Novel insights into protist diversity and niche adaptation using single cell transcriptomics

HENNING ONSBRING
Protists are a polyphyletic group of microbes that represents the vast majority of eukaryotic diversity. Despite this, most sequencing efforts targeting eukaryotes have been focused on animals, fungi and plants. The sequencing bias towards multicellular organisms can partially be explained by the difficulty in cultivating protists, which is needed in traditional sequencing workflows. In this thesis, single-cell RNA sequencing has been used to generate transcriptome data from environmental protists, without being dependent on establishing a culture. These transcriptome data have been used to discover novel protist diversity, as well as exploring the cell biology of two ciliates.

In the first chapter, transcriptomes of cell fragments were generated for the ciliate Stentor. This ciliate is well-known for its ability to repair drastic cellular wounds, and the transcriptomes uncovered genes involved in processes such as cell cycle, signaling and microtubule-based movement to be activated during Stentor regeneration.

Spirostomum semivirescens is another ciliate, whose transcriptome was generated using single-cell RNA sequencing. The transcriptome data suggest that S. semivirescens is using rhodoquinol-dependent fumarate reduction for respiration in environments with low levels of oxygen.

Single-cell RNA sequencing was further used to target cells smaller than Stentor and Spirostomum. By generating 124 transcriptomes of environmental protists, a high number of novel lineages could be identified. The generated transcriptome data included free-living prokinetoplastids, non-photosynthetic euglenids, metamonads and katablepharids.

A few modifications to the single-cell RNA sequencing protocol Smart-seq2 were necessary to generate the 124 transcriptomes of small protists cells. The impact of these modifications to Smart-seq2 was benchmarked using Giardia intestinalis. The generated single-cell transcriptomes revealed that addition of freeze-thaw cycles to Smart-seq2 improved transcript recovery. Finally, we propose a protocol that allows identification of failed cDNA reactions, based only on measuring DNA concentration, without compromising on transcript recovery. Reducing the dependency on quality control will be important if single-cell RNA sequencing would be done in a high-throughput workflow.

In conclusion, single-cell RNA sequencing can be a powerful tool for studying protist diversity and biology. In particular, it has the potential to efficiently uncover protist diversity, provided that a robust and efficient method to isolate single cells from the environment is established.

Keywords: Protists, microbial eukaryotes, cultivation-independent methods, single-cell RNA sequencing, phylogenomics

Henning Onsbring, Department of Cell and Molecular Biology, Box 596, Uppsala University, SE-75124 Uppsala, Sweden.

© Henning Onsbring 2019

ISSN 1651-6214
urn:nbn:se:uu:diva-392618 (http://urn.kb.se/resolve?urn=nbn:se:uu:diva-392618)
List of Papers

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


* Shared first authorship

§ Shared last authorship

Reprints were made with permission from the respective publishers.
Cover designed by Julia Vrager, with inspiration from photos of protists studied in this thesis, and descriptions of the elongated and non-elongated euglenids from Paper III.
Contents

Introduction ................................................................................................ 9
Gaps in our knowledge about eukaryotic diversity can lead to mis-
interpretations .....................................................................................12
Traditional protistology ......................................................................15
Amplicon sequencing from whole community ......................................16
Metagenomics .....................................................................................18
Metatranscriptomics ..........................................................................20
Single-cell genomics .........................................................................22
Single-cell RNA sequencing ...............................................................25
Which method should be used to expand the known diversity in the
eukaryotic tree? ................................................................................27
How should scRNAseq be applied to increase the known diversity in the
eukaryotic tree? .............................................................................30
Future perspectives .............................................................................34
Paper summaries ..............................................................................36
  Paper I. RNA Sequencing of Stentor Cell Fragments Reveals
  Transcriptional Changes during Cellular Regeneration .....................36
  Paper II. Molecular Investigation of the Ciliate Spirostomum
  semivirescens, with First Transcriptome and New Geographical
  Records ...............................................................................................37
  Paper III. Single-cell transcriptomics expands sampled protist diversity
  and provides insights into niche adaptation ........................................38
  Paper IV. An efficient single-cell transcriptomics workflow to assess
  protist diversity and lifestyle ..........................................................39
Svensk sammanfattning .......................................................................41
Acknowledgements .............................................................................43
References ..........................................................................................45
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>cDNA</td>
<td>Complementary DNA</td>
</tr>
<tr>
<td>DHFR</td>
<td>Dihydrofolate reductase</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>FACS</td>
<td>Fluorescence-activated cell sorting</td>
</tr>
<tr>
<td>IL</td>
<td>Isocitrate lyase</td>
</tr>
<tr>
<td>MDA</td>
<td>Multiple displacement amplification</td>
</tr>
<tr>
<td>mRNA</td>
<td>Messenger RNA</td>
</tr>
<tr>
<td>MS</td>
<td>Malate synthase</td>
</tr>
<tr>
<td>ORF</td>
<td>Open reading frame</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase chain reaction</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic acid</td>
</tr>
<tr>
<td>SCG</td>
<td>Single-cell genomics</td>
</tr>
<tr>
<td>scRNAseq</td>
<td>Single-cell RNA sequencing</td>
</tr>
<tr>
<td>TS</td>
<td>Thymidylate synthase</td>
</tr>
</tbody>
</table>
Introduction

This thesis is the product of exploratory work targeting microbial eukaryotes, a polyphyletic group also referred to as protists. It is thought that there is a wide diversity of protists that has not yet been discovered (Keeling and Burki, 2019). This project was initiated in an effort to establish ways to uncover such unknown protist diversity. Generating new data for genomically unexplored protists can increase our understanding of eukaryotic evolution and the metabolic processes that are shaping our environment.

The first time I met my supervisor Thijs Ettema, and got introduced to the wide diversity of microbes in the environment, was during a course in microbial genomics at Uppsala University. During this course, where Thijs was a lecturer, he mentioned the single-cell platform he managed at that time, and how his research group explored unknown microbial diversity. Technologies for generating genome and transcriptome data from single eukaryotic cells were also emerging at this time (Picelli et al., 2013; Yoon et al., 2011). With the opportunities that opened up from single-cell methods, I felt that it was the right time to study microbial diversity without relying on cultivation. I decided to contact Thijs who was applying cultivation-independent methods in his research.

We met once to discuss general scientific interests and already at our second meeting we both agreed that doing single-cell RNA sequencing would be a good project to start with. This was also the project that I started with, and my main focus throughout my thesis work has been testing and establishing workflows to effectively sequence protist transcriptomes to expand the known diversity in the eukaryotic tree (Figure 1).

The first article I published was about which genes are important for the repair of cellular wounds (Paper I). In this study we used the single-cell RNA sequencing method to characterize changes in gene expression from pieces of Stentor polymorphus cells that are recovering from being fragmented. In the next study I generated single-cell transcriptomes of another ciliate, Spirostomum semivirescens, which could be an important species to study to better understand the evolution of the genetic code (Paper II). The cells from Paper I and II are both relatively large, and after adjusting the single-cell RNA sequencing protocol I could eventually start generating transcriptomes of smaller protists (Paper III), which allowed me to uncover previously unknown diversity. I finally describe how to perform the single-cell RNA sequencing protocol adapted for protists in Paper IV. Additionally, Paper IV benchmarks
different method modifications that could improve transcript recovery from low amounts of mRNA.

![Eukaryotic Tree of Life](image)

**Figure 1.** The eukaryotic tree of life. Unrooted eukaryotic tree covering major clades and “orphans”, where orphans are indicated as a single branch. Only a few species are known in the orphan groups, and these orphans have diverged from other major groups early during eukaryotic evolution. Clades containing species referred to as Diaphoretickes are indicated with blue, Amorphea are indicated with red and Excavata in yellow. The tree topology is mainly based on Lax et al. (2018), with exception for the placement of the following groups: Metamonada and Discoba according to Burki (2014), Apusomonadida, Breviatidae, Collodictyonidae and Malawimonadidae as Heiss et al. (2018) and the Ancoracystidae position as reported by Janouskovec et al. (2017).

Protists have been the target during my thesis work since my supervisor and I have assumed there are a lot of such lineages left to discover. The reason for that assumption is that among eukaryotes, the sequencing efforts have been strongly biased towards animals, plants and fungi, while protists actually represent the major part of the eukaryotic diversity (del Campo et al., 2014; Sibbald and Archibald, 2017). Additionally, there could still be several deeply rooted branches in the eukaryotic tree left to discover. In recent years many deeply rooted branches, that we previously did not know existed, have been identified in the eukaryotic tree. Such deeply branching species have been
identified both among species for which morphology was previously reported in the literature and for completely novel findings (Brown et al., 2013; Janouskovec et al., 2017; Lax et al., 2018; Strassert et al., 2019; Torruella et al., 2012; Zhao et al., 2012).

To increase our understanding of eukaryotic evolution we both need to search for species from branches that are deeply rooted in the eukaryotic tree, and improve the taxon sampling in already discovered groups. During this thesis work there have been a lot of sampling efforts, mainly around Uppsala (Sweden), and a majority of the transcriptomes from the collected protists are presented in Paper III. There I show that several lineages, which are distantly related from the already known protists can be found, even with low throughput methods where the samples have been collected from easily accessible environments. This further indicates that there are many unknown protist left to discover, and the discovery rate could be accelerated if the throughput in my workflow is increased.

In the forthcoming text, I will first emphasize why generation of data for novel protist lineages are important, followed by discussing various methods that can be used to uncover protist diversity. I will continue with a brief discussion of the advantages and drawbacks with various cultivation-independent methods. Finally, I will mention different possibilities that could further improve the workflow, which is tested in my papers.
Gaps in our knowledge about eukaryotic diversity can lead to mis-interpretations

The importance of generating new data should not be underestimated. Having less complete data leads to higher chances for a pattern to show up that leads to incorrect conclusions. For example, without the transcriptomes of non-photosynthetic euglenids in Paper III, the gene fusion of malate synthase (MS) and isocitrate lyase (IL) could have been inferred to an early common euglenozoa ancestor (Figure 2). This despite the fact that there are currently 70 genomes of euglenozoa publicly available on NCBI. A closer look at the sequenced genomes reveals that two thirds of the genomes are generated only from the family Trypanosomatidae.

The focus towards Trypanosomatidae is because members of this family are insect parasites that can cause diseases to humans and are killing hundreds of thousands of people every year (Simpson et al., 2006). It can be noted that euglenozoa is a very diverse group, containing taxonomic sub groups such as diplonemids, which are the most diverse taxonomic group of all planktonic eukaryotes in the ocean (Flegontova et al., 2016).

Despite the high diplonemid diversity, that is present in one of our biggest ecosystems on earth, there is only one genome available from this sub group of euglenozoia (Morales et al., 2016). The limited data available for diplonemids is not the only example of bias in the sequencing effort among protists. Similar trends can be seen outside of euglenozoa, a lot of focus has been targeted towards specific lineages, e.g. 90% of the protist genomes sequenced are parasites (Sibbald and Archibald, 2017).

Inferring a gene fusion event of malate synthase and isocitrate lyase to the base of euglenozoa would not lead to any drastic conclusions about eukaryotic evolution. However, a fusion between dihydrofolate reductase (DHFR) and thymidylate synthase (TS) has been used to root the eukaryotic tree (Stechmann and Cavalier-Smith, 2002). Using the DHFR-TS fusion places Opisthokonta and Amoebozoa as the earliest branches in the eukaryotic tree, since DHFR and TS are separated in opisthokonts. The DHFR-TS fusion was found in Apusozoa, and Stechmann and Cavalier-Smith placed this branch as the earliest among the eukaryotes that have the fusion. This root was later questioned when new data got published by Kim et al. (2006). One of the concerns raised by Kim et al. was that apusomonads, that have the DHFR-TS fusion, branch together with opisthokonts.
Figure 2. Malate synthase (MS) phylogenies before and after the addition of additional euglenid data. Most nodes within the clade of MS with eukaryotic origin are not statistically supported, the topology is not congruent with the species tree and euglenozoan taxons are represented by long branches (see Paper III for details). Therefore, the MS phylogeny with only four euglenozoans included will not be in conflict with inferring a MS-IL fusion at the base of euglenozoa. Two independent fusions represent the most parsimonious explanation when more euglenid MS sequences are used to calculate the topology.
Kim et al. demonstrates that the generation of more data is not only important to form new theories; it is also useful for testing what is already proposed. Other benefits with generating more data is that it can provide a set of sequences with higher diversity that can be used to build trees, which topology tend to be more well-supported if additional taxa and data is added (Graybeal, 1998). Missing data can be a problem when marker genes are collected for phylogenies based on multiple concatenated sequences, since single gene trees are used to find the correct copy of each marker. Phylogenies based on multiple concatenated genes, an approach also referred to as phylogenomics, are often used to infer the positions of lineages that are deeply rooted in the eukaryotic tree (Janouskovec et al., 2017; Lax et al., 2018; Strassert et al., 2019). Phylogenomics is likely needed as well to infer the position of the root in the eukaryotic tree. Currently there is no consensus for this position since different methods leads to different placements of the root (Derelle and Lang, 2012; Derelle et al., 2015; He et al., 2014; Katz et al., 2012; Rogozin et al., 2009; Wideman et al., 2013).

The best approach to make progress in finding the position for the true root in the eukaryotic tree is most likely to generate more genome and transcriptome data covering environmental protists.
Traditional protistology

Protistologists focusing on protist evolution have not had the technology to efficiently generate genome and transcriptome data until relatively recently. The first international protistology conference was held in 1961 (Vavra, 2018), which is 43 years before the release of the first commercial sequencing platform that allowed highly parallelized DNA sequencing of reads (Mardis, 2008).

Before phylogenomics researchers have mainly relied on phylogenies based on morphological characters (Cavalier-Smith, 1981, 1999), or used sequences from single genes (Leipe et al., 1993; Sogin et al., 1986). Unfortunately, neither phylogenies based on a single gene nor morphological characters can resolve relationships between species at the deeper nodes (Tekle et al., 2010; Yoon et al., 2008). Additionally, it can also be problematic to define species using morphological characters. If few usable morphological features are found in a taxonomic group, the number of valid species that can be inferred will be limited (Glotova et al., 2018). It is also common that incorrect taxonomic assignments, originally based on morphology, are identified when genetic data becomes available (Lax and Simpson, 2013; Shazib et al., 2016; Williams et al., 2011).

Besides studying protist morphology using different microscopy methods, a lot of effort has been put into culturing different species (Clark and Diamond, 2002; Coleman et al., 1976). Cultivation of protists can be a very challenging task if cells require conditions that are hard to mimic in the lab. For example, some protist requires the presence of other species in their environmental surroundings, or the presence of specific nutrients.

It is hard to estimate the fraction of protist diversity that could grow in the lab, with reasonable investment in time and resources, since failure to establish a culture does not mean that it is impossible to do so. However, with current technology, most species are probably not feasible to grow in the lab (Geisen et al., 2015; Keeling and Campo, 2017). A key benefit of successful cultivation is that it facilitates experiments that require many replicates. Furthermore, a culture makes it possible to extract a high amount of DNA, allowing direct genomic sequencing, and more convenient access to DNA template for amplification of certain genomic regions (Archibald et al., 2000; Kolisko et al., 2008). However, this can only be done if some information about the gene is already known, which is a requirement when designing PCR primers. Also most important of all, a culture will allow characterization of the genome and transcriptome with good coverage.
Amplicon sequencing from whole community

Genomic regions can also be amplified for a large number of species, directly from an environmental sample, i.e. without cultivation. For eukaryotes, the most popular genomic region for this purpose is the 18S rRNA gene, which is highly conserved among eukaryotes, and can be targeted with a universal primer pair (Hugerth et al., 2014; Lopez-Garcia et al., 2001). This approach has been hugely popular in the protist community and has uncovered several lineages such as marine stramenopiles (Massana et al., 2004a; Massana et al., 2004b) and marine alveolates (Lopez-Garcia et al., 2001; Moon-van der Staay et al., 2001). However, the drawback is that this approach only generates data for the 18S gene, which often does not provide enough information for constructing a well-supported and well-inferred phylogenetic tree, especially when it comes to resolving deep branches.

Several genes in the rRNA operon can now be sequenced in one read from whole communities of protists. This is possible due to emerging technologies based on long read sequencing, also allowing improved ability to place taxa in the tree (Jamy et al., 2019). This approach performed by Jamy et al. (2019) is a promising method to more effectively describe diversity, and at the same time calculate phylogenies with a more supported topology than what can be inferred from a sequence as long as a pair-end read.

Further improved taxon sampling can potentially contribute to even more supported phylogenies produced using this method, since the access to 28S rRNA sequences from several taxonomic groups is limited. However, even if studies targeting rRNA genes in whole samples has been a successful approach to detect new species (Hugerth et al., 2014; Jamy et al., 2019; Lopez-Garcia et al., 2001), this approach might miss species which differ significantly from what we already know, since we use what we already know to design the primers.

During my thesis work I have used 18S rRNA targeted primers to characterize the diversity in Valltjärn (a pond in the forest named “Stadsskogen” in Uppsala, Sweden), soil from different parts of Stadsskogen and samples from a peat field close to Uppsala. This data was used in Paper I to conclude that the Stentor polymorphus population in Valltjärn could be considered as monoclonal. Additionally I could also use this approach to identify the peat field north west of Uppsala as an interesting sampling site, since species clustering close to Malawimonas jakobiformis (Figure 3) could be found there. Malawimonas (O’Kelly and Nerad, 2007) is a genus of
evolutionary importance due to the potentially early branching position in the eukaryotic tree (Hampl et al., 2009).

Figure 3. Amplification of the V4 region in the 18S rRNA gene from a peat sample reveal taxa closely related to *Malawimonas*. The maximum likelihood tree is based on the V4 region in the 18S rRNA gene for the environmental protists. Both the 18S and 28S rRNA gene is included for all other taxa in the tree. Bootstrap support for the environmental protists that branch with *Malawimonas jakobiformis* is indicated.

Unfortunately the sampling efforts following the amplicon sequencing did not lead to any cDNA generation for these lineages branching with *Malawimonas*. The 18S rRNA sequences from the *Malawimonas*-related protists had low coverage in the sequencing run, which indicates they are present in low abundance and therefore will be challenging to isolate. They were found both in soil at the surface level and in wet peat about 15 cm below surface level, with the latter environment likely containing low concentrations of oxygen. If this is an anaerobic species it will be an even more interesting finding, since species in the *Malawimonas* genus are aerobic (Heiss et al., 2018). However, this would need to be confirmed by more sampling efforts and experimental work.
Metagenomics

It is difficult to infer well-supported phylogenies or to identify biological processes going on in the environment exclusively from 18S rRNA gene sequences. Metagenomics allows for both the identification of which metabolic capacity species in the environment have, and phylogenomic analysis, without the need for cultivation efforts. A workflow for metagenomics typically starts with extraction of DNA from an environmental sample, followed by sequencing library preparation. Once the sequencing data is generated reads are assembled into assembly contigs, which typically represent parts of the genomes from many different species. To separate these contigs into separate datasets, where each dataset correspond to genomic data from a single organism, a computational step has to be performed that is often referred to as “binning”.

During binning, computational methods are used to identify properties that contigs have in common, which will group contigs from the same genome together into a “bin”. Examples of properties that can be used to bin the assembled contigs are compositional nucleotide sequence differences (Thomas et al., 2012) and differences in coverage (Hug, 2018).

A drawback with the metagenomic approach, for protistologists, is that prokaryotes are generally present in higher abundance in the environment than eukaryotes. Further, eukaryotic genomes are larger and have relatively low density of genes in their genomes compared to prokaryotic genomes (Cuvelier et al., 2010). In addition to the extra complexity of analyzing reads from multiple species, protist genomes can be difficult to assemble (Keeling et al., 2014). Despite these challenges, there are examples of successful metagenomic studies of protists in the literature (Figure 4). For example, a metagenomic approach can make sense if there is a way to enrich the eukaryotic species that you are interested in, which can be achieved using sorting with flow cytometry (Cuvelier et al., 2010). It has also been shown that the conventional workflow for metagenomics can be improved for acquiring eukaryotic genomes by adding an extra step in the computational pipeline to separate the eukaryotic from prokaryotic contigs (West et al., 2018).
Figure 4. Steps in a metagenomic workflow and options to improve the completeness for eukaryotic bins. A conventional metagenomic workflow typically leads to fragmented genomes in the eukaryotic bins. Enrichment of eukaryotes prior to sequencing can be done to improve eukaryotic genome recovery. High completeness for eukaryotic genomes can also be achieved by predicting which assembled contigs are eukaryotic, and then separately bin the contigs predicted as eukaryotic.

Emerging technologies such as long read sequencing can improve the metagenomic assembly by increasing the size of the contigs, which further can lead to binned genomes with higher completeness. Despite these possible technological and computational improvements for obtaining eukaryotes from a metagenomics workflow, the species of interest could be in very low abundance, which is a limitation for using metagenomics to find novel protist lineages.
Metatranscriptomics

While both amplicon sequencing and metagenomics can give an indication of which species you can find in an environment, metatranscriptomics is another possibility for assessing species composition. In metatranscriptomics, cDNA is reverse transcribed from total RNA extracted from a sample using random hexamers as primers. Since around 85% of the RNA in a cell is rRNA (Karpinets et al., 2006), generated cDNA will mostly cover the rRNA operon, allowing for eukaryotic species identification based on the 18S rRNA gene.

The main difference that should be considered between metatranscriptomics and amplicon sequencing is that, because of the lower stability of RNA compared to DNA, metatranscriptomics gives a better view of which species are currently active in an environment (Geisen et al., 2015). Also, metatranscriptomics does not require the design of primers since random hexamers can be used, which possibly allow for detection of novel taxa that primers directed to conserved regions would miss. However, if the goal is to specifically characterize protists there will be a lot of sequencing data generated for prokaryotes when using metatranscriptomics, compared to amplicon sequencing with designed primers. Another limitation of metatranscriptomics when random hexamer primers are used, is that the assembly of large numbers of highly conserved rRNA sequences can be difficult due to sequence similarities between species in a sample (Yuan et al., 2015).

While metagenomics shows the genetic diversity present in an environment, and therefore to some extent gives a picture of microbial community processes, metatranscriptomics is even more connected to ongoing processes in an environment since it captures the genes that are being actively expressed. When adding a poly-A selection step to enrich for mRNA, metatranscriptomics allows researchers to detect functional responses among the eukaryotes in a community when conditions in the environment change. Poly-A enriched metatranscriptomes also mainly show the most highly transcribed genes in a community and therefore give a better insight than metagenomes into which processes are active (Radax et al., 2012). Additionally, it is easier to find eukaryotic genes in a metatranscriptome than in a metagenome since the non-coding regions will not be sequenced.

The main problem with metatranscriptomics is that, while new species can be found, it will most likely be through the detection of novel rRNA sequences. These sequences will not have enough phylogenetic signal to infer a supported phylogenetic tree of 18S rRNA gene sequences if there are no
close relatives known. Also, the binning process is challenging in metatranscriptomics due to the short contigs generated. Metatranscriptomics is therefore unsuitable for the generation of data from novel species for the purpose of establishing a multi-gene dataset for phylogenomic analyses.
Single-cell genomics

Like metagenomics, single-cell genomics (SCG) is an option for generating genomic data from environmental protists without the need for cultivation. However, SCG differs fundamentally from metagenomics, since SCG uses a single cell as input instead of a whole community. To use this method one first needs to isolate individual eukaryotic cells, which can be done either by sorting cells using fluorescence-activated cell sorting (FACS) or manual hand-picking. The genome of the single cell can be amplified after lysis, typically using multiple displacement amplification (MDA) (Nair et al., 2014; Roy et al., 2014; Yoon et al., 2011).

![Diagram of multiple displacement amplification](image)

**Figure 5.** First two cycles of multiple displacement amplification. A genome amplified by MDA tends to have uneven coverage because of the multi-branching structure generated by the Φ29 DNA polymerase.
Random hexamers are often used as primers during MDA, together with the Φ29 DNA polymerase, which amplifies the template DNA into a branching network (Figure 5). After purifying the MDA reaction, the resulting amplified DNA it is often checked with some screening method, e.g. the 18S rRNA gene can be amplified and sequenced. This allows a targeted approach where only certain libraries are prepared for sequencing.

One drawback with MDA is that genome recovery can be low (Table 1). Besides low genome recovery for SCG, another potential problem for eukaryotes is identifying protein-coding genes. Translation of every open reading frame (ORF) in a genome can lead to over a million translated sequences, even if all ORFs translated into less than 80 amino acids are discarded. The high number of ORFs is found since genome data consist of long continuous sequences, unlike transcriptomes where only a few long ORFs exist per contig.

Table 1. Genome recovery achieved when the MDA product of single protist cells were sequenced.

<table>
<thead>
<tr>
<th>Taxonomic affiliation</th>
<th>Genome recovery</th>
<th>Genome recovery reference</th>
<th>Cell size (μm)</th>
<th>Cell size reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Picobiliphyte</td>
<td>Most complete &gt; 50%</td>
<td>Yoon et al. (2011)</td>
<td>2.5-3.8</td>
<td>Yoon et al. (2011)</td>
</tr>
<tr>
<td>MAST-4</td>
<td>Most complete &gt; 70%</td>
<td>Roy et al. (2014)</td>
<td>2–5</td>
<td>Roy et al. (2014)</td>
</tr>
<tr>
<td>*Plasmodium vivax</td>
<td>Mean 32% (8–49%)</td>
<td>Nair et al. (2014)</td>
<td>10</td>
<td>Mikolajczak et al. (2015)</td>
</tr>
<tr>
<td><em>Cryptosporidium</em></td>
<td>Mean 73% (50–92%)</td>
<td>Troell et al. (2016)</td>
<td>3-6</td>
<td>Casemore et al. (1985)</td>
</tr>
<tr>
<td>Diplonemids</td>
<td>Most complete 9%</td>
<td>Gawryluk et al. (2016)</td>
<td>7–25</td>
<td>Gawryluk et al. (2016)</td>
</tr>
<tr>
<td>MAST-4</td>
<td>Mean 17% (2–42%)</td>
<td>Mangot et al. (2017)</td>
<td>2–5</td>
<td>Roy et al. (2014)</td>
</tr>
<tr>
<td>Monosiga brevicollis</td>
<td>Mean 19% (6–36%)</td>
<td>Lopez-Escardo et al. (2017)</td>
<td>3</td>
<td>King et al. (2008)</td>
</tr>
</tbody>
</table>

*Cryptosporidium oocysts contain four sporozoites, each sporozoite has a haploid genome
A high number of translated sequences will lead to computational bottlenecks during further data analysis. Instead of translating every possible frame, ORFs can be predicted with sophisticated mathematical models (Mathe et al., 2002; Stanke and Waack, 2003). Unfortunately, prediction of genes is not straightforward, and commonly applied predictors are dependent on data from known genes (Stanke and Waack, 2003). Dependency on known data could be a problem if all available training sets are represented by species that are distantly related to the taxa whose genes needs to be identified. Additionally, some protists reorder exons (Greslin et al., 1989), or merge exons from separate RNA sequences (Sutton and Boothroyd, 1986) to form the mRNA that will finally be translated. Reordering or merging exons from different transcripts could potentially lead to a genome architecture that is challenging to predict genes from.
Single-cell RNA sequencing

Another possible single-cell based approach is single-cell RNA sequencing (scRNAseq), the method used throughout this thesis. One commonly applied scRNAseq protocol is Smart-seq2 (Picelli et al., 2014). In Smart-seq2 the poly-A tails of mRNA are targeted with oligo-dT primers that carry a 5’ anchor sequence, i.e. a known sequence that can be targeted with primers for amplification. The second anchor sequence is added via template-switching (Zhu et al., 2001), which will allow amplification of full-length mRNA (Figure 6). The amplified cDNA library can, after purification, be processed with the preferred sequencing library preparation protocol to allow generation of transcriptome data.

A drawback with scRNAseq is that RNA is more unstable than DNA, which can complicate some parts of the workflow. This prevents the user from applying harsh lysis conditions that would destroy RNA. Lysis seems to be one of the main problems to get scRNAseq to work for many protist species when using Smart-seq2 as a starting point. This was particularly pronounced when generating single-cell transcriptomes for smaller protist cells during the beginning of my thesis work. After I shifted focus to improving the lysis conditions I could finally generate cDNA libraries from smaller protists with a much higher success rate than before, which allowed me to generate data for Paper III.

An obvious difference between scRNAseq and SCG is that the former can only cover the part of the genome that is transcribed, which makes SCG preferable if the user is also interested in the non-transcribed parts of the genome. However, if the user is interested only in protein-coding genes, scRNAseq might be preferred. In scRNAseq, the sequencing capacity will not be used on repetitive intergenic regions. Instead, the focus will be on sequencing mRNA, which will make the process of assembly and identifying genes much easier.

However, scRNAseq generally leads to lower transcript recovery for smaller protists, which is a drawback. Using scRNAseq, Liu et al. (2017) only recovered about 15% of the transcripts for a dinoflagellate with 15 μm cell size and the transcript recovery was even lower, around 3%, for a haptophyte with a cell size of 8 μm. Meanwhile, Kolisko et al. (2014) achieved around 80% to 100% transcripts recovered, comparing scRNAseq to bulk, for four different ciliate species with cell lengths around 100 μm to 500 μm. This difference suggests that lower transcript recovery should be expected for smaller cells, which is also observed in Paper III.
Figure 6. Generation of full-length cDNA via template switching. The amplified cDNA libraries described in this thesis are made according to the following steps: The cell is first lysed so oligo-dT primers, with a 5’ anchor sequence, can anneal to the poly-A tails of mRNA. In the following step, a reverse transcriptase is used that have a preference for adding a few nontemplate cytosine to the 3’ end of the cDNA after completing reverse transcription of the mRNA molecule. An oligo, that carries two riboguanosines and a locked nucleic acid deoxyguanosine at the terminal end, can then anneal to the nontemplate cytosines at the 3’ end of the cDNA. The oligo that anneal to the nontemplate cytosines allow for reverse transcription to add a second anchor sequence, resulting in a cDNA molecule ready for amplification since it is tagged in both ends.

Possibly the low transcript recovery achieved by Liu et al. (2017) could be improved if the protocol presented in Paper IV would be used. One major difference in the workflow between what Liu et al. used and what was applied in Paper IV is that I did not use any RNA extraction step, where I expect many low copy transcripts will be lost. Additionally, Liu et al. would most likely improve their transcript recovery if they would use the freeze thaw cycles as an extra lysis step that was tested in Paper IV. We could detect 77% of the protein coding genes in the single-cell transcriptomes generated for *Giardia intestinalis* in Paper IV. Since *G. intestinalis* has a cell size around 10 μm (Horlock-Roberts et al., 2017), this result shows that it is possible to achieve high transcript recovery with scRNAseq on small cells.
Which method should be used to expand the known diversity in the eukaryotic tree?

It is important to emphasize that the methods discussed so far in this thesis complement each other for different applications. For some situations there could be multiple suitable methods, but one option is better than the others. Based on the long term goals I have had as a PhD student, it is most interesting for me to answer which method that is most suited for expanding the known diversity in the eukaryotic tree. When I started in 2014 there were no articles published about scRNAseq for protists, and very few in general that used any cultivation-independent genomic or transcriptomic approaches on protists. Therefore, at that time, it was harder to say which method that should be preferred to uncover unknown diversity. It will be easier to answer this question today, since several studies have been published by now that applies different cultivation-independent methods.

One could argue that amplification of rRNA genes from environmental samples would be the best way to explore eukaryotic diversity, and to some extent that might be true. However, there is not much that can be done with such data and it is still dependent on a primer design, which could prevent the user from the most novel discoveries. Ideally, the method used should cover the gene content of the protists, which will allow inferring the phylogenetic position even though close relatives are missing. Due to current limitations in sequencing depths the “meta”-approaches are unfortunately not recommended as a general approach to expand eukaryotic diversity. Metagenomics could be useful in certain situations, e.g. if you have detected a very interesting species in the environment with amplicon sequencing, but you have repeatedly failed to isolate this species manually. However, there are few successful metagenomic studies published where protists are targeted.

The few successful metagenomic studies so far that have targeted protists tend to be based on prior knowledge, such as specific sorting conditions (Cuvelier et al., 2010) or using a training set of already known eukaryotic contigs (West et al., 2018). Depending on prior knowledge could lead to a biased methodology, both experimentally and computationally, that prevents finding cells or genomes that are considerably different from what we know. Finally, the metagenomic approach will likely generate data that is dominated firstly by prokaryotes and secondly by highly abundant protists that are already known.
Based on the technology we have today, scRNAseq and SCG seems to be the most suitable choices for expanding the known diversity in the eukaryotic tree. The key difference between single-cell methods and meta-omics approaches are that the methods that are single-cell based can be used in a more targeted way. Specific cells can be targeted by hand-selection or single cells can be sorted by FACS based on spectroscopic properties. Therefore a protistologist that wants to genomically explore new protist lineages is likely to face the choice: “single-cell RNA sequencing or single-cell genomics?”.

The generated data suggests that both scRNAseq and SCG have pros and cons (Table 1, Paper III). For larger cells it is well documented that scRNAseq works good (Hines et al., 2018; Kolisko et al., 2014; Onsbring et al., 2018). Meanwhile the literature is lacking the same documentation as scRNAseq for SCG with bigger cells. In Table 1 we can notice that these SCG studies target small cells, mainly below 10 μm in length. The only study that performs SCG on larger cells than 10 μm estimated a 9% genome recovery for their best genome (Gawryluk et al., 2016). Additionally, it should be noted that some studies presented in Table 1 have only shared information regarding genome recovery for their best assembly, from an experiment where many cells were sorted into multi-well plates (Roy et al., 2014; Yoon et al., 2011).

Possibly SCG does not work well for larger cells and high completeness is only achieved for a few replicates if a high number of cells are processed. Meanwhile, larger cells generally lead to better transcriptomes when scRNAseq is used (Paper III) and for the very large cells, close to bulk performance is achieved (Kolisko et al., 2014). Therefore, unless non-coding regions are of specific interest, scRNAseq should be the preferred choice for larger cells. SCG could be considered for the smaller cells (below 10 μm), since this approach does not seem to lose performance when cells get smaller like scRNAseq.

It is not clear which methods that would achieve the highest coverage of the coding capacity for cells in the range 10 μm up to 40 μm due to the lack of published SCG studies on such cell sizes. However, if there are only a few cells that are going to be analyzed, both methods could be applied. If a high number of different cells are going to be processed, the scRNAseq workflow offers several advantages. Using transcriptomics will make the detection of genes easier and allow a more convenient assembly process since repetitive regions are not present. It will also be easier to quality control cDNA libraries than MDA products.

Using the scRNAseq protocol with all modifications tested in Paper IV will offer an approach where a negative reaction will have a DNA concentration below the detection limit for commonly applied DNA quantification methods. Further a fragment length analysis can be done to check the quality of the cDNA. Meanwhile in MDA, all reactions will contain a high DNA concentration and the user will usually be forced to assay if a protist genome were amplified with quantitative PCR, which require primers designed based
on what is already known, which should be avoided when searching for novel lineages.
How should scRNAseq be applied to increase the known diversity in the eukaryotic tree?

When the options to characterize unknown protist diversity are reviewed it seems like scRNAseq is one of the more important methods to know how to use. The generation of a high number of transcriptomes in Paper III show that data analysis, such as assembly and gene annotation, can conveniently be done in a standardized way for a wide diversity of cells. It is also shown in Paper III that novel lineages can be found without travelling to remote environments.

Paper III also shows that novel protists lineages can be found by manual methods, without carefully studying the morphology. Before we initiated the study in Paper III we did not know whether we would sequence novel diversity and how much overlap there would be between samples. However, now when we better know what to expect from this approach it is also easier to identify what we can improve to allow covering more diversity with the same investment of resources. For example, it should be expected that a significant amount of sequencing data are generated for taxa that are not novel.

An extra PCR step could prevent sequencing of uninteresting or redundant taxa, in this step an aliquot from each cDNA libraries should be amplified with 18S rRNA gene directed primers, followed by Sanger sequencing. Implementing a Sanger sequencing step would add a cost of consumables around 6.7 USD per sample. At the same time, the consumable cost for sequencing library preparation and sequencing that potentially could be saved, is around 125 USD per sample (Table 2). Based on these estimations 5.4%, or more, of the samples should be discarded based on the Sanger sequencing to motivate the implementation of this step to compensate for the cost of consumables. Given the degree of redundancy and high number of algae that were sequenced in Paper III it is possible that we would have saved consumable cost by adding the Sanger sequencing step.

Besides selection of samples post cDNA synthesis, changes could also be done to how cells are isolated. For example it is possible to pick cells under a higher magnification to allow more accurate identification of cells, which leads to the isolation of fewer uninteresting cells. However hand-selecting cells with an objective higher than 10x will not allow the usage of larger containers as a source to pick from. The vision will be narrowed down to a smaller area with higher magnification, which makes cells harder to find and manipulate with the pipette. Additionally, the smaller container will contain less cells to pick, so fewer cells can be isolated each time a new sample is put
under the microscope. Finally, the washing technique used in Paper III will not be possible to do in a small container with liquid, since isolating cells from a smaller container will force the user to apply extra transfer steps required for washing the cell so it is clean enough for single-cell analysis. Adding more transfer steps of the cell would strongly affect the isolation time in a negative way since many protists tend to stick to the pipette tip and not leave when the liquid in the pipette tip is dispensed.

Table 2. Cost for consumables associated with Sanger and Illumina sequencing. Price information that is from 2018, or older, has been adjusted with a 2% increase per year to account for inflation. For amplification, purification and the Nextera XT workflow an additional 10% is added to the cost to account for spill.

<table>
<thead>
<tr>
<th>Step</th>
<th>Cost per sample (USD)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consumables for amplification</td>
<td>1.3</td>
<td>Schlatter et al. (2015)</td>
</tr>
<tr>
<td>Consumables for purification</td>
<td>1.5</td>
<td>Schlatter et al. (2015) and ZAGENO</td>
</tr>
<tr>
<td>Sanger sequencing</td>
<td>3.9</td>
<td>Eurofins</td>
</tr>
<tr>
<td>Sanger workflow</td>
<td>6.7</td>
<td></td>
</tr>
<tr>
<td>Nextera XT</td>
<td>49</td>
<td>Schlatter et al. (2015) and SNP&amp;SEQ**</td>
</tr>
<tr>
<td>HiSeq Rapid mode PE 250 bp*</td>
<td>76</td>
<td>SNP&amp;SEQ**</td>
</tr>
<tr>
<td><strong>Illumina sequencing</strong></td>
<td>125</td>
<td></td>
</tr>
</tbody>
</table>

*Assuming 96 samples sharing flow cell
**SNP&SEQ Technology Platform, Uppsala University (Sweden)

Conservatively we can assume that each of the changes to the picking procedure from Paper III, that has to be made to use a higher magnification, will negatively impact the cell isolation rate with a factor of two. Then the narrowed view, fewer isolations per prepared sample and higher losses during washing will decrease the rate of cell isolation with a factor of eight. Meanwhile if we take the result in Paper III into consideration, it would only be optimal to discard around 25% of these cells. It is also unlikely that all cells are correctly identified so all uninteresting protists, but not any of the interesting ones, are discarded before cDNA synthesis in Paper III. If the goal is to explore protist diversity without targeting any specific group, only discarding 25% of the cells in Paper III does not seem like an attractive trade-off if the picking procedure would take more than eight times longer.

Isolating cells manually is a major bottleneck in the workflow with respect to time. Another possibility would be exchanging the manual cell picking with FACS. This would allow hundreds of cells to be sorted in just one day. During this thesis work I have put a lot of effort into incorporating FACS into my
scRNAseq workflow. However, sorting protists from an environmental sample is challenging, and our efforts ended in a high fraction of negative reactions after cDNA synthesis.

From our best plate we managed to generate 4 transcriptomes out of 48 wells, which corresponds to only generating positive reactions for 8% of the wells. High amount of failed sorting events has previously been reported by Sieracki et al. (2019), which achieves a yield of 10.8% for 9135 sorted wells. However, higher yield has been reported by Yoon et al. (2011) that could confirm 109 protists sorted into three 96-well plates (i.e. ~45% yield, assuming two columns in each plate were used as controls). Heywood et al. (2011) also achieved a high yield and could confirm amplified protist genomes in 36% of their 242 wells.

Which sample is sorted seems to be important for the outcome, because staff from the same institution achieve yields varying around 11-45% (Heywood et al., 2011; Sieracki et al., 2019; Yoon et al., 2011). Since I have tried sorting from sediment samples, and not oceanic water column like they did in the studies from Bigelow Laboratory (Heywood et al., 2011; Sieracki et al., 2019; Yoon et al., 2011), it is possible that lower yield during the sorting should be expected. Additionally, I have detected sorted protists by fragment length analysis of generated cDNA libraries, which differs from the other studies that use 18S rRNA amplification (Heywood et al., 2011; Sieracki et al., 2019; Yoon et al., 2011). Since protists can be in metabolically inactive life-cycles where the amount of mRNA in the cell is very low (Faghiri and Widmer, 2011), it is possible that a higher yield will be measured if the genome is targeted with PCR.

It should be tested if metabolically inactive cells affect the yield during sorting for scRNAseq, which would allow more informed decisions when choosing between SCG and scRNAseq in a workflow together with FACS. However, it should be noted that the reported FACS yield around 11-45% (Heywood et al., 2011; Sieracki et al., 2019; Yoon et al., 2011) could be underestimated for several reasons. If the FACS yield is based on amplification of the 18S rRNA gene, then successfully sorted cells might have an 18S rRNA region that the primers failed to amplify. Additionally, some protist genomes can have a low 18S rRNA gene copy number (Gong and Marchetti, 2019), which could be absent in the amplified single-cell genome that is associated with poor genome recovery.

The four transcriptomes that I did generate from cells sorted by FACS were not discouraging in terms of the diversity they represented. I found one alveolate, one green algae, one cryptophyte and one amoeba. For the amoeba, In a BLASTN search against the NCBI nt database, the top hit was Schoutedamoeba minuta (van Wichelen et al., 2016) with a 97% 18S rRNA gene identity. Apart from the S. minuta 18S rRNA sequence, the other sequences identified by BLASTN search were at 90% identity, indicating that the amoeba we sorted belong to a poorly characterized part of the eukaryotic tree.
Most likely a workflow using FACS to isolate cells, followed by scRNAseq, would save money spent on consumables by adding a Sanger sequencing step, since more than 5.4% uninteresting transcriptomes should be expected. Besides an 18S rRNA gene sequencing step, I think it is important to use the scRNAseq protocol from Paper IV to easier identify the wells where no cDNA was generated, which would reduce the cost of other downstream quality controls significantly when a high number of failed sorting events are expected.
Future perspectives

I think protistologists in the future should aim for diversifying the approaches used to find new species. Many protistologists put a lot of effort into microscopy and carefully selecting cells, which they also should continue doing. However, it would be beneficial for the scientific community if the search for novel lineages is done with a wider spectrum of approaches. A robust approach for sorting protists from environmental samples should be established to allow covering a lot of protist diversity in a short period of time. Potentially this is not too far away in time since people on conferences often claim that it should be possible with a much higher accuracy than I report, however despite communication with these researchers we could never reach the FACS yield they said were possible.

In a more distant future, performing the scRNAseq in droplets using microfluidics could circumvent the challenges associated with isolating protists via FACS. The positive aspect with microfluidics is that thousands of reactions (Macosko et al., 2015) can be run in parallel and if particles that are not protists enter several droplets it will not be a problem since very low volumes are used. To establish a robust scRNAseq protocol based on microfluidics, that allow the generation of sequence data covering full length mRNA, two technological challenges need to be solved: 1) Cell lysis and 2) Long read sequencing.

Based on previous scRNAseq efforts with only Triton X-100 as lysis, freeze thaw cycles contributed to that a higher number of transcriptomes could be generated in Paper III. Applying freeze thaw cycles to an emulsion of droplets will lead to that droplets merge with others. Merged droplets will prevent separation of sequencing data into separate datasets corresponding to individual cells. Further, the emulsion has to be broken at the end for purification of the cDNA. At this step all cDNA are tagged with oligo-dT primers that carries a unique barcode for each cell, allowing separation of the data during the bioinformatic workflow. In the next step, if an Illumina sequencing platform is used, the cDNA will be cleaved into pieces during tagmentation, separating parts of the transcript from the bar code needed to de-multiplex. Potentially, the tagmented cDNA could then be separated into two reactions, to allow specific amplification the fragments containing the barcode and standard amplification of the tagmented cDNA library.

Specific amplification of the barcode increase coverage on the part of the transcript needed to de-multiplex the metatranscriptomic-like assembly. The potential of this workflow is very unclear until it has been tested. Potentially
it could be better to wait for long read sequencing technologies to improve base call accuracy, which would allow accurate sequencing of whole transcripts without cutting them into pieces (Figure 7).

**Figure 7.** Droplet based workflow allowing the generation of thousands of transcriptomes in parallel. Single protist cells are first encapsulated together with barcoded beads, followed by reverse transcription that attaches cDNA to the beads. When attached cDNA are released and the emulsion is broken, both amplification and purification can follow. Finally, long read sequencing can cover the whole cDNA molecule and the barcode, allowing the user to de-multiplex the data into a high number of single-cell transcriptomes.

In an even more distant future, protist metagenomics could be a very powerful approach, if data is cheaper and long read sequencing have improved. Effective metagenomics for protists would allow researchers to sequence genomes of species that are very hard to separate from the environment, such as soil protists. Therefore, a sequencing technology that generates long reads, and enough data to cover protists with low relative abundance, would allow highly effective exploration of eukaryotic diversity.
Ciliates of the genus *Stentor* are well known for their ability to regenerate drastic cellular wounds. Despite that *Stentors* wound healing ability has been studied for over a century (Nussbaum, 1884), only one molecular study covering one protein involved in regeneration has been described in the literature so far (Slabodnick et al., 2014). To allow identification of a high number of genes involved in *Stentor* regeneration we performed RNA sequencing of individual cell fragments from *Stentor polymorphus*.

The cDNA libraries we sequenced were generated from both anterior and posterior fragments from *Stentor*, allowing insight into the differences in regenerating the oral apparatus, through which *Stentor* feeds, and regeneration of the foothold. Additionally we generated cDNA from different time points to characterize the genes activated at different stages of regeneration.

Biologically relevant clustering of the different replicates can be done based on gene expression levels, which indicates that RNAseq can be done in fractions of *S. polymorphus* cells with a meaningful result. At 10 min post cell-split, the expression profiles between anterior and posterior fragments did not allow us to cluster the data in separate cluster depending on which part of the cell that was lost. The similar expression profile at 10 min indicates that the regenerating cell has not yet identified which part that is missing. In total we found 10,682 upregulated genes, among them 6,433 were only seen upregulated in posterior fragments and 720 genes were exclusively upregulated in anterior fragments. The higher number up upregulated genes specific to the posterior fragments indicates that regeneration of the oral apparatus is transcriptionally more complex for *S. polymorphus* than regenerating the holdfast.

The upregulated genes were mainly involved in signaling, microtubule-based movement and cell division, while the downregulated genes were mainly house-keeping genes. Several expanded kinesin groups that are known for involvement in the cell cycle were identified as upregulated during regeneration, which was not surprising since similarities between cell repair in *Stentor* and mitosis has previously been observed (Guttes and Guttes,
1959). More surprisingly, we could observe several proteins specific to meiosis to be upregulated during regeneration.

Paper II. Molecular Investigation of the Ciliate *Spirostomum semivirescens*, with First Transcriptome and New Geographical Records

The ciliate genus *Spirostomum* consists of eight species, where most have been molecularly characterized (Boscaro et al., 2014; Shazib et al., 2016). However, *Spirostomum semivirescens* has so far been absent from 18S rRNA gene sequencing efforts. In this study we generate single-cell transcriptome data for *S. semivirescens* originating from different sampling sites in both England and Sweden. This will allow us to place *S. semivirescens* in a phylogenetic tree together with other taxa from the same genus, and compare the ciliate cells from different sampling sites. Further we can get an insight into metabolic pathways encoded by the organism and explore other general properties of the genes expressed.

Closest relative to *S. semivirescens* is *S. minus* from clade previously referred to as “minus clade 2” (Boscaro et al., 2014). Further the endosymbiotic algae species are a close relative to *Chlorella vulgaris*. In the transcriptome we found a gene coding for the rhodoquinone biosynthesis protein RquA, indicating that *Spirostomum* uses a rhodoquinol dependent fumarate pathway for respiration under anaerobic conditions. Despite no cyst-form observed, this fragile species have managed to cross long distances and can be found in both England and Sweden.

We found that *S. semivirescens* is using TAA, TAG and TGA as stop codons. However, TGA was rarely used among the highly expressed genes and we could observe a relationship between TGA stop codon frequency and expression level. Interestingly, other ciliates from the class Heterotricha, which *Spirostomum* belongs to, have also reassigned their codon usage (Lozupone et al., 2001) or could be reassigning (Heaphy et al., 2016). Based on the codon usage among closely related ciliates and the observed codon bias, it is possible that *Spirostomum* could be at an early stage of codon reassignment.
Paper III. Single-cell transcriptomics expands sampled protist diversity and provides insights into niche adaptation

A vast majority of the diversity among eukaryotes are represented by protists (del Campo et al., 2014). Even though protists represents most of the diversity in their domain, the sequencing efforts for eukaryotes have been focused towards plants, animals and fungi (Sibbald and Archibald, 2017). The focus on multicellular organisms has generated gaps in our knowledge about eukaryotic diversity.

We manually isolated a high number protist from different sampling sites around Uppsala for single-cell transcriptomics, to expand the known protist diversity. Generating single-cell transcriptomes for the hand-selected environmental protists will both contribute to new knowledge and to get an estimation about what to expect if the same approach is applied in a high throughout setting.

We sequenced 124 transcriptomes, which covered protist diversity from seven of the major clades in the eukaryotic tree. Many of these transcriptomes represented lineages with an 18S rRNA gene that had 90% similarity, or less, to the closest BLASTN hit against NCBI nt database. Therefore, several of the protists found corresponded to distant relatives to what has earlier been observed, even at 18S rRNA gene level. Among the novel lineages there were non-photosynthetic euglenids, free-living prokinetoplastids, katablepharids, ciliates and metamonads.

The generated transcriptomes allowed us to get an insight into the differences between closely related protists that lives fundamentally different life-styles. Such observations include a wider diversity of biosynthesis pathways in osmotrophic euglenids compared to other non-photosynthetic euglenids. Using our ciliate data from the class Litostomatea also allowed us to find that the rumen-dwelling ciliate *Entodinium caudatum* have an increased capacity to metabolize carbohydrates and has lost pathways involved in glycine and serine metabolism. Additionally, we could find that prokinetoplastids have proteins needed to carry out the glyoxylate cycle, which is missing from metakinetoplastids. Further, the proteins in the glyoxylate cycle have gone through evolutionary events such as fusions and transfer among species in euglenozoa.

Our experiment indicates that establishing a high throughput workflow including faster cell isolation and parallelized cDNA synthesis in multi well plates, could further fill in the gaps of the eukaryotic tree and provide more insight into eukaryotic evolution.
Paper IV. An efficient single-cell transcriptomics workflow to assess protist diversity and lifestyle

Single-cell transcriptomics is a useful approach to generate gene content data while avoiding tedious cultivation of protists (Kolisko et al., 2014; Paper III). However, multiple studies report lower performance for single-cell RNA sequencing of smaller cells (Liu et al., 2017; Paper III). Further the Smart-seq2 protocol (Picelli et al., 2014) has a lysis optimized for mammalian cells and there are environmental protists that might not lyse using such conditions.

We tested six different Smart-seq2 modifications that potentially could improve the cDNA synthesis from an input of cells with low RNA content. All these six modifications included freeze thaw cycles for improved lysis. Five of these six modified protocols contained all or one of the following modifications: 1) All volumes used in Smart-seq2 were reduced to half. 2) All primers were 5’ biotinylated. 3) Alternative beads prepared in-house used for purification of cDNA. 4) Concentration of oligo-dT primers reduced by 60%.

All tests were done using single cells of *Giardia intestinalis*, which represents a relatively small protist with a size about 10 μm (Horlock-Roberts et al., 2017). Since the genome of *G. intestinalis* is sequenced this allowed us to collect statistics for gene discovery and coverage. We also generated *de novo* assemblies, using different amounts of data, to test how much data that is needed to find marker genes suitable to concatenate multiple aligned sequences for phylogenomics.

The performance of Smart-seq2 improved by adding freeze-thaw cycles to the workflow, which lead to higher transcript recovery. Most modifications tested did not significantly change the number of genes that were detected. However, when all volumes from Smart-seq2 where reduced to half we detected significantly fewer genes. The *de novo* assemblies indicated that generation of more sequencing data beyond 150 Mbp will not lead to more conserved marker genes in the assembly. The version of Smart-seq2 where only freeze thaw cycles were added lead to recovering significantly more conserved marker genes than four other procedures (standard Smart-seq2, the primers where 5’ biotinylated, concentration of oligo-dT was reduced or when all six modifications tested in this study were applied).

We noted that there were almost no primer dimers when all six modifications of Smart-seq2 that were tested in this study were applied all together at once. This protocol with all changes is still useful, even though we did not see any improved performance, since failed reactions can be identified based on low DNA concentration, making the user less dependent on fragment length analysis. Surprisingly, the version of the protocol that performed best also had the highest amount of primer dimer, which shows that smaller fragments in a cDNA library might not necessarily be a problem before tagmentation and sequencing.
Our experiment shows that some of our changes might not lead to an improvement of gene discovery, but neither do they decrease performance significantly. Meanwhile the tested modifications of Smart-seq2 can save both cost and time and can therefore be considered in a workflow where a high number of single-cell transcriptomes of protists are generated.


I denna avhandling har RNA-sekvensering från enskilda celler använts för att generera transkriptomdata. Fördelen med att studera ett transkriptom, i stället för genom, är att transkriptom endast täcker genuttrycket. Andra delar av genomet, som inte uttrycks, kan försvåra analys av data och bland annat göra de protein-kodande generna svårare att identifiera.


Att kunna sekvsera transkriptom från små protister används i tredje kapitlet där många tidigare okända protister som saknar kända närbesläktade arter analyseras. Dessa tidigare okända protister kunde hittas i vattendrag inom Uppsala kommun. Bland de nyupptäckta arterna fanns frilevande prokinetoplastids, tidigare har samtliga prokinetoplastids som studerats varit beroende av en värd. I studien genererades även flera transkriptom av euglenids, bland dessa transkriptom kunde vi se att euglenids från genuset *Distigma* hade en väldigt bred metabolisk kapacitet. Vissa gener med en funktion inblandad i metaboliska processer visade sig endast kunna uttryckas av *Digstigma* och inte i någon annan känd art ur gruppen euglenozoa, som *Distigma* tillhör. Vidare studier av metaboliska processer uttryckta av *Distigma* avslöjade också två gener inblandade i nedbrytning av kolhydrater, vilka framstod som att de var horisontellt överförda från prokaryoter.

I sista kapitlet beskrivs hur RNA-sekvseringmetoden jag har använt bör utföras på protister om transkriptomdata ska genereras i stor skala. Om nedfrysning-upptinings-cykler används, innan cDNA-syntesen startar, ökar mängden gener som kan kartläggas i ett transkriptom. Förutom nedfrysning-upptinings-cykler så testades även fyra andra variationer av Smart-seq2, som är ett protokoll för enkelcells-RNA-sekvsering. De fyra ändringarna av Smart-seq2 var: 1) Addering av biotin till 5’ änden på primrar. 2) Halvera alla volymer. 3) Minska oligo-dT-koncentrationen med 60%. 4) Striktare rening av amplifierat cDNA.

Om de fyra ändringar av Smart-seq2 som testades i sista kapitlet, tillämpas på samma gång som nedfrysning-upptinings-cykler läggs till, så blir enkelcells-RNA-sekvseringsprocessen mer tid- och kostnadseffektiv. Tid- och kostnadbesparing som känt med Smart-seq2 beror på att negativa cDNA-bibliotek kan identifieras endast med hjälp av att mäta koncentrationen av cDNA. Att minska beroendet av kvalitetskontroll är ett viktigt steg för att storskaligt kunna kartlägga tidigare okänd diversitet bland protister. I sin tur är det viktigt att effektivt kunna utforska okänd protistdiversitet för att få en bättre förståelse för den eukaryota cellens evolution.
Acknowledgements

This thesis would never have reached the current state without advice, help and ideas from numerous people. First of all I want to thank Thijs Ettema for giving me the opportunity to do cutting-edge research in an interesting field. Getting the chance to do this type of exploratory science that I have done during my time as PhD student has been an incredible experience. The supervision has been a perfect balance between freedom and getting pointed in the right direction when needed. I also want to thank Jan Andersson for being my co-supervisor. I have really appreciated our scientific discussions and your feedback.

Being a PhD student in Ettema lab also put me in the same group as the inspiring protistologists Courtney and Laura, which advice and help have been invaluable to complete this thesis. I am also very happy to have worked in the same group as Anja, our scientific discussions during the weekends are one of the highlights during my time at BMC. Another highlight is the recruitment of Mahwash as a master student to the lab. I will never forget our late night of Stentor surgery at EBC to document the regenerating cells at different time points. Mahwash, I really want to thank you for working with me on the Stentor regeneration project, and for all the support you have given me while completing this thesis. The results in this thesis have also relied on support from Anna-Maria and Claudia, thank you for all the help in the clean room! Additionally, I am not sure if any analysis would have been possible at all without Felix. Felix, you have been in the lab since I got recruited, and you have given invaluable support since day one. It has been an honor to end my time as a PhD student at a desk next to you.

The chapters in this thesis would not have been possible without international collaborations with Alexander Tice, Brandon Barton, Genoveva Esteban, Hunter Hines and Matthew Brown. Thank you Alex and Matt for all the nice chats at conferences and joining me, together with Brandon, to prepare the last chapter of this thesis. Genoveva and Hunter, I really enjoyed visiting the UK to learn how to sample protists and later exploring the Spirostomum transcriptome.

Further, during my time as a PhD student I have been surrounded with people that always have an answer or a solution. Daniel, Erik, Jimmy, Jonathan, Katarzyna, Lionel, Maria, Martha and William, thank you for patiently answering all my questions when I have asked for help. Writing this acknowledgement section actually makes me question myself as a free-living
species, since I have not only parasited on the knowledge of the post docs and more senior researchers, I have also mined the brains of other PhD students. **Anders, Disa, Eva, Jennah, Joran, Jun-Hoe** and **Max**, I truly want to thank you for all the help and feedback you have shared. From the Ettema lab, last but not least, **Lina** and **Tom**, I appreciate your efforts with all the cDNA libraries I have thrown at you. The Nextera XT libraries you have delivered have always been of the highest quality.

I thank **Siv** for all the interesting discussions, especially the ones about codon reassignment. I also thank all my other colleagues **Alejandro, Andrea, Christian, Emil, Erik, Feifei, Guilherme, Karl, Kristina, Lisa, Weizhou** and **Zeynep** for making Molecular Evolution into such a great group of people. I also want to thank my lab neighbor **Marcus**, that you let me use your Bioanalyzer has been crucial for getting my last two chapters together. Besides all the nice people from BMC, I have also had the pleasure to meet **Cheng-Jie, Fabien** and **Vasily** on EBC. **Cheng-Jie**, I have appreciated all our discussions and I am also happy that you made me go with the subway, instead of cab, in Moscow. **Fabien**, it is great to have an expert in protistology like you in Uppsala, you both provide great feedback and push interesting science forward. **Vasily**, it has been nice to go sampling with you, and I appreciate all the tricks regarding filtering you have shared with me.

I am grateful to **Simone Picelli** for sharing so much knowledge about single-cell RNA sequencing. Every minute I have spent talking to you have probably saved me 10 hours of experimental work in the lab to test something you already have tested. I am also thankful for **Marshall Wallace** that came all the way from US to Sweden for discussing *Stentor* biology with me. Another great scientist I am happy to have interacted with is **Tatjana Haitina**, you taught me so much when I was an engineering student and really made me prepared for my PhD position.

**Cecilia, Julia, Joacim** and **Oliwer**, it is nice to see that people working outside of my field could help me with this thesis. Cecilia and Joacim, I appreciate that you stepped in and proofread some parts of the thesis just before submission. Julia, great job with the cover! Oliwer, our road trip lead to finding some of the protists which are discussed in detail in the third chapter of this thesis.

Finally, I want to thank my **family** for all the support. Especially my father has done a significant impact on this thesis, since he suggested stopping at the pond where all the elongated euglenids and the *Mayorella* species where found, which later were discussed in Paper III.
References


Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 1853

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”.)