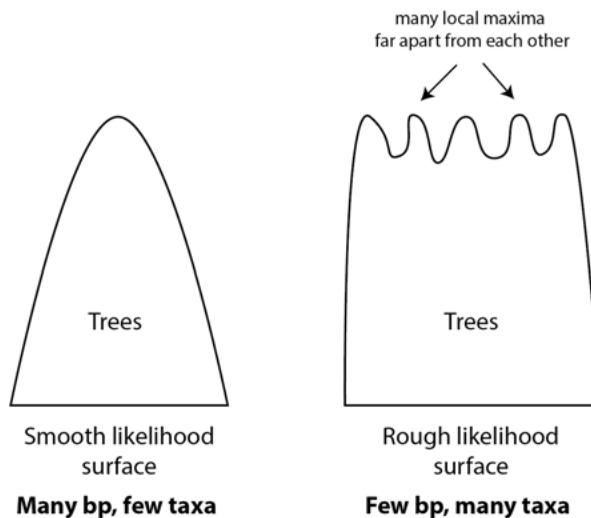
Alveolates (MALVs) (López-García et al. 2001; Moon-Van Der Staay et al. 2001; Massana et al. 2004a,c).

Later, the emergence of 454 pyrosequencing quickly followed by Illumina and Ion Torrent technologies, enabled the generation of millions of reads corresponding to short ( $\leq 500$  bp) V4 or V9 fragments of the 18S gene, resulting in the bulk of metabarcoding surveys that we are familiar with today (Amaral-Zettler et al. 2009). While these short-reads capture a much greater proportion of the eukaryotic community, they contain little phylogenetic signal (Box 1). We thus saw a general shift away from phylogeny-based analyses (though there are some exceptions e.g. Dunthorn et al. 2014a; Mahé et al. 2017; Lewitus et al. 2018; Lentendu and Dunthorn 2021). Current short-read metabarcoding sequencing studies do detect novel diversity in environments, but cannot precisely determine its phylogenetic affiliation. For example, one-third of the diversity sequenced by a global marine expedition (Tara Oceans) could not be taxonomically assigned to any group, including 11 cosmopolitan OTUs (de Vargas et al. 2015). Therefore, Sanger sequencing is still the technology of choice for exploring group-specific diversity where obtaining longer sequences to enable phylogenetic inference is desirable (e.g. Lara et al. 2016; Ward et al. 2018), though we note that this role might be fulfilled by long-read metabarcoding in the future.

Poor taxonomic annotation of divergent sequences is not the only problem that arises due to the little phylogenetic signal of short-reads. Indeed, it can also be challenging to use this data to investigate questions of an evolutionary nature, which by nature, often require phylogenetic inference (Dunthorn et al. 2014a). Some of these challenges are overcome by the development of a suite of tools dedicated to phylogenetic placement (Matsen et al. 2010; Berger et al. 2011; Barbera et al. 2019; Czech et al. 2020). Additionally, the development of long-read metabarcoding is now allowing taxon-rich, robust, phylogenetic inference (**Papers I and II**), thus alleviating some of the challenges traditionally faced by metabarcoding.

**Box 1. Why is it difficult to infer phylogenies from short-read metabarcoding data?**

Short-read metabarcoding datasets typically contain many thousands of taxa, and very few bp. The first reason why it can be difficult to infer phylogenies is that there is simply not enough data! Second, it can be difficult to parallelize the tree search algorithm because there are so many taxa, greatly inflating the tree space. And finally, due to the “shape” of the dataset (many taxa, few bp), accuracy suffers (Bininda-Emonds et al. 2001). This is because the tree space has a “rough likelihood surface” (see figure below), where a large number of local optima exist that cannot be distinguished from each other statistically (Stamatakis et al. 2020).



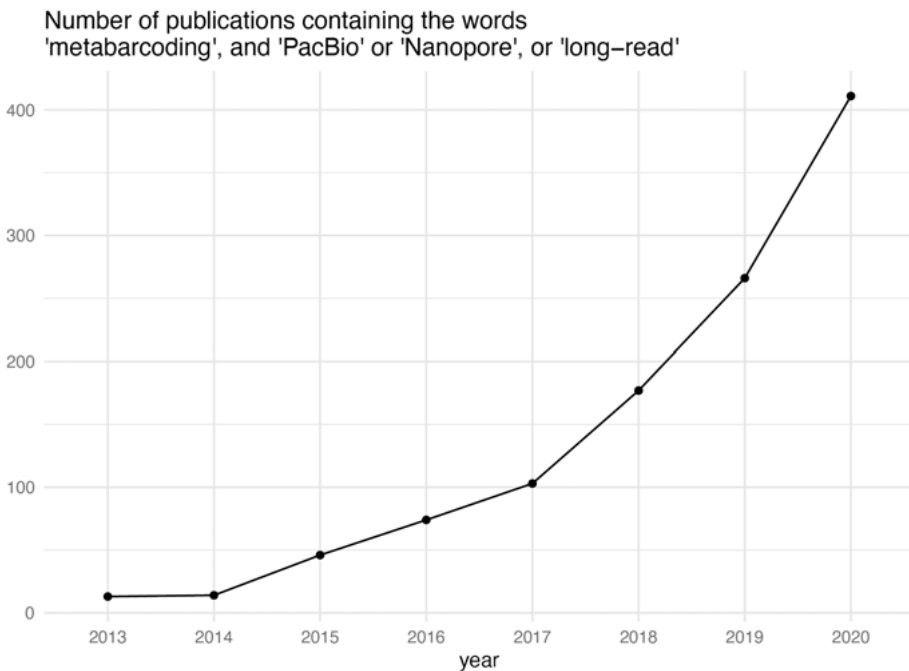
A phylogenomics dataset can be considered a “well-shaped” dataset, while a short-read metabarcoding dataset can be considered a “badly-shaped” dataset, with many local optima. Figure adapted from a lecture given by A. Stamatakis (Stamatakis 2018).

# Long-read metabarcoding

The first long-read metabarcoding study appeared in 2013, in which Mosher et al. amplified the entire bacterial 16S gene (1500 bp) on the Pacific Biosciences (PacBio) RS platform (Mosher et al. 2013). This study estimated error rates to be as high as 17-18%, far too high to reliably assess microbial diversity. Since then, third-generation sequencing technologies, namely PacBio and Nanopore, have improved drastically and have much lower error rates that are comparable to Sanger sequencing (Loit et al. 2019). And while the microbial ecology community has yet to broadly adopt these technologies for metabarcoding, long-read metabarcoding is becoming increasingly popular (Figure 4), and several proof-of-concept studies have been published recently, including **Paper I** of this thesis (Schloss et al. 2016; Wagner et al. 2016; Heeger et al. 2018; Tedersoo et al. 2018; Martijn et al. 2019; Jamy et al. 2020).

These sequencing technologies generate reads of length 1500-6000 bp (Heeger et al. 2018; Orr et al. 2018; Jamy et al. 2020), long enough to contain the complete 18S gene (covering both V4 and V9), or even the entire ribosomal DNA operon (Figure 3), including the ITS region and the 28S gene. The increased length and phylogenetic signal of these reads presents a number of advantages. First, the increased genetic information allows better taxonomic classification of environmental sequences compared to the short V4 or V9 sequences, regardless of the taxonomic annotation method used (Heeger et al. 2018; Chung et al. 2020; Jamy et al. 2020; Jeong et al. 2021). Second, these reads can potentially be used to rapidly populate reference databases, not just for the 18S gene, but for the ITS region and 28S gene as well. The long-reads contain all these markers linked together, which is useful for establishing associations between them. Third, these reads can also be used to infer taxon-rich reference phylogenies on to which existing short-read data can be placed, as we show in **Paper II** of this thesis. Fourth, by inferring phylogenetic trees, we can visualize and assess the phylogenetic diversity of an environmental sample (**Paper I**), and potentially discover novel clades along with resolving their position in the tree of life (Jamy et al. 2020; Kolaříková et al. 2021; Strasser et al. 2021). Finally, these taxon-rich phylogenies can allow us to explore eukaryotic biodiversity in an eco-evolutionary context (**Papers II and III**; see “How has evolution shaped protist diversity?”), thus opening up an exciting avenue of research. As always, there are potential pitfalls associated with the method which remain to be critically assessed. For instance, long-

range PCR are more prone to chimera formation (though this problem can be somewhat alleviated by reducing the number of PCR cycles; Heeger et al. 2018). It seems likely however, that the benefits far outweigh the risks.



**Figure 4.** A trend of the increasing popularity of long-read metabarcoding. The number of studies each year in Google Scholar containing the words “metabarcoding” and “PacBio”/“Nanopore”/“long-read” was plotted.

## Which sequencing platform should be used?

There are three main options for sequencing long-read amplicons: Nanopore, PacBio, and Synthetic Long Reads (SLR; now commercially available as LoopSeq). Each has its own sets of strengths and weaknesses and ultimately choosing the right sequencing platform depends on the study design, the cost, the taxonomic scope of the study, and the genetic marker being used. Below, I will briefly present the three sequencing technologies and discuss when they are best suited considering their pros and cons.

## PacBio

Like Nanopore, PacBio sequencing too has a high raw error rate, estimated to be around 14% (Rhoads and Au 2015). However, unlike Nanopore, this high error rate can be largely overcome with Circular Consensus Sequencing (CCS; Figure 5) (Schloss et al. 2016). Briefly, during sequencing library prep, hairpin adapters are ligated to the ends of amplicons, thus generating circularized DNA molecules (SMRT bell template; Figure 5). During sequencing, the polymerase goes round the DNA molecule several times and each complete run through the forward or reverse strand is termed a “pass”. A circular consensus sequence (CCS) is generated from all passes of a molecule during post-processing. This CCS has a substantially lower error rate since the errors are randomly distributed, and is comparable to Sanger sequencing (Loit et al. 2019). Furthermore, this error rate tends to decrease as the number of passes increases (Schloss et al. 2016). This makes PacBio sequencing suitable for assessing highly complex communities, and it is for this reason that we opted for this platform for the research presented in this thesis. In **Paper I**, we present data from three soil communities sequenced on three SMRT (single molecule real time) cells on Sequel I, while in **Paper II**, we present additional data from 18 samples that were sequenced on four SMRT cells on the newer Sequel II instrument.



**Figure 5.** The panel on the left shows a double stranded DNA SMRT bell template (forward and reverse strands depicted in orange and green respectively), circularized by hairpin adapters (yellow). The polymerase molecule (in grey) sequences the circular template multiple times, generating subreads or “passes” shown in the panel on the right. Each pass has a high proportion of randomly distributed sequencing errors, which are removed when the CCS is generated during post-processing.

## Nanopore

The MinION and other sequencers from Oxford Nanopore, famous for their portability and rapid processing of samples, operate by unwinding the DNA double helix and driving the single stranded DNA through a protein nanopore, inducing small voltage fluctuations with every base that passes through. It is these voltage changes that are measured and used to infer the sequence of nucleic acids (Branton et al. 2008). The main drawback of Nanopore sequencing

is its high raw error rate which is currently around 11 to 15% (Loit et al. 2019). The raw error rate can be remediated through generating consensus sequences from multiple reads (Loit et al. 2019; Baloglu et al. 2021), however this means that it is unsuitable for assessing complex communities, at least until technological developments lower the error rate. Currently, Nanopore sequencing is best suited for assessing the presence of pathogens (for which rapid processing is desirable), when sequencing in the field is desirable, or for sequencing simple communities or targeted lineages (Quick et al. 2015; Parker et al. 2017; Pomerantz et al. 2018; Loit et al. 2019; Strassert et al. 2021).

## LoopSeq

Synthetic Long Reads (SLRs) or LoopSeq is the newest technology available for long-read metabarcoding. This method does not actually generate long reads, but relies on Illumina sequencing of barcoded molecules (with barcodes indicating the origin of each molecule). As the name implies, the resulting sequences are stitched together using the barcode information into synthetic long reads. As a result it has the lowest error rates, around 0.005% (Callahan et al. 2021). At present, there are two kits available, one enabling 16S sequencing for prokaryotes, and another targeted for 18S and ITS sequencing for fungi, and is therefore unsuitable for studies interested in the entire rDNA operon.

# Insights into protist diversity from metabarcoding

Over the past decade, metabarcoding has opened a window to the microbial world. We've certainly come a long way from the early days of environmental sequencing, from the first studies in 2001 generating ~30 sequences on average, to large scale sampling efforts such as the TARA Oceans project which alone generated hundreds of millions of reads from the world's oceans (López-García et al. 2001; Moon-Van Der Staay et al. 2001; de Vargas et al. 2015). Both global and local biodiversity surveys have captured snapshots of the eukaryotic community at different times and places in our oceans, lakes, soils, and even extreme environments such as hot springs and acid mine drainage (Bates et al. 2013; Massana et al. 2015; de Vargas et al. 2015; Duarte 2015; Kopf et al. 2015; Pernice et al. 2016; Khomich et al. 2017; Mahé et al. 2017; Oliverio et al. 2018, 2020; Xue et al. 2018; Boenigk et al. 2018; Giner et al. 2019; Annenkova et al. 2020; Seppey et al. 2020; Wolf and Vis 2020; Luan et al. 2020). Putting together these snapshots provided by metabarcoding has truly revolutionized our understanding of protist diversity and how it is shaped. Below I briefly summarize some of the key findings revealed by metabarcoding. This list is by no means exhaustive, and I focus in particular on findings that are relevant to my thesis.

## Protists are diverse and form complex interactions

One of the most striking findings of metabarcoding surveys is just how diverse protists are. For instance, the Tara Oceans global survey estimated that our oceans are home to around ~150,000 OTUs. Some of this diversity was unknown to science before environmental sequencing studies, such as MALVs (marine alveolates)(Moon-Van Der Staay et al. 2001; Guillou et al. 2008), MASTs (marine stramenopiles)(López-García et al. 2001; Massana et al. 2002, 2004b) and MASHOL (marine small holozoan clade)(Arroyo et al. 2020). Others were lineages known to science, but their diversity and abundance had previously been underestimated. This was for examples the case for diplomonads in the surface and deep oceans (de Vargas et al. 2015; Flegontova et al. 2016), and for apicomplexans in neotropical forest soils (Eric Ma E et al. n.d.). Overall, it is heterotrophs and not autotrophs that seem to dominate each environment both in terms of abundance and diversity (with the possible

exception of freshwater systems)(Singer et al. 2021). This hyper-dominance of heterotrophs indicates that protists are heavily involved in trophic interactions, and indeed, this is supported by network and co-occurrence analyses (Lima-Mendez et al. 2015; Oliverio et al. 2020). It is also speculated that it is these interactions or symbioses, both among protists, and with other organisms, that might have been the driving factor for the vast protist diversity we see today(de Vargas et al. 2015). Some of these interactions are already being documented by combining environmental sequencing data with imaging techniques such as FISH (fluorescent in-situ hybridization) (e.g. (Massana et al. 2009; Chambouvet et al. 2019; Piwosz et al. 2021)), and single-cell genomics (Martinez-Garcia et al. 2012).

It should be noted that a big chunk of this protist diversity belongs to the so-called “rare biosphere”. All samples are usually composed of a few hyperabundant taxa, and a long tail of rare taxa (Dunthorn et al. 2014b; Logares et al. 2015). While sequencing and PCR errors may contribute somewhat to this rare biosphere, it is established that these rare taxa are real and metabolically active (though their impact in the ecosystem remains unclear) (Logares et al. 2015). Some of these taxa might be conditionally rare, becoming abundant when environmental conditions become suitable again, or they can remain permanently rare, never exceeding a certain abundance (Logares et al. 2014, 2015).

## How is protist diversity structured?

Metabarcoding studies are also revealing how protist diversity is structured in space and time. According to a recent synthesis of community ecology (Vellend 2010, 2016), there are four main processes that determine community assembly: (1) *selection* from environmental factors such as salinity, pH, predators etc. can filter out ill-adapted species from a community; (2) *dispersal* of species can allow them to enter new communities, (3) communities can fluctuate over time simply through ecological *drift*; (4) and new species are introduced over evolutionary time through the process of *speciation*. All of these processes are thought to act in conjunction, however, the relative importance of each process in different environments and scales is yet to be established (Santoferrara et al. 2020).

One of the questions resolved by metabarcoding was whether “everything is everywhere”, as elegantly formulated in the Baas Beeking hypothesis (the complete statement is “everything is everywhere, but the environment selects”)(Becking 1934). Under this hypothesis, it was posited that protists have unlimited dispersal abilities, meaning that a protist taxon can be found in all its preferred habitats, no matter how geographically distant. Numerous studies have now shown that this assumption is (almost entirely) false, and that most

protists do display biogeographies, at least in part due to dispersal limitation (Bates et al. 2013; de Vargas et al. 2015; Lentendu et al. 2018; Singer et al. 2019). For instance, a global soil study found 671 out of 672 OTUs to have restricted geographic ranges. Therefore, in order to capture the full diversity of protists, one must sample from a wide variety of locations.

Another key finding revealed by a meta-analysis of metabarcoding studies is that at the global level, protist communities can be divided into marine, and non-marine communities, with freshwater communities more similar to those in soil than in oceans (Singer et al. 2021). (Indeed, it was previously thought that for protists, freshwater and soil habitats are not so different (Robertson et al. 1997), however this has since been disputed (Sieber et al. 2020)). Many lineages dominant in marine habitats (e.g. Radiolaria) are absent in terrestrial settings, and vice-versa. Altogether this points to salinity being one of the main factors structuring protist communities at the global scale (Logares et al. 2009; Singer et al. 2021) (also see **Paper II**). When comparing the marine and terrestrial realms, the latter have been found to contain far more diversity, especially in soils (Singer et al. 2021). Terrestrial communities have also been found to be more heterogenous (Singer et al. 2021), indicating a larger role of dispersal limitation, and reflecting the many microhabitats found in soils and freshwaters (Oloo et al. 2016), as opposed to oceans where protist communities are more homogenized due to the mixing of waters (Richter et al. 2019).

Finally, metabarcoding studies can also test whether general ecological patterns known in plants and animals are also displayed by protists. One example of this is the latitudinal diversity gradient (LDG)—a poleward decline of species richness—which is one of the most well document ecological patterns in macroorganisms (Hillebrand 2004). The LDG was only very recently detected in marine surface waters (Ibarbalz et al. 2019), but notably has not been discovered in deep sea waters, or in global soils (Oliverio et al. 2020).

# How has evolution shaped protist diversity?

## Integrating ecology and evolution

Several fundamental questions arise from the patterns described in the previous section. For instance, *why* do we observe an increase in species diversity near the equator in surface oceans? Answering such questions requires an integration of ecological and evolutionary theory (Cavender-Bares et al. 2009; Hernández et al. 2013; McPeck 2017; McGill et al. 2019). This is because ecological patterns are not independent of the evolutionary history of lineages; it is evolution that influences how species interact with their environment. At the same time, ecological interactions (with the biotic and abiotic environment) shape the evolutionary trajectories of lineages. Or as George Evelyn Hutchinson so eloquently put it: the evolutionary play is carried out in the ecological theater (Hutchinson 1965). Ecology and evolution (and in particular macroecology and macroevolution) are often treated as distinct fields in biology<sup>1</sup>, but it is an artificial divide.

Investigating questions at the cross roads of ecology and evolution often requires phylogenetic inference, and for protists, the metabarcoding data that is routinely used for ecological analyses is rarely used for inferring phylogenies due its limited phylogenetic signal (Berger et al. 2011) (but see Dunthorn et al. 2014a; Lewitus et al. 2018; Lentendu and Dunthorn 2021). One might ask why it is not better to use phylogenomic data, which allows robust tree calculations, to study questions in ecology and evolution. The answer is that metabarcoding data offers the comprehensive view of protist diversity, and is often associated with excellent metadata on location, time, and measurements of abiotic variables.

In this thesis, we introduce long-read metabarcoding of the rDNA of broad eukaryotic diversity which generates data with greater phylogenetic signal (**Paper I**).

---

<sup>1</sup> It should be pointed out that there are several examples of work that link ecology and evolution such as the niche conservatism hypothesis which states that ecological niches evolve slowly along a phylogeny, and closely related species are therefore more likely to occupy similar niches (Peterson et al. 1999; Ackerly 2003). Another example is the field of paleontology which for investigates how diversity varied through different climate periods and environmental conditions over time (Hagino and Young 2015).

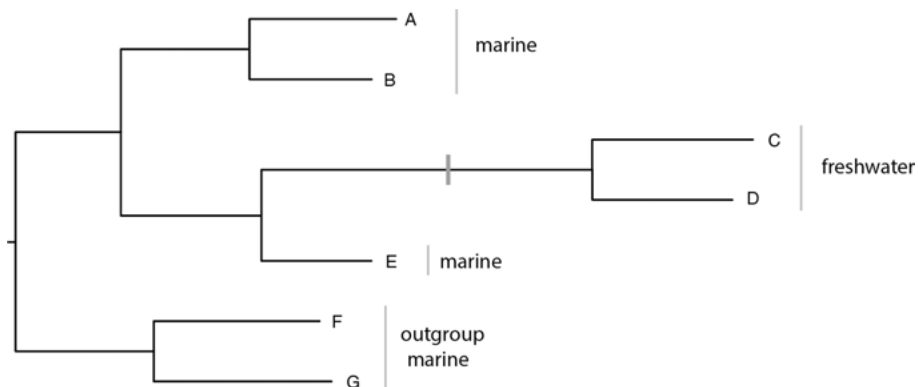
In addition, we incorporate the vast, existing short-read metabarcoding data in **Paper II** to investigate whether the distinctive protist groups in marine and terrestrial environments the result of a few habitat transitions that have occurred rarely in evolutionary history as is suggested by several studies (Von Der Heyden et al. 2004; Logares et al. 2007b, 2009; Carr et al. 2017)? Or is this diversity the result of multiple, frequent, more recent habitat transitions (Logares et al. 2007a, 2008; Žerdoner Čalasan et al. 2019; Annenkova et al. 2020)? Did eukaryotic life originate in the oceans?

In **Paper III**, we investigate whether terrestrial and marine habitats exert different pressures on their inhabitants, resulting in systematic differences in the rate of molecular evolution in terrestrial and marine protists. Furthermore, we investigate which mechanism is responsible for generating the latitudinal diversity gradient in marine surface protists. • A popular explanation is the “kinetic energy hypothesis” which posits that the higher temperature in the tropics results in higher metabolic rates and faster generation times, resulting in an increase in the rate of molecular evolution, and ultimately an increase in speciation rate (Allen et al. 2002, 2006b; Brown et al. 2004). Indeed faster rates of evolution in tropical regions have been documented for various macroorganisms (Davies et al. 2004; Wright et al. 2006; Lumbsch et al. 2008; Orton et al. 2019) as well as for foraminifera (Allen et al. 2006a). But can the kinetic energy hypothesis be generalized for all protist groups? Or do other hypotheses explaining the LDG fit better?

In order to investigate these questions, we used several phylogeny-based methods and analyses. In the next few sections, I present a brief overview of the methods used in this thesis.

# What can phylogenies tell us?

A phylogenetic tree is a hypothesis about the evolutionary relationships of taxa, usually inferred from sequence data. There are two pieces of information contained in a phylogeny which will be illustrated with the toy example in Figure 6. (1) The topology of the tree indicates that A is more closely related to B than to C. Similarly, A and B are more closely related to C, D and E than to F and G. And so on. (2) The branch lengths indicate the number of substitutions that have taken place in the respective lineages. Here, the freshwater clade (C and D) is fast evolving, or has a higher evolutionary rate (See **Inferring rates of evolution**).



**Figure 6.** A phylogenetic tree with 7 taxa, where taxa F and G are the outgroup. All taxa are marine except for C and D which are found in freshwater.

We can also infer that the ancestral lineage of all seven taxa was likely marine (using maximum parsimony), and that there was a transition from marine to freshwater somewhere along the marked branch (See **Ancestral state reconstruction**).

## Phylogenetic placement

Classifying environmental diversity based on phylogenies has been demonstrated to be far more reliable than similarity-based methods (Berger et al.

2011). However, the avalanche of reads generated from Illumina and 454 studies are impossible to deal with in the same way as Sanger sequences due to computational complexity and the relatively little phylogenetic signal due to short length (Matsen et al. 2010; Berger et al. 2011). To overcome this issue, software for “phylogenetic placement” such as pplacer (Matsen et al. 2010) and EPA (Evolutionary Placement Algorithm; (Berger et al. 2011), were developed as early as 2010, and continue to be developed (Barbera et al. 2019, 2020; Czech et al. 2020).

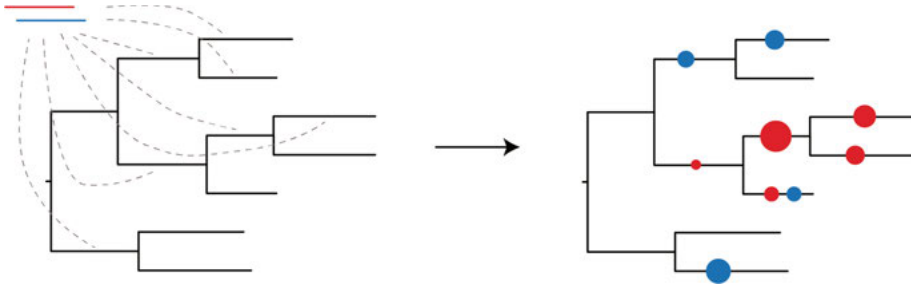
In the context of metabarcoding, phylogenetic placement “places” short reads (or queries) independently on a reference tree inferred from full length sequences. Briefly, the algorithm works by inserting the query sequences on all edges of the reference tree and calculating the associated probability for every placement location. The query is then placed on the branch with the highest probability. The end result that can be visualized is a reference tree onto which all query sequences have been grafted.

The uncertainty of these placements can be measured by two metrics: likelihood weight ratio (LWR) and expected distance between placement locations (EDPL). The LWR is simply:

$$\frac{\text{the likelihood of query placed in the most probable position}}{\text{sum of likelihood for all possible positions}}$$

That is, a high LWR indicates that the query could be confidently placed on one edge to the exclusion of all others. However, a low LWR does not always translate to a high uncertainty of placement, and for this reason, the metric is supplemented by EDPL (Figure 7). EDPL calculates the sum of the distances between optimal placements weighted by their probability; that is a low EDPL indicates that all optimal placements are located close to each other in a small subclade of the reference tree, while a high EDPL indicates that placements are widely scattered over the tree and thus cannot be confidently placed (Figure 7; Matsen et al. 2010; Mahé et al. 2017).

Phylogenetic placement can thus put query sequences in a phylogenetic context, and while few studies on eukaryotes have employed it, it is rapidly gaining popularity in prokaryotic diversity studies (Hugerth and Andersson 2017). It should be noted that longer reads can be placed more accurately than short reads (Berger et al. 2011; Quick et al. 2015). With this in mind, an important application of phylogenetic placement is taxonomic annotation (used in **Paper I**) (Kozlov et al. 2016; Czech et al. 2020). Briefly, taxonomic information at the tips of the tree are used to derive annotations for all inner nodes of the reference tree, after which placement of queries is carried out. We can use the



**Figure 7.** Placement of two sequences (red and blue) on a reference phylogeny. The probability for insertion in all possible positions is calculated (left panel) and the placement results depicted in the right panel. The red sequence represents a case where the exact placement edge cannot be confidently determined (i.e. low LWR). But the sequence can be confidently associated with a clade as the red circles are clustered in a small region of the tree (low EDPL). On the other hand, the blue sequence has a high EDPL, and it cannot be associated with any group in the tree.

information from the most optimal placements of a query to compute its taxonomy as well as the confidence scores for each taxonomic rank (Kozlov et al. 2016). Thus, for example, a query placed confidently in a subclade of the genus *Stentor* (Ciliophora) will be annotated as such, while a query placed on a deep branch might only be annotated as “Eukaryota”, as might a query that cannot be placed in any major group with confidence.

While phylogenetic placement has many applications and overcomes many of the limitations of short reads, it should be emphasized that its performance is highly dependent on the reference tree and missing taxa can result in erroneous or less informative results. Thus a current challenge is generating full-length sequences that can be used as reference taxa in the reference tree. We tackle this issue in **Paper I** of this thesis.

## Ancestral State Reconstruction

Ancestral state reconstruction allows us to study how traits evolve, and peer into the past and see what conditions may have been like during evolutionary events. It does so by considering the traits of extant taxa and the relationships between the taxa. The traits in consideration may be discrete traits (such as habitat type) or continuous (e.g. body mass, or temperature).

The most common method for reconstructing ancestral states of discrete characters is maximum parsimony (MP). MP chooses a state at an ancestral node such that the total number of evolutionary changes is minimized (Swofford and Maddison 1987). If trait change is rare, MP is a suitable option. However,

there are a number of reasons why the alternative of model-based methods based on maximum likelihood and Bayesian frameworks might be preferable. First, because MP does not take branch lengths into account, it might perform poorly if the rate of trait change is high or if branch lengths are long (Collins et al. 1994; Maddison 1994). Second, model-based methods give estimates of uncertainty which can tell us how confident we can be about the ancestral state of a node. Third, the parameters of model-based methods are themselves of interest. As an example, they can tell us whether changes from A to B of a trait are more likely than changes from B to A, and how many times more likely. In the following test, I briefly describe how models of ancestral state reconstruction work for both discrete and continuous data.

## Modelling discrete data

Let's imagine we wish to study the evolution of habitat across a phylogeny. For the sake of simplicity, let's assume it has two states: marine (M) and terrestrial (T). The most common model for discrete characters is the Mk-like model, and here we can use it to estimate the parameters  $q_{MT}$  (instantaneous rate of change from marine to terrestrial) and  $q_{TM}$  (instantaneous rate of change from terrestrial to marine) (Pagel 1994; Lewis 2001). The Mk model is a Markov model, meaning that the probability of change from one state to another depends only on the current state and is not influenced by its evolutionary trajectory. You may notice, that this model looks very similar to models of sequence evolution such as the Jukes Cantor or GTR model, and indeed they are analogous. Our model can be summarised with the following transition matrix or  $\mathbf{Q}$  matrix:

$$\begin{bmatrix} -q_{MT} & q_{MT} \\ q_{TM} & -q_{TM} \end{bmatrix}$$

Our parameters of interest are in the non-diagonals part of the matrix, while the value in the diagonals serves to have the sum of each row equal to zero. Once the transition rates have been estimated, we can obtain a matrix describing the probability of change over a branch of length  $t$ .

$$P(t) = e^{\mathbf{Q}t}$$

Note that the branch length can either be in units of time or in units of genetic change (number of substitutions per site). To illustrate, let us assume our analyses resulted in the following  $\mathbf{Q}$  matrix:

$$\begin{bmatrix} -0.6 & 0.6 \\ 0.2 & -0.2 \end{bmatrix}$$

That is  $q_{MT}=0.6$ , and  $q_{TM}=0.2$ . If we start in a marine state, what is the probability that we will be in a terrestrial state given a branch length of  $t=0.5$ ?

$$P(t) = e^{Qt} = \exp \begin{bmatrix} -0.6 & 0.6 \\ 0.2 & -0.2 \end{bmatrix} \cdot 0.5 = \begin{bmatrix} 0.29 & 0.21 \\ 0.07 & 0.43 \end{bmatrix}$$

From the probability matrix, we can see that the probability of change from marine to terrestrial along a branch length of 0.5 is 0.21. On the other hand the probability of staying in a marine state is 0.29. We can similarly estimate probabilities given varying branch lengths.

### Bayesian Methods

Ancestral states are often reconstructed along a single phylogeny, and make the assumption that the given phylogeny represent the true evolutionary history of the clade. However, phylogenies are rarely known with certainty. A Bayesian approach for modelling trait evolution can take a sample of trees as input (Pagel et al. 2004). The output is a posterior probability distribution for each parameter integrated over all the given trees. A Bayesian approach can therefore take phylogenetic uncertainty into account. It also has the added advantage of being able to estimate models that are very complex (for instance estimating  $q_{MT}$  and  $q_{TM}$  separately for all the different specified clades in a tree).

### Modelling continuous data

For traits that vary continuously, we can model trait evolution using random-walk or Brownian motion models. Under these models, traits change values randomly, and in no particular direction. Over a small amount of time (or genetic distance)  $t$ , a trait changes with a mean change of zero, and variance  $\sigma^2$ . That is, after time  $t$ , the value of a trait is  $(t \cdot \sigma^2)$ .

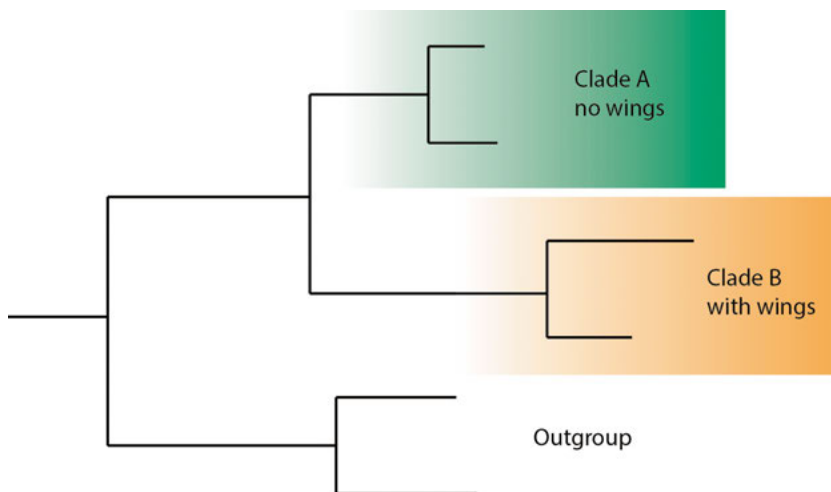
### Inferring rates of evolution

Molecular evolution does not progress in a simple, clock-like manner. Instead, rates of molecular evolution have been found to vary considerably between lineages (Moran 1996; Duffy et al. 2008; Smith and Donoghue 2008). The variation in evolutionary rates has been linked to a number of factors such as lifestyle (with parasitic taxa often evolving faster than their non-parasitic

relatives), generation time (faster generation times often correspond to faster evolutionary rates) and climate (tropical species tend to accumulate substitutions faster than species in cooler climates) (Andreasen and Baldwin 2001; Allen et al. 2006; Wright et al. 2006; Smith and Donoghue 2008; Bromham et al. 2013). Furthermore, faster rates of molecular evolution have also been linked to higher diversification rates (Pagel et al. 2006; Lanfear et al. 2010a). Identifying the underlying factors responsible for variation in evolutionary rates can therefore shed light on evolutionary processes (Lanfear et al. 2010b). This can be done in one of two ways: sister clade analyses, or whole tree methods. The latter are better developed for testing the link between substitution rates and continuously varying traits.

### Sister clade analyses

Phylogenetically independent sister-pair analyses are the simplest method for comparative analyses of evolutionary rates and is illustrated in Figure 8.



**Figure 8.** Illustration of sister clade analysis. Average branch length is measured for each clade and compared.

We select sister-clades, such as the one shown in Figure 8, with each clade associated with a variable of interest—in this case, whether the taxa have wings or not. Since, by definition, the two clades diverged from the ancestral node at the same time, they have had equal time to accumulate substitutions (Bromham n.d.; Lanfear et al. 2010b). By comparing the average branch lengths, we can infer that Clade B is evolving faster than Clade A. A simple sign test or Wilcoxon signed rank test can reveal whether faster evolutionary rates are associated with the presence of wings. The disadvantage of this

method is that it only uses a subset of the tree as information, and therefore has low statistical power.

### Correlated evolution of substitution rates and continuous traits

Sister clade analyses (or phylogenetic independent contrasts) can also be used to investigate the link between continuous traits and evolutionary rate, however whole tree methods (as the one used in the programme Coevol) are more powerful (Lartillot and Poujol 2011). In this method, the continuous trait and substitution rates are jointly modelled using Brownian motion. The strength and direction of coupling between trait values and substitution rate values is then estimated in a Bayesian MCMC framework.

# Integrating metabarcoding data in a phylogenetic framework

We have now learnt how comparative phylogenetic analyses can be used to track the evolution of traits, and examine the link between traits and evolutionary rates. But how can we incorporate metabarcoding data in a phylogenetic framework given the low phylogenetic signal, especially short-read sequences.

Long-read metabarcoding certainly overcomes this issue to a certain extent. Using the 18S and 28S together has been demonstrated to provide an even stronger phylogenetic signal which can resolve more “difficult” relationships (**Paper I**) (Marande et al. 2009; Zhao et al. 2012). (Marande et al. 2009; Zhao et al. 2012). However, even long-read metabarcoding cannot resolve deep-branching nodes, or at least not with robust support. We suggest incorporating published phylogenomic data by constraining groups such as Sar, which have conclusively been resolved as a robust group, to be monophyletic (Burki et al. 2007, 2020).

Constraining trees can certainly be quite controversial. For instance Alexis Stamatakis says on the RAxML website: “I personally have a strong dislike for constraint trees because the bias the analysis a priori using some biological knowledge that may not necessarily represent the signal coming from the data one is analyzing.” However we argue that constraining trees can be appropriate given that this is the only way of inferring biologically meaningful taxon rich phylogenies. Furthermore, phylogenetic constraints have been successfully used by several recent studies (Louca et al. 2018; Varga et al. 2019).

Additionally, we advocate using trees inferred with long-read metabarcoding data (or the full 18S sequence) as backbone phylogenies into which short-read metabarcoding data can be incorporated. This approach is used in **Paper II and III** of this thesis and in a recent study (Lewitus et al. 2018).

# Research Aims

The overarching aim of this thesis is to use metabarcoding data in a phylogenetic framework to enable inferences about eukaryotic evolution and ecology.

More specifically, we first aimed to develop a method for long-read metabarcoding of the rDNA operon targeting broad eukaryotic communities, from primer design, to read curation, phylogeny-aware taxonomic annotation, and inferring robust phylogenies of environmental data (**Paper I**).

Using the method developed in Paper I and existing short-read metabarcoding data, we aimed to investigate global rates and patterns of marine-terrestrial transitions across the eukaryotic tree of life (**Paper II**). Additionally, we inferred the ancestral habitat of major eukaryotic clades, and of the last common eukaryotic ancestor.

The third aim of this thesis was to investigate determinants of the rate of molecular evolution in protists, again using long-read and short-read metabarcoding data. Specifically, we focused on whether there was a systematic difference in the evolutionary rates of marine and terrestrial protists, and whether protists in tropical marine waters have faster evolutionary rates than protists in cooler waters (**Paper III**).

For **Paper IV** we shifted focus to highlight another use of long-read metabarcoding datasets. We aimed to identify the protist parasite of the endangered freshwater pearl mussel in Sweden, using Sanger sequencing of host tissue and imaging techniques. We further assessed the prevalence of the parasite in environmental samples using long-read metabarcoding data.

# Paper Summaries

## Paper I. Long-read metabarcoding of eukaryotes

This study focused on developing a new method for long-read metabarcoding to overcome the limitations of short-read metabarcoding: namely limited phylogenetic signal, and poor identification of highly divergent sequences. We aimed to analyse the generated data in a phylogenetic framework.

- We successfully amplified the rDNA operon (spanning ~4500 bp) from three soil samples and sequenced amplicons on the PacBio Sequel platform, generating roughly 113,000 CCS reads.
- Long-read sequencing technologies are relatively new and can have high raw error rates. We developed a curation pipeline to deal with the biases specific to these long reads and obtained 650 high quality OTUs.
- We developed a phylogeny-aware method to assign taxonomy to the 650 OTUs which showed more accuracy than similarity search based methods. Sequences could be annotated to the appropriate taxonomic rank based on their position in the phylogeny, instead of arbitrary similarity thresholds. Taxonomy was assigned based on the 18S gene alone, and then transferred to the 28S gene as they are physically connected on the same molecule.
- We inferred a robust phylogeny spanning broad eukaryotic diversity based on the combined 18S and 28S genes, allowing us to infer the evolutionary relationships between the environmental sequences themselves.
- We conclude that long-read metabarcoding allows us to analyse environmental diversity in an evolutionary context.

## Paper II. Habitat evolution across eukaryotic tree of life

This study implements the method developed in Paper I and generated long-read metabarcoding data from a range of habitats including soils, freshwater, and marine waters. We used this data in conjunction with available short-read data (V4 region) to infer the role of the salt barrier in shaping eukaryote evolution.

- We generated nearly 10 million CCS reads from different environments which we clustered into 16,821 OTUs and taxonomically annotated.
- We inferred a global 18S-28S phylogeny (with a set of constraints retrieved from phylogenomic data) which spanned all major eukaryotic groups. Consistent with previous studies, the phylogenetic distinction between marine and terrestrial (freshwater and soil) communities was apparent.
- We incorporated short-read metabarcoding data, using the long-read phylogenies as backbone constraints to infer comprehensive clade-specific phylogenies. The short-read data provided information about transitions that were “missed” by the long-read data alone.
- Rates of transitioning across the salt barrier varied across eukaryotic lineages. Fungi were found to have the fastest transition rates, and may be generalists. Within protists, golden algae and diatoms are most adept at crossing the salt barrier. We detected several hundred transition events, more than anticipated.
- Most transition events detected occurred relatively recently in evolutionary history.
- The largest eukaryotic supergroups, TSAR and Amorphea, likely arose in different habitats, and early eukaryotes likely lived in non-marine environments.

### Paper III. Rate of molecular evolution in protists

In this study we set out to investigate whether habitat (marine and terrestrial) impacted the rate of molecular evolution. Additionally, we investigated whether marine surface species had higher evolutionary rates in the tropics, which might explain the latitudinal diversity gradient recently detected in marine protists.

- We used the long-read metabarcoding dataset and sister-clade analysis to examine the impact of habitat on evolutionary rates. We detected no systematic difference between the two habitats.
- There might be a weak link between transitioning to a new habitat and accelerated rates of evolution, however the result was not statistically significant.
- We inferred phylogenies with V9 metabarcoding data from the Tara Oceans survey (constrained by backbone phylogenies of long-read sequences). Contrary to expectations, we observed faster evolutionary rates in the tropics for only some of the groups examined.
- Our results cast doubt on the generality of the kinetic energy hypothesis which links high temperatures to faster evolutionary rates and

subsequently to higher speciation rates to explain the latitudinal diversity gradient.

## Paper IV. A gregarine parasite infecting freshwater pearl mussels

The endangered freshwater pearl mussel has declining population in Sweden. We identify the protist parasite associated with several mortality events in Swedish rivers.

- Using environmental sequencing of host tissue we obtain 18S molecular data of the parasite. Phylogenetic analyses reveal it to be a gregarine (Apicomplexa), specifically related to the genus *Nematopsis* which is known to infect tadpoles.
- We describe the parasite using a combination of histology, in-situ hybridization, and electron microscopy.
- We investigate the environmental prevalence of the parasite using existing long-read and short-read metabarcoding datasets. Surprisingly, we detect the parasite in long-read datasets, revealing it to be present in Swedish lakes at low abundance, but do not detect it in V4 or V9 datasets.

## Conclusions and Future Perspectives

When I first started my PhD in 2017, there were only a few published studies (four to my knowledge) that had tested the possibility of long-read metabarcoding (Mosher et al. 2013, 2014; Schloss et al. 2016; Wagner et al. 2016), and none of them on them had benchmarked the protocol on eukaryotes, or tested the possibility of sequencing the more than the SSU gene. The general consensus at the time was that Nanopore and (mostly) PacBio represented a potentially good opportunity to overcome the limitations of short-reads, but that researchers should be wary of the high error rates and use strict curation pipelines. Since then, technology has improved rapidly, and with Sequel I we obtained curated sequences with an error rate of 0.17% (Jamy et al. 2020). Within two years, we were able to sequence a larger set of samples on the Sequel II, and the difference in through-put and quality was astounding. We obtained 100-times more reads (3 million reads/SMRT cell on Sequel II compared to 30,000 reads/SMRT cell on Sequel I), with much higher quality (highest average quality of reads on Sequel II = Q60 vs. Q40 on Sequel I). Most researchers I've come across still have the thought of "high error rates" when they words "PacBio" and "Nanopore", however this trend will gradually fade as these technologies further improve and become more cost-effective and accessible to more researchers. In order for long-read metabarcoding to be adopted more widely by the microbial ecology research community, it will be essential to further develop the semi-automatic pipelines presented in Papers I and II of this thesis into fully automatic pipelines for read curation and taxonomic annotation.

One of the main aims of this thesis was to emphasize the phylogenetic potential of these long-reads. Fittingly, one of the most exciting tasks in my PhD was getting to scroll through large phylogenies of environmental sequences (or tree gazing as I like to call it). It is incredibly exciting and humbling to see the vast number of putative novel clades, and how little eukaryotic diversity has been described. It was also extremely interesting to observe *how* samples clustered, whether it was by sample, or by habitat, or in some cases, no pattern at all. Inferring such phylogenies and performing model-based analyses on them opens the door to studying more eco-evolutionary questions (such as in Papers II and III). Here in this thesis, we focused on habitat evolution and the rates of molecular evolution. We predict that future studies will use this combination of long-read and short-read metabarcoding to study protist diversity

in an eco-evolutionary framework. Important unanswered questions include: How have diversification rates of different eukaryotic clades varied through time? How do diversification dynamics differ in marine and terrestrial habitats, and why do terrestrial habitats host a larger amount of diversity? Are biotic interactions shaped by shared evolutionary history?

In summary, the work presented in this thesis represents another step towards using metabarcoding data in eco-evolutionary studies by allowing more robust phylogenies to be inferred. And I greatly look forward to all the exciting studies on protist diversity that are sure to follow.

# Popular Science Summary

What do humans have in common with the single-celled, microscopic *Paramecium*? The answer is that they are both eukaryotes, meaning that they keep their DNA coiled inside a nucleus in their cells. It is estimated that there are 8.7 million species of eukaryotes on earth. Some of these species are organisms we are familiar with, such as animals, plants and fungi. However, the vast majority of eukaryotes like *Paramecium*, are invisible to the naked eye and can't be seen without a microscope. These microbes are collectively called protists, and include things like diatoms, green algae, red algae, amoebas, ciliates, and parasites like the malaria causing agent, *Plasmodium*. Protists are found everywhere on earth, and studying them is important because they perform many roles that keep ecosystems running. For example, they produce nearly half the oxygen on earth. They eat bacteria. They get hunted by bigger protists and by animals. They cause several important diseases (malaria is one example).

But how do we study protists if they are so small? One quick method is **metabarcoding**. This involves collecting a sample from an environment, be it pond water, soil, ocean water etc.—and then extracting DNA from all the protists living in the sample. Researchers then sequence a small part of gene found in all eukaryotes, called the 18S gene. Each species has a unique sequence of A, C, G and T's, like a fingerprint or a barcode. This genetic barcode can then be used to identify all protist species found in a sample. However with current sequencing technologies, we only obtain around 500 bases of DNA or less, which is not always enough to tell apart one species from another. In **Paper I** of this thesis, we present a new method of metabarcoding which sequences around 5000 bases of DNA, roughly 10 times longer. Using this method we are better able to identify species in a given environmental sample. What is most exciting about this method however, is that we can use the longer sequences to generate a tree of life (as depicted on the cover of this thesis). How does this work? Well, organisms more closely related to each other have more similar DNA sequences. With 5000 bases of DNA, we can estimate a more accurate tree of life. This opens an exciting avenue of research because it allows us to study how eukaryotes have evolved.

In **Paper II** of this thesis, we sequence samples from different habitats such as soils, lakes and oceans. By constructing a tree of life from all these habitats,

we can estimate how often protists have switched habitats during their evolution. Moving from the oceans to land (and vice versa) is a major evolutionary change. In animals, such habitat shifts are known to be extremely rare, with the most prominent example being the move of our fishy ancestors to land, which eventually gave rise to amphibians, reptiles, birds and mammals. Our results show that similar habitat shifts in protists are uncommon, but more frequent than previously thought. Furthermore, our results indicate that eukaryotes first arose in freshwaters, soils, or other non-marine habitats.

In **Paper III** of this thesis, we investigate a phenomenon called the “kinetic energy hypothesis”. Under this hypothesis, it is thought that organisms living in the tropics accumulate changes in their DNA at a greater speed than organisms living in cooler climates. This is because the higher temperature of the tropics is thought to cause DNA to mutate faster. However, our results show that this hypothesis cannot be generalized to all groups of organisms.

Finally, in **Paper IV** of this this thesis, we show that our method can be used to explore parasites hiding in our surroundings. In particular, we show that parasites infecting the endangered Swedish freshwater pearl mussel are related to *Plasmodium*, the malaria pathogen, and can be found in Swedish lakes.

Overall, we hope that the methods presented in this thesis will help us answer more questions about the wonderful microbes that run our world.

# Svensk sammanfattning

Vad har människor gemensamt med encelliga, mikroskopiska *Paramecium*? Svaret är att båda är eukaryoter, vilket betyder att de har sitt DNA ihoprullat inuti en kärna i deras celler. Uppskattningsvis finns det 8.7 miljoner arter eukaryoter på jorden. Några av dessa organismer är vi väl bekanta med, såsom djur, växter och svampar. Trots det är de allra flesta eukaryoter, likt *Paramecium*, osynliga utan hjälpmedel. Dessa mikrober kallas protister och inkluderar bland annat diatomeer, gröna- och röda alger, amöbor, infusionsdjur och parasiter såsom *Plasmodium*, vilken orsakar malaria. Protister finns överallt på jorden och att studera dem är viktigt eftersom de bidrar till att upprätthålla fundamentala ekosystem. Till exempel producerar protister nära hälften av allt syre på jorden. De äter även bakterier och blir själva byten till större protister och djur. Protister ligger även bakom många betydande sjukdomar varav redan nämnda malaria är ett exempel.

Men hur studerar vi protister om de är så små? En snabb metod kallas **meta-barcoding**, vilken utförs genom att man tar ett miljöprov från t.ex. jord, söt- eller saltvatten, för att sedan extrahera DNA från alla levande protister i provet. Forskare sekvenserar sedan en liten del av 18S, en gen som finns i alla eukaryoter. Varje art har en unik sekvens av A, C, G och T vilket fungerar som ett fingeravtryck eller streckkod (*barcode*). Den genetiska streckkoden kan sedan användas för att identifiera alla protistarter i ett prov. Med nuvarande sekvenseringsteknologi kan vi endast läsa av cirka 500 baser av DNA, vilket inte är tillräckligt för att särskilja arter från varandra. I **Artikel I** av denna avhandling, presenterar vi en ny metod av **metabarcoding** där vi sekvenserar 5000 baser av DNA, alltså tio gånger längre sekvenser. Genom att använda denna metod kan vi lättare identifiera arter i ett miljöprov. Vad som är mest spännande med denna metod är att vi kan rekonstruera *livets träd* (vilket schematiskt kan ses på omslaget av denna avhandling). Hur fungerar detta? Organismer som är mer släkt med varandra har mer lika DNA sekvenser. Med 5000 baser av DNA, kan vi rekonstruera en mer exakt avbildning av livets träd. Detta öppnar nya vägar för vidare forskning eftersom det möjliggör studier av eukaryoternas evolution.

I **Artikel II** av denna avhandling, sekvenserar vi prover från olika habitat såsom jord, sjöar och hav. Genom att rekonstruera släktskapsträd bland protister från olika habitat kan vi uppskatta hur ofta protister genom den

evolutionära historien har bytt habitat. Att ta sig från hav till land (och vice-versa) är en omfattande evolutionär förändring. Hos djur är sådan habitatförändringar extremt ovanliga, där det mest prominenta exemplet är flytten av våra fisklika förfäder från vatten till land, vilket sedan gav upphov till amfibier, reptiler, fåglar och däggdjur. Våra resultat visar att liknande habitat-skiften hos protister är ovanliga men förekommer mer frekvent än vad vi tidigare har trott. Dessutom indikerar våra resultat att eukaryoter uppkom först i färskvatten, jord eller dylika icke-marina habitat.

I **Artikel III** av denna avhandling undersöker vi den så kallade "kinetiska energi-hypotesen". Enligt denna hypotes ackumulerar organismer som lever i tropikerna fler förändringar i sin arvs massa per tidsenhet än organismer som lever i kyligare klimat. Anledningen spekuleras vara att högre temperaturer i tropikerna leder till att arvs massan muteras i högre takt. Vi visar dock att denna hypotes inte stämmer för alla grupper av organismer.

Avslutningsvis, i **Artikel IV** av denna uppsats visar vi att vår metod kan användas för att undersöka parasiter som gömmer sig i våra omgivningar. Framförallt, visar vi att parasiter som infekterar den hotade svenska flodpärlmusslan är besläktade med malariaparasiten *Plasmodium*, och kan påvisas i svenska sjöar.

Vi hoppas att metoderna som presenteras i denna avhandling kommer att hjälpa oss att svara på fler frågor om hur de underbara mikroberna styr vår värld.

# خلاصہ برائے پاپولر سائنس

انسانوں اور خوردبینی یک خلوی ، پیرامیسیم (paramecium) میں کیا مشترک ہے ؟ اس کا جواب ہے کہ وہ دونوں یوکاریوٹس (eukaryotes) ہیں ، یعنی وہ اپنے ڈی این اے کو اپنے خلیوں میں کسی نیوکلئس کے اندر خم زدہ رکھتے ہیں۔ ایک تخمینے کے مطابق، زمین پر یوکاریوٹس کی 87 لاکھ اقسام ہیں۔ ان میں سے کچھ انواع حیاتیات سے ہم واقف ہیں ، جیسے جانور ، پودے اور فطریات ۔ تاہم ، یوکاریوٹس کی ایک وسیع اکثریت جیسے پیرامیسیم ، سادہ آنکھ کے لئے پوشیدہ ہیں اور انہیں کسی خوردبین کے بغیر نہیں دیکھا جاسکتا ہے۔ ان مائکروبز کو اجتماعی طور پر پروٹسٹس (protists) کہا جاتا ہے ، اور ان میں ڈائٹومز ، سبز طحالب ، سرخ طحالب ، امیبا، سائلیٹ، اور ملیریا کا باعث بننے والے طفیلی عامل، پلازموڈیم، شامل ہیں ۔ پروٹسٹس کرہ ارض پر ہر جگہ پائے جاتے ہیں ، اور ان کا مطالعہ کرنا ضروری ہے کیونکہ وہ بہت سے اہم کردار ادا کرتے ہیں جو ماحولیاتی نظام کو چلاتے رہتے ہیں۔ مثلاً وہ زمین پر نصف آکسیجن پیدا کرتے ہیں۔ وہ بیکٹیریا کھاتے ہیں۔ وہ بڑے پروٹسٹس اور جانوروں کی غذا ہیں۔ وہ کئی اہم بیماریوں مثلاً ملیریا، کا باعث بنتے ہیں ۔

لیکن ہم پروٹسٹس کا مطالعہ کیسے کریں ، اگر وہ اتنے چھوٹے ہیں ؟ ایک فوری طریقہ میٹابارکوڈنگ ہے۔ اس میں کسی ایک ماحول ، چاہے یہ تالاب کا پانی ، مٹی ، سمندری پانی وغیرہ ہو ، سے نمونے کا حصول ، اور پھر نمونے میں رہنے والے تمام پروٹسٹس سے ڈی این اے نکالنا شامل ہے۔ اس کے بعد محققین، جین کا ایک چھوٹا سا حصہ ترتیب سلاسل کرتے ہیں، جو تمام یوکاریوٹس میں پائے جاتے ہیں اور S18 جین کہلاتے ہیں ۔ ہر ایک نوع میں فنر پرنت یا بار کوڈ کی ترتیب (A, C, G, Ts) کا منفرد انداز ہوتا ہے۔ اس جینیاتی بارکوڈ کا استعمال پھر نمونے میں پائی جانے والی تمام پروٹسٹ انواع کی شناخت کے لئے کیا جاسکتا ہے۔ تاہم ، حالیہ ترتیبی سلاسل ٹیکنالوجیز کے ذریعہ ، ہم صرف ڈی این اے کے تقریباً 500 یا اس سے کم قاعدے حاصل کرتے ہیں ، جو ایک نوع کو دوسری نوع سے ہر بار جدا بتانے کے لئے ، ناکافی ہوتے ہیں ۔ اس مقالہ کے قسطاس اول میں ، ہم ایک میٹابارکوڈنگ کانیا طریقہ پیش کرتے ہیں ، جس سے ڈی این اے کے تقریباً 5000 قاعدے حاصل ہوتے ہیں، تقریباً پہلے سے دس گنا طویل۔ اس طریقے کا استعمال کرتے ہوئے ، کسی دیئے گئے ماحولیاتی نمونے میں انواع کی نشاندہی کی صلاحیت خوب تر ہو گئی ہے۔ تاہم اس طریقہ کار کے بارے میں سب سے زیادہ دلچسپ بات یہ ہے کہ ہم شجر حیات کی انشاء کے لئے طویل ترتیبی سلسلہ استعمال کرسکتے ہیں (جیسا کہ اس مقالہ کی جلد پردکھایا گیا ہے)۔ یہ کیسے کام کرتا ہے ؟ بخوبی کہ، ایک دوسرے سے قریبی مربوط حیاتیات کے ڈی این اے کی ترتیبات سلاسل مشابہ ہوتی ہیں۔ ڈی این اے کے 5000 قواعد کے ساتھ ، ہم شجر حیات کا زیادہ درست اندازہ لگا سکتے ہیں۔ اس سے تحقیق کی ایک

بیجان انگیز راہ کھل گئی ہے کیونکہ اس سے ہمیں یہ مطالعہ کرنے کی اجازت ملتی ہے کہ یوکاریوٹس کیسے ارتقاء پزیر ہوئے۔

اس مقالہ کے **قرطاس** **دوئم** میں ہم مٹیوں، جھیلوں اور سمندروں جیسی مختلف زیستگاہوں سے نمونے ترتیب دیتے ہیں۔ ان تمام زیستگاہوں سے شجریات بنا کر ہم اندازہ لگا سکتے ہیں کہ ان کے ارتقاء کے دوران پروٹسٹس نے کتنی بار زیستگاہوں کو تبدیل کیا ہے۔ سمندروں سے زمین کی طرف جانا (اور برعکس) ایک بڑی ارتقائی تبدیلی ہے۔ جانوروں میں اس طرح کی زیستگاہوں کی تبدیلیاں انتہائی نایاب ہیں، اس کی سب سے نمایاں مثال ہمارے ماہی مانند اجداد کا زمین کی طرف جانا ہے، جس نے آخر کار دوزیستانوں، خزندگانوں، جانوروں، پرندوں اور ممالیہ کو جنم دیا۔ ہمارے نتائج سے پتا چلتا ہے کہ پروٹسٹس میں زیستگاہوں کی تبدیلی غیر معمولی ہے، لیکن پہلے کی سوچ سے کہیں زیادہ بکثرت۔ مزید برآں، ہمارے نتائج بتاتے ہیں کہ یوکاریوٹس سب سے پہلے تازہ پانیوں، مٹیوں یا دوسرے غیر سمندری زیستگاہوں میں پیدا ہوئے۔

اس مقالہ کے **قرطاس** **سوئم** میں، ہم "متحرک توانائی مفروضہ" نامی ایک امر کی تحقیق کرتے ہیں۔ اس مفروضے کے تحت، یہ تصور کیا جاتا ہے کہ استوائی خطہ میں رہنے والے حیاتیات ٹھنڈے آب و ہوا میں رہنے والے حیاتیات کی نسبت زیادہ تیزی سے اپنے ڈی این اے میں تبدیلیاں جمع کرتے ہیں خیال کیا جاتا ہے کہ استوائی خطہ کا نسبتاً بلند درجہ حرارت ڈی این اے کو تیزی سے بدلنے کا باعث بنتا ہے۔ تاہم، ہمارے نتائج بتاتے ہیں کہ حیاتیات کے تمام گروہوں میں اس رجحان عمل کو عام نہیں کہا جاسکتا ہے۔

آخر میں، اس مقالہ کے **قرطاس** **چہارم** میں، ہم یہ دیکھاتے ہیں کہ ہمارے طریقہ کار کو اپنے گردونواح میں روپوش طفیلیوں کو دریافت کرنے کے لئے استعمال کیا جاسکتا ہے۔ بالخصوص، ہم یہ ظاہر کرتے ہیں کہ معدومیت سے پرخطر، سویڈش میٹھے پانی والی صدف کو مصائب زدہ کرنے والے طفیلی، پلازموڈیم سے تعلق رکھتے ہیں، جو ملیریا پیتھوجن ہے اور سویڈش جھیلوں میں پایا جاسکتا ہے۔

مجموعی طور پر، ہم امید کرتے ہیں کہ اس مقالہ میں پیش کیے گئے طریقے ہماری دنیا کو چلا نے والے حیرت انگیز مائکروبز کے بارے میں مزید سوالات کے جوابات دینے میں ہمارے مددگار ہوں گے۔

مہوش جامی

# Acknowledgements

During my PhD studies, I've been extremely fortunate to have the support and company of so many fantastic mentors, friends and family members. This thesis would not have been possible without you!

First and foremost, I would like to thank my supervisor and mentor Fabien Burki. Four years ago, I don't think that either of us thought that our work together would wade into newer territories in environmental sequencing! What started as a little side project evolved into several years of work, and I am so grateful that you gave me so much scientific freedom. Thank you for all the really enjoyable meetings, for encouraging me to become more independent, for helping me to formulate ideas more clearly, and for all your support in the more bumpy parts of the ride. It has truly been a pleasure being part of your group!

During my time in Fabien's group, I've also had the opportunity to spend time with a really amazing set of lab members. (I do hope we get to have our Thursday pancake lunches back soon — I have missed them over the past year). Jürgen and Iker, you were the first people to welcome me into the group, and I have really enjoyed your company. Iker thank you for being my phylogenetics guru over the years. I've learnt so much from you! (I cringe now when I hear the word "basal" or "higher organisms" so I think your job here is done ;) ). Thank you also for the delicious food and fikas! P.S. I'm writing this acknowledgements section while sitting on the couch that I inherited from you ☺. Jürgen, it's been really fun hanging out with you! I hope I opened your mind a little bit to fantasy book and movies (or maybe I turned you off even more haha). I really admire how thorough you are in your research. We were soon joined in the lab by Vasily. Vasily, I honestly think you are a bit of a magician behind the microscope. Thank you for several lovely afternoons peering down at tardigrades, ciliates, telonemids, and of course centrohelids! Ioana you joined the lab and brought your infectious enthusiasm for biology (and lichens!) with you. And it was great to finally have another person in the lab who understood Harry Potter references ;) . You made lab meetings really fun by engaging with everyone's work and I really admire your creativity. Thank you for stopping by my office so frequently in the last month, and also for your help with Paper IV. Elena, though you were only in the lab for a few months, you made a big impact! Thank you for your work on optimizing the

long-range PCRs. Anders, I'm amazed at your knowledge about oysters, and mussels, and fish, and at your gardening skills. Good luck with the rest of your PhD! Max, I'm happy that we got to host you in our lab. Thanks for introducing us to Slack, for looking after Noisy and for general advice on bioinformatics problems whenever I needed help. Megan and Caesar, it's also been really lovely having you both in the group ☺. Charlie! Omg I loved working with you! It was really fun to have someone to bounce ideas with and also share excitement with when we found something cool (and let's be honest, we found lots of cool things in our project together ☺). Paper II would not have been what it is without you! Finally, Miguel you brought your expertise of ecology and metabarcoding when you joined the lab. Thank you for hosting a really useful workshop on networks and being incredibly patient as I continuously questioned why anyone would want to use something other than phylogenies. At the end you even managed to convince me.

I've also been lucky to work with great collaborators over the years. First, David Bass, thanks to you and Rachel for generating the first batch of data for our project together. Second, I would like to thank members of the Stamatakis lab: Pierre, Lucas, Alexey and Alexis. Your work on the phylogeny-aware taxonomic annotation was an invaluable addition to Paper I. Pierre and Lucas, I'm particularly grateful that you always responded so fast anytime I had a question about phylogenetic placement. We were also lucky to have Daniel Vaultot as a collaborator. PR<sup>2</sup> is such a treasure for the protist community, and I can't wait for metaPR<sup>2</sup> to be released. Thank you for your huge contribution to the transitions project. I would also like to think the people who provided us with the coolest samples to sequence, from soils in Puerto Rico to the depths of the Mariana Trench. Thanks to Anna Rosling, Sari Peura, and Hongmei Jing for sharing these samples with us. I would like to thank Olga Pettersson for taking a chance on sequencing our long-amplicons on the new Sequel II, and NGI Uppsala in general for providing sequencing facilities. Finally, Jesper, you weren't officially a collaborator, but I feel your name should be included in this list here because of how much you engaged with my work. Thank you for all the discussions and for getting me to think a bit more from a pop-gen perspective. Thank you also for helping me out with awk so often! I still don't understand the syntax (is there even any logic to it?) but you are an awk king and make it look so easy.

SystBio! What a wonderful environment to work in! Hanna (and previously, Martin and Sandie), thank you for creating such a warm and welcoming atmosphere for PhD students.

Martin, I've really enjoyed developing the Evolutionary Patterns course with you. Thank you for taking the time to demystify and explain the fungal tree of life to me. The same for Aaron and Jenni! You three have been my go-to fungal people anytime I had questions. Aaron, we started in SystBio at the same time (give or take a few months). Thank you for pancakes, and maple syrup, and all round great company! Jenni, you're incredibly witty and funny. Thank you for checking in on me when I was close to submitting, and for cheering me on ("Halfway and one more step"). Diem, I admire your super-human organisational skills. I really appreciate that you took the time on several occasions to help me with my presentations, and then also giving me feedback every time I presented. Ivain, it was really nice having you around, and you quickly identified all of us at SystBio as being "taxonomically-sensitive" haha. Mikael and Petra, I remember you both as my favourite teachers when I was a Masters student in 2014, and I'm sure it's no coincidence that I continued working with phylogenetic trees. Nahid, thank you for your kindness and for taking care of us. Alex, it is always a pleasure to talk to you. I'd like to thank the rest of the staff at SystBio: Anushree, Octavio, Mats, David, Magnus, Christoffer, Sanja, Leif, and Inga for their presence in the corridor.

A massive thank you to all the PhD students (past and present) at SystBio for sharing this journey with me. It's been so so lovely celebrating each other's successes with fikas and with great spexes over the years. Lore, I still remember getting a hug from you on my first day here—and I think that set the tone for our friendship moving forward. Markus, it's been great sharing an office with you (the most zen office in the corridor). Thanks for putting up with my sprawling mess as I took over the board, and the couch with my stuff. Saneaaa, it was so nice to have someone else from Pakistan in the same corridor. Thanks for all the laughs and conversations and gifs. Brendan, you're so incredibly creative! Thank you for bringing in amazing pies, and for discussions on metabarcoding. Sarina, I had a lot of fun at your art-ies and all the other events you organized. Raquel, your zoom writing sessions helped me trudge through writing on bad days. On that note, thank you Vale for the shut-up-and-write sessions. Anneli, it is always fun to talk to you! Thanks also to Juma, Stella, Faheema, (big) Jesper, as well as Ivar and Petter (aka new generation). It's been a pleasure working with you all!

I would also like to acknowledge the lab of Thijs Ettema where I first fell in love with protists. Thanks to Thijs for taking me in as a Masters student (and then encouraging me to apply for the position with Fabien). I must also thank Henning for being an amaazing supervisor (and pastor of the protistan church). I had an incredible amount of fun sampling protists with you, doing single cell surgery on *Stentor* (!) and then photographing the cell fragments. I was intimidated by the command line when I first started, but you taught me to pipe

commands like a pro. Courtney, you were also incredibly helpful anytime I had questions. Jennah, I've always found your work to be so inspiring. Thanks for the dinners, mocktails and the nights in Cesky Krumlov. Joran, thank you for the discussions and help with the curation pipeline for long-read metabarcoding.

I am extremely grateful to all my friends who made my PhD studies a much more enjoyable experience, and for being there during more discouraging times.

I am greatly indebted to Jesper. Thanks for pulling me away from the computer for table tennis breaks (even though I mostly got slaughtered), for chocolate muffins and Pucko to fuel my writing, for chasing butterflies (maybe someday we'll finally upload them on Artportalen), the great grey owl, and goose. Thanks also for translating the science summary to Swedish!

Lore and Mercè, I am SO glad I shared this experience with you both! I can't even begin to imagine what these four years would have been like without our evenings at Ai Japansk, Årummet and at EBC. Lore, I think we felt an instant connection during the MEME summer school in 2016. So much so that I remember we used to joke that you were half-Aaron, half-me. Thank you for the laughs, hugs, dancing (remember Tommy at the Korean expo?), and for listening and providing perspective when I was down.. Mercè, likewise, you were there from the very start. Thank you for all the lovely conversations, for widening my views, your honesty, hugs, and the head-tilt selfies (of course xD). I am constantly amazed at your energy! You've made EBC a much more vibrant place for me and for everyone else with beervolutions, GoT evenings, and all the other events.

MEME friends! You will always hold a very special place in my heart and I am so happy that we've managed to keep in touch despite being on different continents, and in different time zones ("If that makes sense?"). Anya, I was always comforted by the fact that you were relatively close by. Thank you for hosting me at your place, it was one of the highlights of my first year in Sweden <3. Khalis and Étienne, it still feels surreal that we were actually able to meet up, and then cook and hang out as if no time had passed. Thank you, all three of you, for the postcards, and for cheering me on from a distance. Thanks also to Ülkü who hosted me for a very special weekend in Oxford.

Manolis, we started this together, and we are finishing this together! Thank you for being my partner in figuring-out-logistics-of-thesis-submission, and for checking in on me the last few months. Karin, I'm sure you've heard this before, but your enthusiasm is so infectious! Caro, thank you for feeding me

lomo saltado (I still can't make it as good as you). I've really enjoyed our Cesky Krumlov dinner nights (along with Jennah). Of course, I must also mention the "EBC Randoms", a rag-tag group of friends. Alex, Willian, TJ, Bere, Fede, Car, Lore, Mercè, André, Fotini, and Moos, thank you for your company on many occasions. (Moos and Lore, I hope someday you will forgive me for spoiling Avengers Infinity War five minutes before the movie started). I would also like to thank all the other PhD students at EBC over the years who've made my workplace so much fun: Luciana, Laura, Madee, Venkat, Ghazal, Bianca and Kevin. To my students in Evolutionary Patterns (all four cohorts), teaching you, and getting to know each other was a very special experience. I ended up learning a lot during our classes!

Tack Emy and Johan for the fun weekend walks, and otter sightings, and teaching me a thing or two about plants.

Haleemah and Anum, I'm so glad we've kept in touch! Thank you for all the pep-talks at the right time.

Sara, I sincerely thank you for the kindness you showed me when we were flat mates.

Ulla, Stig, Josefine, and Oscar, thank you for all the **great** food, company, and for making me feel at home in Sweden.

Last, but not least, I want to sincerely thank my family. Being so far away from home is not always easy, so I'd like to thank all my cousins, and especially Sobia, Rabia, Maaheen, Misha and Meher for the Eid Zoom calls. Sobia in particular, you have always been one text message away. To my Nana, Nano, and Dadi, thank you for all the love, indulging me when I visit and for your duas.

I cannot even begin to express my gratitude to the Aamirs (Aamir mamu, Nighat mami, Faraaz, Shehryaar and Qasim) for taking me into your family in Australia (more than 10 years ago now!). Thank you for three wonderful years—I wouldn't be here in Sweden if it wasn't for you.

Tahir, thank you for sooo many things: for making this thesis cover (aur uff kya cover banaya hai yaar :p), for helping me with all the little things near the end of thesis submission, for your messages, and for watching all four seasons of Avatar the Last Airbender with me online.

And most importantly, my parents, Nyma and Naveed. Amma, thank you for all the duas and nafals (and counting the pages haha), and Abbu for translating the popular science summary in Urdu. Words will never be enough but, thank you, for always supporting me through thick and through thin, and for helping me grow into the person that I am today. Thank you for everything.

# References

- Ackerly, D. D. 2003. Community assembly, niche conservatism, and adaptive evolution in changing environments. *Int. J. Plant Sci.* 164.
- Allen, A. P., J. F. Gillooly, V. M. Savage, and J. H. Brown. 2006. Kinetic effects of temperature on rates of genetic divergence and speciation. *Proc. Natl. Acad. Sci.* 103:9130–9135.
- Amaral-Zettler, L. A., E. A. McCliment, H. W. Ducklow, and S. M. Huse. 2009. A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA Genes. *PLoS One* 4:e6372.
- Andreasen, K., and B. G. Baldwin. 2001. Unequal evolutionary rates between annual and perennial lineages of checker mallows (*Sidalcea*, Malvaceae): Evidence from 18S-26S rDNA internal and external transcribed spacers. *Mol. Biol. Evol.* 18:936–944.
- Annenkova, N. V., C. R. Giner, and R. Logares. 2020. Tracing the Origin of Planktonic Protists in an Ancient Lake. *Microorganisms* 8:543.
- Arroyo, A. S., R. Iannes, E. Baptiste, and I. Ruiz-Trillo. 2020. Gene similarity networks unveil a potential novel unicellular group closely related to animals from the tara oceans expedition. *Genome Biol. Evol.* 12:1664–1678.
- Baldauf, S. L., A. J. Roger, I. Wenk-Siefert, and W. F. Doolittle. 2000. A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science* (80-. ). 290:972–977.
- Baloğlu, B., Z. Chen, V. Elbrecht, T. Braukmann, S. MacDonald, and D. Steinke. 2021. A workflow for accurate metabarcoding using nanopore MinION sequencing. *Methods Ecol. Evol.* 12:794–804.
- Barbera, P., L. Czech, S. Lutterop, and A. Stamatakis. 2020. SCRAPP: A tool to assess the diversity of microbial samples from phylogenetic placements Running Title: SCRAPP. , doi: 10.1101/2020.02.28.969980.
- Barbera, P., A. M. Kozlov, L. Czech, B. Morel, D. Darriba, T. Flouri, and A. Stamatakis. 2019. EPA-ng: massively parallel evolutionary placement of genetic sequences. *Syst. Biol.* 68:365–369.
- Bass, D., G. M. Ward, and F. Burki. 2019. Ascetosporea. *Curr. Biol.* 29:R7–R8.
- Bates, S. T., J. C. Clemente, G. E. Flores, W. A. Walters, L. W. Parfrey, R. Knight, and N. Fierer. 2013. Global biogeography of highly diverse protistan communities in soil. *ISME J.* 7:652–659.
- Becking, B. 1934. *Geobiologie of inleiding tot de milieukunde* .
- Berger, S. A., D. Krompass, and A. Stamatakis. 2011. Performance, Accuracy, and Web Server for Evolutionary Placement of Short Sequence Reads under Maximum Likelihood. *Syst. Biol.* 60:291–302.
- Bininda-Emonds, O. R., S. G. Brady, J. Kim, and M. J. Sanderson. 2001. Scaling of accuracy in extremely large phylogenetic trees. *Pac. Symp. Biocomput.* 547–558.

- Bochdansky, A. B., M. A. Clouse, and G. J. Herndl. 2017. Eukaryotic microbes, principally fungi and labyrinthulomycetes, dominate biomass on bathypelagic marine snow. *ISME J.* 11:362–373.
- Boenigk, J., S. Wodniok, C. Bock, D. Beisser, C. Hempel, L. Grossmann, A. Lange, and M. Jensen. 2018. Geographic distance and mountain ranges structure freshwater protist communities on a european scale. *Metabarcoding and Metagenomics* 2:e21519.
- Branton, D., D. W. Deamer, A. Marziali, H. Bayley, S. A. Benner, T. Butler, M. Di Ventra, S. Garaj, A. Hibbs, X. Huang, S. B. Jovanovich, P. S. Krstic, S. Lindsay, X. S. Ling, C. H. Mastrangelo, A. Meller, J. S. Oliver, Y. V. Pershin, J. M. Ramsey, R. Riehn, G. V. Soni, V. Tabard-Cossa, M. Wanunu, M. Wiggin, and J. A. Schloss. 2008. The potential and challenges of nanopore sequencing.
- Bromham, L. n.d. Substitution Rate Analysis and Molecular Evolution.
- Bromham, L., P. F. Cowman, and R. Lanfear. 2013. Parasitic plants have increased rates of molecular evolution across all three genomes. *BMC Evol. Biol.* 13:126.
- Brown, M. W., A. A. Heiss, R. Kamikawa, Y. Inagaki, A. Yabuki, A. K. Tice, T. Shiratori, K.-I. Ishida, T. Hashimoto, A. G. B. Simpson, and A. J. Roger. 2018. Phylogenomics Places Orphan Protistan Lineages in a Novel Eukaryotic Super-Group. *Genome Biol. Evol.* 10:427–433.
- Burki, F. 2014. The Eukaryotic Tree of Life from a Global Phylogenomic Perspective. *Cold Spring Harb. Perspect. Biol.* 6:a016147–a016147.
- Burki, F., A. J. Roger, M. W. Brown, and A. G. B. Simpson. 2020. The New Tree of Eukaryotes.
- Burki, F., K. Shalchian-Tabrizi, M. Minge, Å. Skjæveland, S. I. Nikolaev, K. S. Jakobsen, and J. Pawlowski. 2007. Phylogenomics reshuffles the eukaryotic supergroups. *PLoS One* 2:e790.
- Callahan, B. J., D. Grinevich, S. Thakur, M. A. Balamotis, and T. Ben Yehezkel. 2021. Ultra-accurate microbial amplicon sequencing with synthetic long reads. *Microbiome* 9:130.
- Caporaso, J. G., J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Pêa, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunencko, J. Zaneveld, and R. Knight. 2010. QIIME allows analysis of high-throughput community sequencing data.
- Caron, D. A., and S. K. Hu. 2019. Are We Overestimating Protistan Diversity in Nature?
- Carr, M., D. J. Richter, P. Fozouni, T. J. Smith, A. Jeuck, B. S. C. Leadbeater, and F. Nitsche. 2017. A six-gene phylogeny provides new insights into choanoflagellate evolution. *Mol. Phylogenet. Evol.* 107.
- Cavender-Bares, J., K. H. Kozak, P. V. A. Fine, and S. W. Kembel. 2009. The merging of community ecology and phylogenetic biology. *Ecol. Lett.* 12:693–715.
- Chambouvet, A., A. Monier, F. Maguire, S. Itoiz, J. del Campo, P. Elies, B. Edvardsen, W. Eikreim, and T. A. Richards. 2019. Intracellular Infection of Diverse Diatoms by an Evolutionary Distinct Relative of the Fungi. *Curr. Biol.* 29:4093–4101.e4.
- Chung, N., M. W. Van Goethem, M. A. Preston, F. Lhota, L. Cerna, F. Garcia-Pichel, V. Fernandes, A. Giraldo-Silva, H. S. Kim, E. Hurowitz, M. Balamotis, I. Wu, and T. Ben-Yehezkel. 2020. Accurate Microbiome Sequencing with Synthetic Long Read Sequencing. *bioRxiv* 2020.10.02.324038.

- Collins, T. M., P. H. Wimberger, and G. J. P. Naylor. 1994. Compositional Bias, Character-State Bias, and Character-State Reconstruction Using Parsimony. *Syst. Biol.* 43:482.
- Corliss, J. O. 2004. Why the World Needs Protists! *J. Eukaryot. Microbiol.* 51:8–22.
- Creer, S., K. Deiner, S. Frey, D. Porazinska, P. Taberlet, W. K. Thomas, C. Potter, and H. M. Bik. 2016. The ecologist's field guide to sequence-based identification of biodiversity.
- Czech, L., P. Barbera, and A. Stamatakis. 2020. Genesis and Gappa: Processing, analyzing and visualizing phylogenetic (placement) data. *Bioinformatics* 36:3263–3265.
- De Luca, D., R. Piredda, D. Sarno, and W. H. C. F. Kooistra. 2021. Resolving cryptic species complexes in marine protists: phylogenetic haplotype networks meet global DNA metabarcoding datasets. *ISME J.* 1–12.
- de Vargas, C., S. Audic, N. Henry, J. Decelle, F. Mahé, R. Logares, E. Lara, C. Berney, N. Le Bescot, I. Probert, M. Carmichael, J. Poulain, S. Romac, S. Colin, J.-M. Aury, L. Bittner, S. Chaffron, M. Dunthorn, S. Engelen, O. Flegontova, L. Guidi, A. Horák, O. Jaillon, G. Lima-Mendez, J. Lukeš, S. Malviya, R. Morard, M. Mulot, E. Scalco, R. Siano, F. Vincent, A. Zingone, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, T. O. Tara Oceans Coordinators, S. G. Acinas, P. Bork, C. Bowler, G. Gorsky, N. Grimsley, P. Hingamp, D. Iudicone, F. Not, H. Ogata, S. Pesant, J. Raes, M. E. Sieracki, S. Speich, L. Stemann, S. Sunagawa, J. Weissenbach, P. Wincker, E. Karsenti, C. Vargas, S. Audic, N. Henry, J. Decelle, F. Mahé, R. Logares, E. Lara, C. Berney, N. Bescot, I. Probert, M. Carmichael, J. Poulain, and S. Romac. 2015. Eukaryotic plankton diversity in the sunlit ocean. *Science* (80-. ). 348:1261605.
- Del Campo, J., M. E. Sieracki, R. Molestina, P. Keeling, R. Massana, and I. Ruiz-Trillo. 2014. The others: Our biased perspective of eukaryotic genomes.
- Delmont, T. O., M. Gaia, D. D. Hinsinger, P. Fremont, A. F. Guerra, A. M. Eren, C. Vanni, A. Kourlaiev, L. d'Agata, Q. Clayssen, E. Villar, K. Labadie, C. Cruaud, J. Poulain, C. da Silva, M. Wessner, B. Noel, J. M. Aury, C. de Vargas, C. Bowler, E. Karsenti, E. Pelletier, P. Wincker, O. Jaillon, S. Sunagawa, S. G. Acinas, P. Bork, E. Karsenti, C. Bowler, C. Sardet, L. Stemann, C. de Vargas, P. Wincker, M. Lescot, M. Babin, G. Gorsky, N. Grimsley, L. Guidi, P. Hingamp, O. Jaillon, S. Kandels, D. Iudicone, H. Ogata, S. Pesant, M. B. Sullivan, F. Not, L. Karp-Boss, E. Boss, G. Cochrane, M. Follows, N. Poulton, J. Raes, M. Sieracki, and S. Speich. 2020. Functional repertoire convergence of distantly related eukaryotic plankton lineages revealed by genome-resolved metagenomics.
- Derelle, R., and B. F. Lang. 2012. Rooting the eukaryotic tree with mitochondrial and bacterial proteins. *Mol. Biol. Evol.* 29:1277–1289.
- Derelle, R., G. Torruella, V. Klimeš, H. Brinkmann, E. Kim, Č. Vlček, B. F. Lang, and M. Eliáš. 2015. Bacterial proteins pinpoint a single eukaryotic root. *Proc. Natl. Acad. Sci.* 112:E693–E699.
- Dobell, C., and A. Van Leeuwenhoek. 1932. Antony van Leeuwenhoek and his “Little animals”; Being Some Account of the Father of Protozoology and Bacteriology and His Multifarious Discoveries in These Disciplines.
- Duarte, C. M. 2015. Seafaring in the 21st century: The Malaspina 2010 circumnavigation expedition.
- Duffy, S., L. A. Shackelton, and E. C. Holmes. 2008. Rates of evolutionary change in viruses: Patterns and determinants.

- Dunthorn, M., J. Otto, S. A. Berger, A. Stamatakis, F. Mahé, S. Romac, C. De Vargas, S. Audic, B. Consortium, A. Stock, F. Kauff, and T. Stoeck. 2014a. Placing Environmental Next-Generation Sequencing Amplicons from Microbial Eukaryotes into a Phylogenetic Context. *Mol. Biol. Evol.* 31:993–1009.
- Dunthorn, M., T. Stoeck, J. Clamp, A. Warren, and F. Mahé. 2014b. Ciliates and the rare biosphere: A review. Pp. 404–409 in *Journal of Eukaryotic Microbiology*.
- Eric Ma E, F., C. De Vargas, D. Bass, L. Czech, A. Stamatakis, E. Lara, D. Singer, J. Mayor, J. Bunge, S. Sernaker, T. Siemensmeyer, I. Trautmann, S. Romac, E. Berney, A. Kozlov, E. A. D. Mitchell, C. V. W. Seppey, E. Egge, G. Lentendu, R. Wirth, G. Trueba, and M. Dunthorn. n.d. Soil Protists in Three Neotropical Rainforests are Hyperdiverse and Dominated by Parasites. , doi: 10.1101/050997.
- Elbrecht, V., and F. Leese. 2015. Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass-sequence relationships with an innovative metabarcoding protocol. *PLoS One* 10:e0130324.
- Flegontova, O., P. Flegontov, S. Malviya, S. Audic, P. Wincker, C. de Vargas, C. Bowler, J. Lukeš, and A. Horák. 2016. Extreme Diversity of Diplonemid Eukaryotes in the Ocean. *Curr. Biol.* 26:3060–3065.
- Frøsløv, T. G., R. Kjølner, H. H. Bruun, R. Ejrnæs, A. K. Brunbjerg, C. Pietroni, and A. J. Hansen. 2017. Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. *Nat. Commun.* 8:1–11.
- Gaonkar, C. C., W. H. C. F. Kooistra, C. B. Lange, M. Montresor, and D. Sarno. 2017. Two new species in the *Chaetoceros socialis* complex (Bacillariophyta): *C. sporotruncatus* and *C. dichatoensis*, and characterization of its relatives, *C. radicans* and *C. cinctus*. *J. Phycol.* 53:889–907.
- Gawryluk, R. M. R., D. V. Tikhonenkov, E. Hehenberger, F. Husnik, A. P. Mylnikov, and P. J. Keeling. 2019. Non-photosynthetic predators are sister to red algae.
- Giner, C. R., V. Balagué, A. K. Krabberød, I. Ferrera, A. Reñé, E. Garcés, J. M. Gasol, R. Logares, and R. Massana. 2019. Quantifying long-term recurrence in planktonic microbial eukaryotes. *Mol. Ecol.* 28:923–935.
- Gottschling, M., L. Czech, F. Mahé, S. Adl, and M. Dunthorn. 2020. The windblown: possible explanations for dinophyte DNA in forest soils. *bioRxiv.org* 2020.08.07.242388.
- Guillou, L., D. Bachar, S. Audic, D. Bass, C. Berney, L. Bittner, C. Boutte, G. Burgaud, C. de Vargas, J. Decelle, J. del Campo, J. R. Dolan, M. Dunthorn, B. Edvardsen, M. Holzmann, W. H. C. F. Kooistra, E. Lara, N. Le Bescot, R. Logares, F. Mahé, R. Massana, M. Montresor, R. Morard, F. Not, J. Pawlowski, I. Probert, A.-L. Sauvadet, R. Siano, T. Stoeck, D. Vaultot, P. Zimmermann, and R. Christen. 2012. The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Res.* 41:D597–D604.
- Guillou, L., M. Viprey, A. Chambouvet, R. M. Welsh, A. R. Kirkham, R. Massana, D. J. Scanlan, and A. Z. Worden. 2008. Widespread occurrence and genetic diversity of marine parasitoids belonging to *Syndiniales* (*Alveolata*). *Environ. Microbiol.* 10:3349–3365.
- Hagino, K., and J. R. Young. 2015. Biology and paleontology of coccolithophores (Haptophytes). Pp. 311–330 in *Marine Protists: Diversity and Dynamics*.
- Harrison, J. B., J. M. Sunday, and S. M. Rogers. 2019. Predicting the fate of eDNA in the environment and implications for studying biodiversity.

- Hartikainen, H., O. S. Ashford, C. Berney, B. Okamura, S. W. Feist, C. Baker-Austin, G. D. Stentiford, and D. Bass. 2014. Lineage-specific molecular probing reveals novel diversity and ecological partitioning of haplosporidians. *ISME J.* 8:177–186.
- He, D., O. Fiz-Palacios, C.-J. Fu, J. Fehling, C.-C. Tsai, and S. L. Baldauf. 2014. An Alternative Root for the Eukaryote Tree of Life. *Curr. Biol.* 24:465–470.
- Hebert, P. D. N., A. Cywinska, S. L. Ball, and J. R. DeWaard. 2003. Biological identifications through DNA barcodes. *Proc. R. Soc. B Biol. Sci.* 270:313–321.
- Heeger, F., E. C. Bourne, C. Baschien, A. Yurkov, B. Bunk, C. Spröer, J. Overmann, C. J. Mazzoni, and M. T. Monaghan. 2018. Long-read DNA metabarcoding of ribosomal RNA in the analysis of fungi from aquatic environments. *Mol. Ecol. Resour.* 18:1500–1514.
- Hernández, C. E., E. Rodríguez-Serrano, J. Avaria-Llautureo, O. Inostroza-Michael, B. Morales-Pallero, D. Boric-Bargetto, C. B. Canales-Aguirre, P. A. Marquet, and A. Meade. 2013. Using phylogenetic information and the comparative method to evaluate hypotheses in macroecology. *Methods Ecol. Evol.* 4:401–415.
- Hillebrand, H. 2004. On the generality of the latitudinal diversity gradient.
- Hillis, D. M., and M. T. Dixon. 1991. Ribosomal DNA: Molecular evolution and phylogenetic inference. *Q. Rev. Biol.* 66:411–446.
- Hugerth, L. W., and A. F. Andersson. 2017. Analysing Microbial Community Composition through Amplicon Sequencing: From Sampling to Hypothesis Testing. *Front. Microbiol.* 8:1561.
- Hutchinson, G. E. 1965. The Ecological Theater and the Evolutionary Play.
- Ibarbalz, F. M., N. Henry, M. C. Brandão, S. Martini, G. Busseni, H. Byrne, L. P. Coelho, H. Endo, J. M. Gasol, A. C. Gregory, F. Mahé, J. Rigonato, M. Royo-Llonch, G. Salazar, I. Sanz-Sáez, E. Scalco, D. Soviadan, A. A. Zayed, A. Zingone, K. Labadie, J. Ferland, C. Marec, S. Kandels, M. Picheral, C. Dimier, J. Poulain, S. Pisarev, M. Carmichael, S. Pesant, S. G. Acinas, M. Babin, P. Bork, E. Boss, C. Bowler, G. Cochrane, C. de Vargas, M. Follows, G. Gorsky, N. Grimsley, L. Guidi, P. Hingamp, D. Iudicone, O. Jaillon, L. Karp-Boss, E. Karsenti, F. Not, H. Ogata, N. Poulton, J. Raes, C. Sardet, S. Speich, L. Stemmann, M. B. Sullivan, S. Sunagawa, P. Wincker, E. Pelletier, L. Bopp, F. Lombard, and L. Zinger. 2019. Global Trends in Marine Plankton Diversity across Kingdoms of Life. *Cell* 179:1084-1097.e21.
- Jamy, M., R. Foster, P. Barbera, L. Czech, A. Kozlov, A. Stamatakis, G. Bending, S. Hilton, D. Bass, and F. Burki. 2020. Long-read metabarcoding of the eukaryotic rDNA operon to phylogenetically and taxonomically resolve environmental diversity. *Mol. Ecol. Resour.* 20.
- Jeong, J., K. Yun, S. Mun, W. H. Chung, S. Y. Choi, Y. do Nam, M. Y. Lim, C. P. Hong, C. H. Park, Y. Ahn, and K. Han. 2021. The effect of taxonomic classification by full-length 16S rRNA sequencing with a synthetic long-read technology. *Sci. Rep.* 11:1727.
- Jewari, C. Al, and S. L. Baldauf. 2021. The impact of incongruence and exogenous gene fragments on estimates of the eukaryote root. *bioRxiv* 2021.04.08.438903.
- Khomich, M., H. Kauserud, R. Logares, S. Rasconi, and T. Andersen. 2017. Planktonic protistan communities in lakes along a large-scale environmental gradient. *FEMS Microbiol. Ecol.* 93:231.

- Kolaříková, Z., R. Slavíková, C. Krüger, M. Krüger, and P. Kohout. 2021. PacBio sequencing of Glomeromycota rDNA: a novel amplicon covering all widely used ribosomal barcoding regions and its applicability in taxonomy and ecology of arbuscular mycorrhizal fungi. *New Phytol.* 231:490–499.
- Koonin, E. V. 2010. The origin and early evolution of eukaryotes in the light of phylogenomics. *Genome Biol.* 11:209.
- Kopf, A., M. Bicak, R. Kottmann, J. Schnetzer, I. Kostadinov, K. Lehmann, A. Fernandez-Guerra, C. Jeanthon, E. Rahav, M. Ullrich, A. Wichels, G. Gerdt, P. Polymenakou, G. Kotoulas, R. Siam, R. Z. Abdallah, E. C. Sonnenschein, T. Cariou, F. O’Gara, S. Jackson, S. Orlic, M. Steinke, J. Busch, B. Duarte, I. Caçador, J. Canning-Clode, O. Bobrova, V. Marteinsen, E. Reynisson, C. M. Loureiro, G. M. Luna, G. M. Quero, C. R. Löscher, A. Kremp, M. E. DeLorenzo, L. Øvreås, J. Tolman, J. LaRoche, A. Penna, M. Frischer, T. Davis, B. Katherine, C. P. Meyer, S. Ramos, C. Magalhães, F. Jude-Lemeilleur, M. L. Aguirre-Macedo, S. Wang, N. Poulton, S. Jones, R. Collin, J. A. Fuhrman, P. Conan, C. Alonso, N. Stambler, K. Goodwin, M. M. Yakimov, F. Baltar, L. Bodrossy, J. Van De Kamp, D. M. F. Frampton, M. Ostrowski, P. Van Ruth, P. Malthouse, S. Claus, K. Deneudt, J. Mortelmans, S. Pitois, D. Wallom, I. Salter, R. Costa, D. C. Schroeder, M. M. Kandil, V. Amaral, F. Biancalana, R. Santana, M. L. Pedrotti, T. Yoshida, H. Ogata, T. Ingleton, K. Munnik, N. Rodriguez-Ezpeleta, V. Berteaux-Lecellier, P. Wecker, I. Cancio, D. Vault, C. Bienhold, H. Ghazal, B. Chaouni, S. Essayeh, S. Ettamimi, E. H. Zaid, N. Boukhatem, A. Bouali, R. Chahboune, S. Barrijal, M. Timinouni, F. El Otmani, M. Bennani, M. Mea, N. Todorova, V. Karamfilov, P. Ten Hoopen, G. Cochrane, S. L’Haridon, K. Can Bizsel, A. Vezzi, F. M. Lauro, P. Martin, R. M. Jensen, J. Hinks, S. Gebbels, R. Rosselli, F. De Pascale, R. Schiavon, A. Dos Santos, E. Villar, S. Pesant, B. Cataletto, F. Malfatti, R. Edirisinghe, J. A. Herrera Silveira, M. Barbier, V. Turk, T. Tinta, W. J. Fuller, I. Salihoglu, N. Serakinci, M. C. Ergoren, E. Bresnan, J. Iriberry, P. A. F. Nyhus, E. Bente, H. E. Karlsen, P. N. Golyshin, J. M. Gasol, S. Moncheva, N. Dzhenbekova, Z. Johnson, C. D. Sinigalliano, M. L. Gidley, A. Zingone, R. Danovaro, G. Tsiamis, M. S. Clark, A. C. Costa, M. El Bour, A. M. Martins, R. Eric Collins, A. L. Ducluzeau, J. Martinez, M. J. Costello, L. A. Amaral-Zettler, J. A. Gilbert, N. Davies, D. Field, and F. O. Glöckner. 2015. The ocean sampling day consortium.
- Kozlov, A. M., J. Zhang, P. Yilmaz, F. O. Glöckner, and A. Stamatakis. 2016. Phylogeny-aware identification and correction of taxonomically mislabeled sequences. *Nucleic Acids Res.* 44:5022–5033.
- Krehenwinkel, H., M. Wolf, J. Y. Lim, A. J. Rominger, W. B. Simison, and R. G. Gillespie. 2017. Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. *Sci. Rep.* 7:1–12.
- Lanfear, R., S. Y. W. Ho, D. Love, and L. Bromham. 2010a. Mutation rate is linked to diversification in birds. *Proc. Natl. Acad. Sci. U. S. A.* 107:20423–20428.
- Lanfear, R., J. J. Welch, and L. Bromham. 2010b. Watching the clock: Studying variation in rates of molecular evolution between species.
- Lara, E., L. Roussel-Delif, B. Fournier, D. M. Wilkinson, and E. A. D. Mitchell. 2016. Soil microorganisms behave like macroscopic organisms: patterns in the global distribution of soil euglyphid testate amoebae. *J. Biogeogr.* 43:520–532.
- Lartillot, N., and R. Poujol. 2011. A Phylogenetic Model for Investigating Correlated Evolution of Substitution Rates and Continuous Phenotypic Characters. *Mol. Biol. Evol.* 28:729–744.

- Lax, G., Y. Eglit, L. Eme, E. M. Bertrand, A. J. Roger, and A. G. B. Simpson. 2018. Hemimastigophora is a novel supra-kingdom-level lineage of eukaryotes. *Nature* 564:410–414.
- Lentendu, G., and M. Dunthorn. 2021. Phylogenetic relatedness drives protist assembly in marine and terrestrial environments. *Glob. Ecol. Biogeogr.* 00:geb.13317.
- Lentendu, G., F. Mahé, D. Bass, S. Rueckert, T. Stoeck, and M. Dunthorn. 2018. Consistent patterns of high alpha and low beta diversity in tropical parasitic and free-living protists. *Mol. Ecol.* 27:2846–2857.
- Lewis, P. O. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst. Biol.* 50:913–925.
- Lewitus, E., L. Bittner, S. Malviya, C. Bowler, and H. Morlon. 2018. Clade-specific diversification dynamics of marine diatoms since the Jurassic. *Nat. Ecol. Evol.* 2:1715–1723.
- Lima-Mendez, G., K. Faust, N. Henry, J. Decelle, S. Colin, F. Carcillo, S. Chaffron, J. C. Ignacio-Espinosa, S. Roux, F. Vincent, L. Bittner, Y. Darzi, J. Wang, S. Audic, L. Berline, G. Bontempi, A. M. Cabello, L. Coppola, F. M. Cornejo-Castillo, F. D’Ovidio, L. De Meester, I. Ferrera, M. J. Garet-Delmas, L. Guidi, E. Lara, S. Pesant, M. Royo-Llonch, G. Salazar, P. Sánchez, M. Sebastian, C. Souffreau, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, G. Gorsky, F. Not, H. Ogata, S. Speich, L. Stemmann, J. Weissenbach, P. Wincker, S. G. Acinas, S. Sunagawa, P. Bork, M. B. Sullivan, E. Karsenti, C. Bowler, C. De Vargas, J. Raes, E. Boss, M. Follows, N. Grimsley, P. Hingamp, D. Iudicone, O. Jaillon, L. Karp-Boss, U. Krzic, E. G. Reynaud, C. Sardet, M. Sieracki, and D. Velayoudon. 2015. Determinants of community structure in the global plankton interactome. *Science* (80-. ). 348.
- Logares, R., S. Audic, D. Bass, L. Bittner, C. Boutte, R. Christen, J.-M. Claverie, J. Decelle, J. R. Dolan, M. Dunthorn, B. Edvardsen, A. Gobet, W. H. C. F. Kooistra, F. Mahé, F. Not, H. Ogata, J. Pawlowski, M. C. Pernice, S. Romac, K. Shalchian-Tabrizi, N. Simon, T. Stoeck, S. Santini, R. Siano, P. Wincker, A. Zingone, T. A. Richards, C. de Vargas, and R. Massana. 2014. Patterns of rare and abundant marine microbial eukaryotes. *Curr. Biol.* 24:813–821.
- Logares, R., J. Bråte, S. Bertilsson, J. L. Clasen, K. Shalchian-Tabrizi, and K. Rengefors. 2009. Infrequent marine–freshwater transitions in the microbial world. *Trends Microbiol.* 17:414–422.
- Logares, R., N. Daugbjerg, A. Boltovskoy, A. Kremp, J. Laybourn-Parry, and K. Rengefors. 2008. Recent evolutionary diversification of a protist lineage. *Environ. Microbiol.* 10:1231–1243.
- Logares, R., J.-F. Mangot, and R. Massana. 2015. Rarity in aquatic microbes: placing protists on the map. *Res. Microbiol.* 166:831–841.
- Logares, R., K. Rengefors, A. Kremp, K. Shalchian-Tabrizi, A. Boltovskoy, T. Tengs, A. Shurtleff, and D. Klaveness. 2007a. Phenotypically different microalgal morphospecies with identical ribosomal DNA: A case of rapid adaptive evolution? *Microb. Ecol.* 53.
- Logares, R., K. Shalchian-Tabrizi, A. Boltovskoy, and K. Rengefors. 2007b. Extensive dinoflagellate phylogenies indicate infrequent marine-freshwater transitions. *Mol. Phylogenet. Evol.* 45:887–903.
- Loit, K., K. Adamson, M. Bahram, R. Puusepp, S. Anslan, R. Kiiker, R. Drenkhan, and L. Tedersoo. 2019. Relative performance of Oxford Nanopore MinION vs. Pacific Biosciences Sequel third-generation sequencing platforms in identification of agricultural and forest pathogens. *bioRxiv* 592972.

- López-García, P., F. Rodríguez-Valera, C. Pedrós-Alió, and D. Moreira. 2001. Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature* 409:603–607.
- Louca, S., P. M. Shih, M. W. Pennell, W. W. Fischer, L. W. Parfrey, and M. Doebeli. 2018. Bacterial diversification through geological time. *Nat. Ecol. Evol.* 2:1458–1467.
- Luan, L., Y. Jiang, M. Cheng, F. Dini-Andreote, Y. Sui, Q. Xu, S. Geisen, and B. Sun. 2020. Organism body size structures the soil microbial and nematode community assembly at a continental and global scale. *Nat. Commun.* 11:1–11.
- Maddison, D. R. 1994. Phylogenetic methods for inferring the evolutionary history and processes of change in discretely valued characters.
- Mahé, F., C. De Vargas, D. Bass, L. Czech, A. Stamatakis, E. Lara, D. Singer, J. Mayor, J. Bunge, S. Sernaker, T. Siemensmeyer, I. Trautmann, S. Romac, C. Berney, A. Kozlov, E. A. D. Mitchell, C. V. W. Seppey, E. Egge, G. Lentendu, R. Wirth, G. Trueba, and M. Dunthorn. 2017. Parasites dominate hyperdiverse soil protist communities in Neotropical rainforests. *Nat. Ecol. Evol.* 1:0091.
- Mahé, F., T. Rognes, C. Quince, C. de Vargas, and M. Dunthorn. 2014. Swarm: Robust and fast clustering method for amplicon-based studies. *PeerJ* 2014:e593.
- Marande, W., P. López-García, and D. Moreira. 2009. Eukaryotic diversity and phylogeny using small- and large-subunit ribosomal RNA genes from environmental samples. *Environ. Microbiol.* 11:3179–3188.
- Martijn, J., A. E. Lind, M. E. Schön, I. Spiertz, L. Juzokaite, I. Bunikis, O. V. Pettersson, and T. J. G. Ettema. 2019. Confident phylogenetic identification of uncultured prokaryotes through long read amplicon sequencing of the 16S-ITS-23S rRNA operon. *Environ. Microbiol.* 1462-2920.14636.
- Martin, C. 2015. Biology's dark matter. *Curr. Biol.* 25:pR301–R307.
- Martinez-Garcia, M., D. Brazel, N. J. Poulton, B. K. Swan, M. L. Gomez, D. Masland, M. E. Sieracki, and R. Stepanauskas. 2012. Unveiling in situ interactions between marine protists and bacteria through single cell sequencing. *ISME J.* 6:703–707.
- Massana, R. 2015. Getting specific: making taxonomic and ecological sense of large sequencing data sets. *Mol. Ecol.* 24:2904–2906.
- Massana, R., V. Balagué, L. Guillou, and C. Pedrós-Alió. 2004a. Picoeukaryotic diversity in an oligotrophic coastal site studied by molecular and culturing approaches. *FEMS Microbiol. Ecol.* 50:231–243.
- Massana, R., J. Castresana, V. Balagué, L. Guillou, K. Romari, A. Groisillier, K. Valentin, and C. Pedrós-Alió. 2004b. Phylogenetic and ecological analysis of novel marine stramenopiles. *Appl. Environ. Microbiol.* 70:3528–34.
- Massana, R., J. Castresana, V. Balague, L. Guillou, K. Romari, A. Groisillier, K. Valentin, C. Pedros-Alio, V. Balagué, L. Guillou, K. Romari, A. Groisillier, K. Valentin, and C. Pedrós-Alió. 2004c. Phylogenetic and ecological analysis of novel marine stramenopiles. *Appl. Environ. Microbiol.* 70:3528–3534.
- Massana, R., A. Gobet, S. Audic, D. Bass, L. Bittner, C. Boute, A. Chambouvet, R. Christen, J.-M. M. Claverie, J. Decelle, J. R. Dolan, M. Dunthorn, B. Edvardsen, I. Forn, D. Forster, L. Guillou, O. Jaillon, W. H. C. F. C. F. C. F. Kooistra, R. Logares, F. Mahé, F. Not, H. Ogata, J. Pawlowski, M. C. Pernice, I. Probert, S. Romac, T. Richards, S. Santini, K. Shalchian-Tabrizi, R. Siano, N. Simon, T. Stoeck, D. Vault, A. Zingone, and C. de Vargas. 2015. Marine protist diversity in European coastal waters and sediments as revealed by high-throughput sequencing. *Environ. Microbiol.* 17:4035–4049.

- Massana, R., L. Guillou, B. Díez, and C. Pedrós-Alió. 2002. Unveiling the organisms behind novel eukaryotic ribosomal DNA sequences from the ocean. *Appl. Environ. Microbiol.* 68:4554–8.
- Massana, R., F. Unrein, R. Rodríguez-Martínez, I. Forn, T. Lefort, J. Pinhassi, and F. Not. 2009. Grazing rates and functional diversity of uncultured heterotrophic flagellates. *ISME J.* 3:588–595.
- Matsen, F. A., R. B. Kodner, and E. V. Armbrust. 2010. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* 11:538.
- McGill, B. J., J. M. Chase, J. Hortal, I. Overcast, A. J. Rominger, J. Rosindell, P. A. V Borges, B. C. Emerson, R. Etienne, M. J. Hickerson, D. L. Mahler, F. Massol, A. McGaughan, P. Neves, C. Parent, J. Patiño, M. Ruffley, C. E. Wagner, and R. Gillespie. 2019. Unifying macroecology and macroevolution to answer fundamental questions about biodiversity. *Glob. Ecol. Biogeogr.* 1–12:geb.13020.
- McPeck, M. A. 2017. *Evolutionary Community Ecology*.
- Moon-Van Der Staay, S. Y., R. De Wachter, and D. Vaulot. 2001. Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature*, doi: 10.1038/35054541.
- Mora, C., D. P. Tittensor, S. Adl, A. G. B. Simpson, and B. Worm. 2011. How Many Species Are There on Earth and in the Ocean? *PLoS Biol.* 9:e1001127.
- Moran, N. A. 1996. Accelerated evolution and Muller's ratchet in endosymbiotic bacteria. *Proc. Natl. Acad. Sci. U. S. A.* 93:2873–8.
- Mosher, J. J., E. L. Bernberg, O. Shevchenko, J. Kan, and L. A. Kaplan. 2013. Efficacy of a 3rd generation high-throughput sequencing platform for analyses of 16S rRNA genes from environmental samples. *J. Microbiol. Methods* 95:175–181.
- Mosher, J. J., B. Bowman, E. L. Bernberg, O. Shevchenko, J. Kan, J. Korlach, L. A. Kaplan, and L. A. Kaplan. 2014. Improved performance of the PacBio SMRT technology for 16S rDNA sequencing. *J. Microbiol. Methods* 104:59–60.
- Mukherjee, I., Y. Hodoki, and S. ichi Nakano. 2015. Kinetoplastid flagellates overlooked by universal primers dominate in the oxygenated hypolimnion of Lake Biwa, Japan. *FEMS Microbiol. Ecol.* 91:83.
- Mukherjee, I., M. M. Salcher, A. Ş. Andrei, V. S. Kavagutti, T. Shabarova, V. Grujčić, M. Haber, P. Layoun, Y. Hodoki, S. ichi Nakano, K. Šimek, and R. Ghai. 2020. A freshwater radiation of diplomonads. *Environ. Microbiol.* 22:4658–4668.
- Newmaster, S. G., A. J. Fazekas, and S. Ragupathy. 2006. DNA barcoding in land plants: Evaluation of rbcL in a multigene tiered approach.
- Obiol, A., C. R. Giner, P. Sánchez, C. M. Duarte, S. G. Acinas, and R. Massana. 2020. A metagenomic assessment of microbial eukaryotic diversity in the global ocean. *Mol. Ecol. Resour.* 20:718–731.
- Oliverio, A. M., S. Geisen, M. Delgado-Baquerizo, F. T. Maestre, B. L. Turner, and N. Fierer. 2020. The global-scale distributions of soil protists and their contributions to belowground systems. *Sci. Adv.* 6:eaax8787.
- Oliverio, A. M., J. F. Power, A. Washburne, S. C. Cary, M. B. Stott, and N. Fierer. 2018. The ecology and diversity of microbial eukaryotes in geothermal springs. *ISME J.* 12:1918–1928.
- Oloo, F., A. Valverde, M. V. Quiroga, S. Vikram, D. Cowan, and G. Mataloni. 2016. Habitat heterogeneity and connectivity shape microbial communities in South American peatlands. *Sci. Rep.* 6:1–8.

- Orr, R. J. S., S. Zhao, D. Klaveness, A. Yabuki, K. Ikeda, M. M. Watanabe, and K. Shalchian-Tabrizi. 2018. Enigmatic Diphyllatea eukaryotes: culturing and targeted PacBio RS amplicon sequencing reveals a higher order taxonomic diversity and global distribution. *BMC Evol. Biol.* 18:115.
- Pagel, M. 1994. Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. *Proc. R. Soc. B Biol. Sci.* 255:37–45.
- Pagel, M., A. Meade, and D. Barker. 2004. Bayesian Estimation of Ancestral Character States on Phylogenies. *Syst. Biol.* 53:673–684.
- Pagel, M., C. Venditti, and A. Meade. 2006. Large punctuational contribution of speciation to evolutionary divergence at the molecular level. *Science* (80-. ). 314:119–121.
- Parker, J., A. J. Helmstetter, D. Devey, T. Wilkinson, and A. S. T. Papadopoulos. 2017. Field-based species identification of closely-related plants using real-time nanopore sequencing. *Sci. Rep.* 7:8345.
- Pernice, M. C., C. R. Giner, R. Logares, J. Perera-Bel, S. G. Acinas, C. M. Duarte, J. M. Gasol, and R. Massana. 2016. Large variability of bathypelagic microbial eukaryotic communities across the world's oceans. *ISME J.* 10:945–958.
- Peterson, A. T., J. Soberón, and V. Sánchez-Cordero. 1999. Conservatism of ecological niches in evolutionary time. *Science* (80-. ). 285:1265–1267.
- Pitsch, G., E. P. Bruni, D. Forster, Z. Qu, B. Sonntag, T. Stoeck, and T. Posch. 2019. Seasonality of planktonic freshwater ciliates: Are analyses based on V9 regions of the 18S rRNA gene correlated with morphospecies counts? *Front. Microbiol.* 10:248.
- Piwosz, K., I. Mukherjee, M. M. Salcher, V. Grujčić, and K. Šimek. 2021. CARD-FISH in the Sequencing Era: Opening a New Universe of Protistan Ecology.
- Pomerantz, A., N. Peñafiel, A. Arteaga, L. Bustamante, F. Pichardo, L. A. Coloma, C. L. Barrio-Amorós, D. Salazar-Valenzuela, and S. Prost. 2018. Real-time DNA barcoding in a rainforest using nanopore sequencing: Opportunities for rapid biodiversity assessments and local capacity building. *Gigascience* 7:1–14.
- Quast, C., E. Priesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glöckner. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 41:D590–D596.
- Quick, J., P. Ashton, S. Calus, C. Chatt, S. Gossain, J. Hawker, S. Nair, K. Neal, K. Nye, T. Peters, E. De Pinna, E. Robinson, K. Struthers, M. Webber, A. Catto, T. J. Dallman, P. Hawkey, and N. J. Loman. 2015. Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of *Salmonella*. *Genome Biol.* 16:114.
- Rhoads, A., and K. F. Au. 2015. PacBio Sequencing and Its Applications. *Genomics. Proteomics Bioinformatics* 13:278–289.
- Richter, D. J., R. Watteaux, T. Vannier, J. Leconte, P. Frémont, G. Reygondeau, N. Maillet, N. Henry, G. Benoit, A. Fernández-Guerra, S. Suweis, R. Narci, C. Berney, D. Eveillard, F. Gavory, L. Guidi, K. Labadie, E. Mahieu, J. Poulain, S. Romac, S. Roux, C. Dimier, S. Kandels, M. Picheral, S. Searson, T. O. Coordinators, S. Pesant, J.-M. Aury, J. R. Brum, C. Lemaitre, E. Pelletier, P. Bork, S. Sunagawa, L. Karp-Boss, C. Bowler, M. B. Sullivan, E. Karsenti, M. Mariadassou, I. Probert, P. Peterlongo, P. Wincker, C. de Vargas, M. R. d'Alcalá, D. Iudicone, O. Jaillon, and T. O. Coordinators. 2019. Genomic evidence for global ocean plankton biogeography shaped by large-scale current systems. *bioRxiv* 867739.

- Robertson, G. P., K. M. Klingensmith, M. J. Klug, E. A. Paul, J. R. Crum, and B. G. Ellis. 1997. SOIL RESOURCES, MICROBIAL ACTIVITY, AND PRIMARY PRODUCTION ACROSS AN AGRICULTURAL ECOSYSTEM.
- Rognes, T., T. Flouri, B. Nichols, C. Quince, and F. Mahé. 2016. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 4:e2584.
- Santoferrara, L., F. Burki, S. Filker, R. Logares, M. Dunthorn, and G. B. McManus. 2020. Perspectives from Ten Years of Protist Studies by High-Throughput Metabarcoding. *J. Eukaryot. Microbiol.* 67:612–622.
- Santoferrara, L. F. 2019. Current practice in plankton metabarcoding: Optimization and error management. *J. Plankton Res.* 41.
- Sarno, D., W. H. C. F. Kooistra, L. K. Medlin, I. Percopo, and A. Zingone. 2005. Diversity in the genus *Skeletonema* (Bacillariophyceae). II. An assessment of the taxonomy of *S. costatum*-like species with the description of four new species. *J. Phycol.* 41:151–176.
- Schloss, P. D., M. L. Jenior, C. C. Koumpouras, S. L. Westcott, and S. K. Highlander. 2016. Sequencing 16S rRNA gene fragments using the PacBio SMRT DNA sequencing system. *PeerJ* 4:e1869.
- Schloss, P. D., S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, J. W. Sahl, B. Stres, G. G. Thallinger, D. J. Van Horn, and C. F. Weber. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* 75:7537–7541.
- Schoch, C. L., K. A. Seifert, S. Huhndorf, V. Robert, J. L. Spouge, C. A. Levesque, W. Chen, F. B. Fungal Barcoding Consortium, F. B. C. A. Fungal Barcoding Consortium Author List, E. Bolchacova, K. Voigt, P. W. Crous, A. N. Miller, M. J. Wingfield, M. C. Aime, K.-D. An, F.-Y. Bai, R. W. Barreto, D. Begerow, M.-J. Bergeron, M. Blackwell, T. Boekhout, M. Bogale, N. Boonyuen, A. R. Burgaz, B. Buyck, L. Cai, Q. Cai, G. Cardinali, P. Chaverri, B. J. Coppins, A. Crespo, P. Cubas, C. Cummings, U. Damm, Z. W. de Beer, G. S. de Hoog, R. Del-Prado, B. Dentinger, J. Diéguez-Uribeondo, P. K. Divakar, B. Douglas, M. Dueñas, T. A. Duong, U. Eberhardt, J. E. Edwards, M. S. Elshahed, K. Fliegerova, M. Furtado, M. A. García, Z.-W. Ge, G. W. Griffith, K. Griffiths, J. Z. Groenewald, M. Groenewald, M. Grube, M. Gryzenhout, L.-D. Guo, F. Hagen, S. Hambleton, R. C. Hamelin, K. Hansen, P. Harrold, G. Heller, C. Herrera, K. Hirayama, Y. Hirooka, H.-M. Ho, K. Hoffmann, V. Hofstetter, F. Högnabba, P. M. Hollingsworth, S.-B. Hong, K. Hosaka, J. Houbraken, K. Hughes, S. Huhtinen, K. D. Hyde, T. James, E. M. Johnson, J. E. Johnson, P. R. Johnston, E. B. G. Jones, L. J. Kelly, P. M. Kirk, D. G. Knapp, U. Kõljalg, G. M. Kovács, C. P. Kurtzman, S. Landvik, S. D. Leavitt, A. S. Liggenstoffer, K. Liimatainen, L. Lombard, J. J. Luangsa-ard, H. T. Lumbsch, H. Maganti, S. S. N. Maharachchikumbura, M. P. Martin, T. W. May, A. R. McTaggart, A. S. Methven, W. Meyer, J.-M. Moncalvo, S. Mongkolsamrit, L. G. Nagy, R. H. Nilsson, T. Niskanen, I. Nyilasi, G. Okada, I. Okane, I. Olariaga, J. Otte, T. Papp, D. Park, T. Petkovits, R. Pino-Bodas, W. Quaedvlieg, H. A. Raja, D. Redecker, T. L. Rintoul, C. Ruibal, J. M. Sarmiento-Ramírez, I. Schmitt, A. Schübler, C. Shearer, K. Sotome, F. O. P. Stefani, S. Stenroos, B. Stielow, H. Stockinger, S. Suetrong, S.-O. Suh, G.-H. Sung, M. Suzuki, K. Tanaka, L. Tedersoo, M. T. Telleria, E. Tretter, W. A. Untereiner, H. Urbina, C. Vágvölgyi, A. Vialle, T. D. Vu, G. Walther, Q.-M. Wang, Y. Wang, B. S. Weir, M. Weiß, M. M. White, J. Xu, R. Yahr, Z. L. Yang, A. Yurkov, J.-C. Zamora,

- N. Zhang, W.-Y. Zhuang, and D. Schindel. 2012. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc. Natl. Acad. Sci. U. S. A.* 109:6241–6.
- Schön, M. E., V. V. Zlatogursky, R. P. Singh, C. Poirier, S. Wilken, V. Mathur, J. F. H. Strasser, J. Pinhassi, A. Z. Worden, P. J. Keeling, T. J. G. Ettema, J. G. Wideman, and F. Burki. 2021. Picozoa are archaeplastids without plastid. *bioRxiv* 2021.04.14.439778.
- Seppey, C. V. W., O. Broennimann, A. Buri, E. Yashiro, E. Pinto-Figueroa, D. Singer, Q. Blandenier, E. A. D. Mitchell, H. Niculita-Hirzel, A. Guisan, and E. Lara. 2020. Soil protist diversity in the Swiss western Alps is better predicted by topo-climatic than by edaphic variables. *J. Biogeogr.* 47:866–878.
- Sieber, G., D. Beisser, C. Bock, and J. Boenigk. 2020. Protistan and fungal diversity in soils and freshwater lakes are substantially different. *Sci. Rep.* 10.
- Sierra, R., S. J. Cañas-Duarte, F. Burki, A. Schwelm, J. Fogelqvist, C. Dixelius, L. N. González-García, G. H. Gile, C. H. Slamovits, C. Klopp, S. Restrepo, I. Arzul, and J. Pawlowski. 2016. Evolutionary Origins of Rhizarian Parasites. *Mol. Biol. Evol.* 33:980–983.
- Simpson, A. G. B., and D. J. Patterson. 1999. The ultrastructure of *Carpodomonas membranifera* (Eukaryota) with reference to the “excavate hypothesis.” *Eur. J. Protistol.* 35:353–370.
- Singer, D., E. A. D. Mitchell, R. J. Payne, Q. Blandenier, C. Duckert, L. D. Fernández, B. Fournier, C. E. Hernández, G. Granath, H. Rydin, L. Bragazza, N. G. Koronotova, I. Goia, L. I. Harris, K. Kajukalo, A. Kosakyan, M. Lamentowicz, N. P. Kosykh, K. Vellak, and E. Lara. 2019. Dispersal limitations and historical factors determine the biogeography of specialized terrestrial protists. *Mol. Ecol.* 28:3089–3100.
- Singer, D., C. V. W. Seppey, G. Lentendu, M. Dunthorn, D. Bass, L. Belbahri, Q. Blandenier, D. Debroas, G. A. de Groot, C. de Vargas, I. Domaizon, C. Duckert, I. Izaguirre, I. Koenig, G. Mataloni, M. R. Schiaffino, E. A. D. Mitchell, S. Geisen, and E. Lara. 2021. Protist taxonomic and functional diversity in soil, freshwater and marine ecosystems. *Environ. Int.* 146.
- Smith, S. A., and M. J. Donoghue. 2008. Rates of molecular evolution are linked to life history in flowering plants. *Science* (80-. ). 322:86–89.
- Stamatakis, A. 2018. Phylogeny-aware Analysis & Post-Analysis of Short Reads.
- Stamatakis, A., A. M. Kozlov, and A. Kozlov. 2020. Efficient Maximum Likelihood Tree Building Methods. P. 1.2:18 in C. Scornavacca, F. Delsuc, and N. Galtier, eds. *Phylogenetics in the Genomic Era*.
- Strasser, J. F. H., M. Jamy, A. P. Mylnikov, D. V. Tikhonenkov, and F. Burki. 2019. New phylogenomic analysis of the enigmatic phylum Telonemia further resolves the eukaryote tree of life. *Mol. Biol. Evol.* 36.
- Strasser, J. F. H., C. Wurzbacher, V. Hervé, T. Antany, A. Brune, and R. Radek. 2021. Long rDNA amplicon sequencing of insect-infecting nephridiophagids reveals their affiliation to the Chytridiomycota and a potential to switch between hosts. *Sci. Rep.* 11:396.
- Swofford, D. L., and W. P. Maddison. 1987. Reconstructing ancestral character states under Wagner parsimony. *Math. Biosci.* 87:199–229.
- Taberlet, P., A. Bonin, L. Zinger, and E. Coissac. 2018. Environmental DNA: For biodiversity research and monitoring.
- Tedersoo, L., A. Tooming-Klunderud, and S. Anslan. 2018. PacBio metabarcoding of Fungi and other eukaryotes: errors, biases and perspectives. *New Phytol.* 217:1370–1385.

- Varga, T., K. Krizsán, C. Földi, B. Dima, M. Sánchez-García, S. Sánchez-Ramírez, G. J. Szöllősi, J. G. Szarkándi, V. Papp, L. Albert, W. Andreopoulos, C. Angelini, V. Antonín, K. W. Barry, N. L. Bougher, P. Buchanan, B. Buyck, V. Bense, P. Catcheside, M. Chovatia, J. Cooper, W. Dämon, D. Desjardin, P. Finy, J. Geml, S. Haridas, K. Hughes, A. Justo, D. Karasiński, I. Kautmanova, B. Kiss, S. Kocsubé, H. Kotiranta, K. M. LaButti, B. E. Lechner, K. Liimatainen, A. Lipzen, Z. Lukács, S. Mihaltcheva, L. N. Morgado, T. Niskanen, M. E. Noordeloos, R. A. Ohm, B. Ortiz-Santana, C. Ovrebo, N. Rác, R. Riley, A. Savchenko, A. Shiryaev, K. Soop, V. Spirin, C. Szebenyi, M. Tomšovský, R. E. Tulloss, J. Uehling, I. V. Grigoriev, C. Vágvölgyi, T. Papp, F. M. Martin, O. Miettinen, D. S. Hibbett, and L. G. Nagy. 2019. Megaphylogeny resolves global patterns of mushroom evolution. *Nat. Ecol. Evol.* 3:668–678.
- Vellend, M. 2010. Conceptual synthesis in community ecology. *Q. Rev. Biol.* 85:183–206.
- Vellend, M. 2016. *The Theory of Ecological Communities*.
- Von Der Heyden, S., E. E. Chao, and T. Cavalier-Smith. 2004. Genetic diversity of goniomonads: An ancient divergence between marine and freshwater species. *Eur. J. Phycol.* 39:343–350.
- Wagner, J., P. Coupland, H. P. Browne, T. D. Lawley, S. C. Francis, and J. Parkhill. 2016. Evaluation of PacBio sequencing for full-length bacterial 16S rRNA gene classification. *BMC Microbiol.* 16:274.
- Ward, G. M., S. Neuhauser, R. Groben, S. Ciaghi, C. Berney, S. Romac, and D. Bass. 2018. Environmental Sequencing Fills the Gap Between Parasitic Haplosporidians and Free-living Giant Amoebae. *J. Eukaryot. Microbiol.* 65:574–586.
- West, P. T., A. J. Probst, I. V. Grigoriev, B. C. Thomas, and J. F. Banfield. 2018. Genome-reconstruction for eukaryotes from complex natural microbial communities. *Genome Res.* 28:569–580.
- Wolf, D. I., and M. L. Vis. 2020. Stream Algal Biofilm Community Diversity Along An Acid Mine Drainage Recovery Gradient Using Multimarker Metabarcoding. *J. Phycol.* 56:11–22.
- Wright, S., J. Keeling, and L. Gillman. 2006. The road from Santa Rosalia: A faster tempo of evolution in tropical climates.
- Xue, Y., H. Chen, J. R. Yang, M. Liu, B. Huang, and J. Yang. 2018. Distinct patterns and processes of abundant and rare eukaryotic plankton communities following a reservoir cyanobacterial bloom. *ISME J.* 12:2263–2277.
- Žerdoner Čalasan, A., J. Kretschmann, and M. Gottschling. 2019. They are young, and they are many: dating freshwater lineages in unicellular dinophytes. *Environ. Microbiol.* 21.
- Zhao, S., F. Burki, J. Brate, P. J. Keeling, D. Klaveness, and K. Shalchian-Tabrizi. 2012. Collodictyon—an ancient lineage in the tree of eukaryotes. *Mol. Biol. Evol.* 29:1557–1568.



# Acta Universitatis Upsaliensis

*Digital Comprehensive Summaries of Uppsala Dissertations  
from the Faculty of Science and Technology 2054*

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. (Prior to January, 2005, the series was published under the title "Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology".)

Distribution: [publications.uu.se](http://publications.uu.se)  
urn:nbn:se:uu:diva-446935



ACTA  
UNIVERSITATIS  
UPSALIENSIS  
UPPSALA  
2021