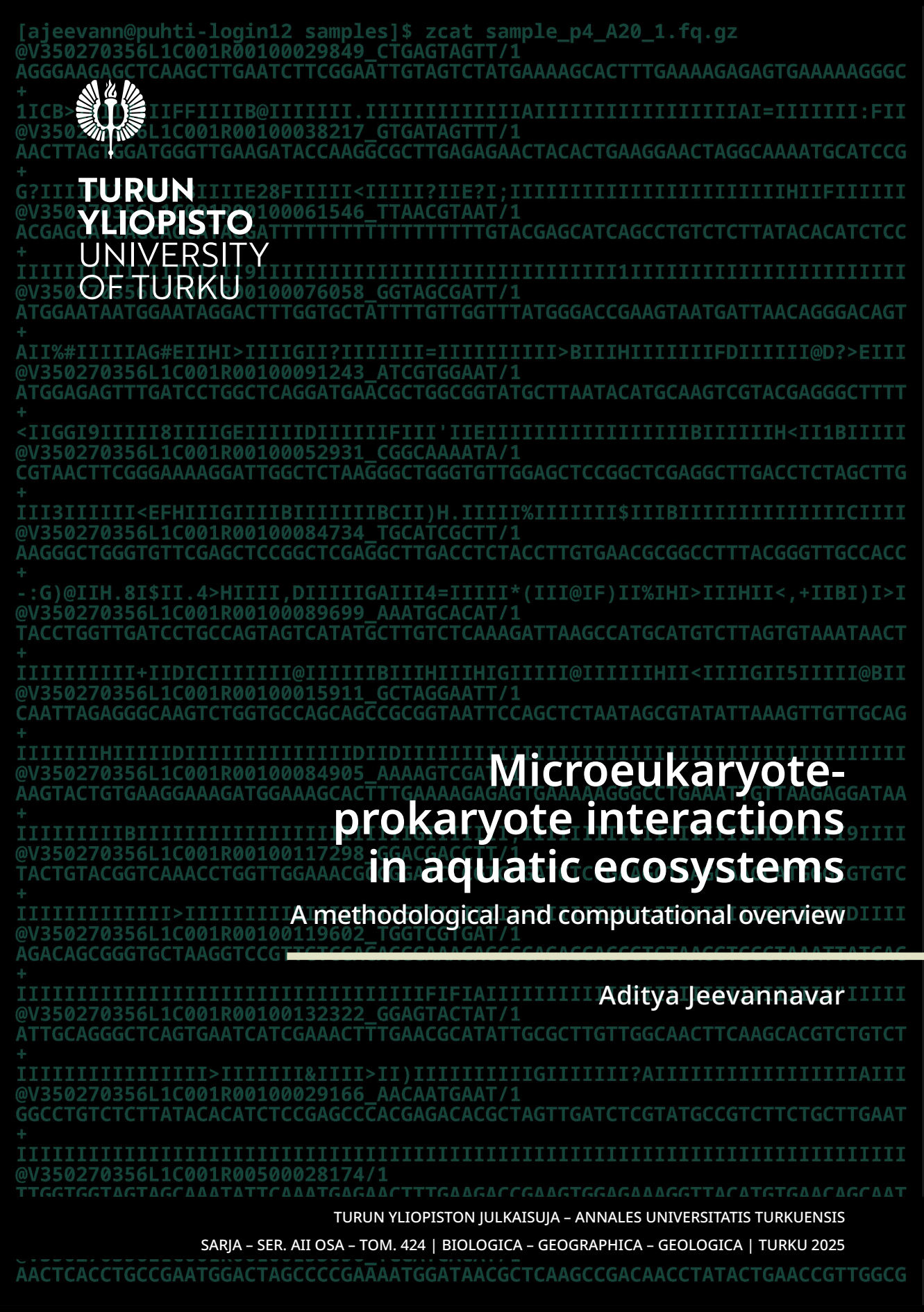

TURUN YLIOPISTO
UNIVERSITY OF TURKU

Microeukaryote-prokaryote interactions in aquatic ecosystems

A methodological and computational overview

Aditya Jeevannavar

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS

SARJA – SER. AII OSA – TOM. 424 | BIOLOGICA – GEOGRAPHICA – GEOLOGICA | TURKU 2025

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

MICROEUKARYOTE-PROKARYOTE INTERACTIONS IN AQUATIC ECOSYSTEMS

A methodological and computational overview

Aditya Jeevannavar

TURUN YLIOPISTON JULKAISUJA ANNALES UNIVERSITATIS TURKUENSIS
SARJA SER. AII OSA TOM. 424 | BIOLOGICA GEOGRAPHICA GEOLOGICA | TURKU 2025

University of Turku

Faculty of Science
Department of Biology
Biology
Doctoral programme in Biology, Geography, and Geology (BGG)

Supervised by

Academy Research Fellow and Docent,
Manu Tamminen
Department of Biology
University of Turku
Turku, Finland

Professor, Teppo Hiltunen
Department of Biology
Univeristy of Turku
Turku, Finland

Reviewed by

Docent, Conny Sjöqvist
Environmental and Marine Biology
Åbo Akademi University
Turku, Finland

Docent, Maliheh Mehrshad
Division of Microbial Ecology
Swedish University of Agricultural Sciences
Uppsala, Sweden

Opponent

Professor, Micah Dunthorn
Natural History Museum
University of Oslo
Oslo, Norway

The originality of this publication has been checked in accordance with the University of Turku quality assurance system using the Turnitin OriginalityCheck service.

Cover Image: Aditya Jeevannavar

ISBN 978-952-02-0497-6 (PRINT)
ISBN 978-952-02-0498-3 (PDF)
ISSN 0082-6979 (PRINT)
ISSN 2343-3183 (ONLINE)
Painosalama, Turku, Finland 2025

To sequences that let us imagine
entire lives that we cannot see with our eyes

UNIVERSITY OF TURKU
Faculty of Science
Department of Biology
Biology
JEEVANNAVAR, ADITYA: Microeukaryote-prokaryote interactions in aquatic ecosystems
Doctoral dissertation, 140 pp.
Doctoral programme in Biology, Geography, and Geology (BGG)
December 2025

ABSTRACT

Microorganisms are the most dominant forms of life in Earth's history. Interactions between prokaryotic and eukaryotic microorganisms play key roles in ecosystem functioning and can take many forms. Traditionally, such interactions have been studied either by isolating individual microeukaryotes and observing them and their symbionts under a microscope, or by identifying all microeukaryotes and prokaryotes in a community and analysing correlation patterns of their co-occurrence. The former methods fail to scale to the numbers and complexity typical of natural systems, while the latter fail to capture individual interactions and their functional context. This thesis aims to add single-cell transcriptomics as a high-throughput high-resolution tool to the microbial ecology toolbox to characterize natural microeukaryote function and their interactions with prokaryotes.

First, the thesis presents a study using amplicon sequencing for identifying microeukaryotes and prokaryotes. Analyses of their compositions across time and environmental conditions revealed that changes in the prokaryotic community had a causal impact on the eukaryotic community and the environment. However, neither functional activity nor individual- or species-level interactions could be elucidated from amplicon sequences. Thus, the thesis presents subsequent studies using single-cell transcriptomics to isolate individual microeukaryotes and sequence their transcriptomes to infer population-level taxonomic identity, sub-population-level functional profiles and metabolic states, and individual-level prokaryotic associations. Gene expression and associated prokaryote compositions enabled the identification of cells potentially close to dormancy or under bacterial colonisation.

Most environmental microeukaryotes, including many in this thesis, are thought to be uncultivable in laboratory conditions and are absent from reference genome or transcriptome databases. This thesis aggregates computational methods–*de novo* assembly of microeukaryote transcriptomes, transcript annotation using sequence and protein structure homology, differential expression and abundance analyses, and metabolic mapping–for such environmental microeukaryotes. By demonstrating single-cell transcriptomics' and these computational methods' efficacy among environmental mixotrophic microeukaryotes without cultures or references, I have set the stage to scale this reference-free culture-free method to the natural microeukaryotic diversity and study microeukaryote-prokaryote functions and interactions in the stressful, heterogeneous, and ephemeral conditions that they occur in.

KEYWORDS: Biodiversity, Protists, Single-cell, Symbiosis, Transcriptomics

TURUN YLIOPISTO
Matemaattis-luonnontieteellinen tiedekunta
Biologian Laitos
Biologia
JEEVANNAVAR, ADITYA: Mikroeukaryootti-prokaryootti vuorovaikutusten vesiekosysteemeissä
Väitöskirja, 140 s.
Biologian, maantieteen ja geologian tohtoriohjelma (BGG)
Joulukuu 2025

TIIVISTELMÄ

Mikro-organismit–bakteerit ja mikroeukaryootit–ovat maapallon historian hallitsevimpia elämänmuotoja ja näiden fyysiset vuorovaikutukset ovat keskeisiä ekosysteemien toiminnalle. Perinteisesti tällaisia vuorovaikutuksia on tutkittu esimerkiksi eristämällä yksittäisiä mikroeukaryootteja ja havainnoimalla niitä ja niiden bakteerisymbiontteja mikroskoopilla, tai vaihtoehdoisesti etsimällä korrelaatioita mikroeukaryoottien ja bakteereiden esiintymisessä erilaisissa ympäristöissä. Mainituista edelliset menetelmät eivät skaalaudu luonnonympäristöille tyypilliseen määrään ja monimutkaisuuteen, kun taas jälkimmäiset eivät kykene havainnoimaan yksittäisiä vuorovaikutuksia ja niiden funktionaalista kontekstia. Tämä väitöskirja esittää menetelmiä korkean resoluution työkaluiksi mikrobiekologian menetelmävalikoimaan, tarkoituksena selvittää luonnonympäristöjen mikroeukaryoottien toimintaa ja niiden vuorovaikutuksia bakteereiden kanssa.

Väitöskirjan ensimmäisessä tutkimuksessa selvitettiin mikroeukaryoottien ja bakteereiden merkitystä rehevöitymiselle kokeellisessa makean veden ympäristössä. Tutkimuksessa yhteisörakenteen aikasarja-analyysi amplikonisekvensointia käyttäen paljasti, että muutoksilla bakteeriyhteisössä oli kausaalinen vaikutus eukaryoottiyhteisöön ja tätä kautta ympäristön rehevöitymiseen. Mikrobien funktionaalista aktiivisuutta tai yksilö- ja lajitasoisia vuorovaikutuksia ei kuitenkaan voitu selvittää käytössä olleella amplikonisekvensointimenetelmällä. Tämän vuoksi väitöskirja esittelee kaksi jatkotutkimusta, joissa käytettiin yksisolutason transkriptomisekvensointia tarkoituksena päätellä yksittäisten solujen toiminnalliset tilat sekä yksittäisten solujen fyysiset vuorovaikutukset bakteerisolujen kanssa. Näissä tutkimuksissa onnistuttiin selvittämään ennestään tuntemattomien yksisoluisten organismien toimintaan liittyviä seikkoja, esimerkiksi merkkejä lepotilasta tai käynnissä olevasta bakteerikolonisaatiosta.

Valtaosaa ympäristön mikroeukaryooteista ei ole mahdollista viljellä laboratoriossa, eikä tietoa niistä ole tallennettu vertailutietokantoihin. Väitöskirjani esittelee menetelmiä, joilla tällaisia huonosti tunnettuja mikroeukaryotteja voidaan tutkia modernein DNA-sekvensointimenetelmin hyödyntäen transkriptomien *de novo*-rakentamista, DNA-sekvenssidatan tulkintaa sekvenssi- ja proteiinirakenteiden perusteella ja metabolista kartoitusta. Nämä tekniikat avaavat mahdollisuuksia aivan uudenlaisille ympäristön mikroeukaryoottien funktionaalisille tutkimuksille sekä mikroeukaryootti-bakteeri vuorovaikutusten selvittämiselle, ja tuottavat täysin uutta tietoa näiden mikroskooppisten mutta ekosysteemien kannalta keskeisten organismien toiminnasta ympäristöissään.

ASIASANAT: Biodiversiteetti, Protistit, Transkriptomiikka, Yksisoluinen

Acknowledgements

My sincerest gratitude goes to my supervisor Dr. Manu Tamminen, who brought me into his research group when I wasn't even sure if academic research was the track for me and showed me how wonderful it can be to do research at a university. I first met him as a student in his *Bioinformatics for Biologists* course in 2019. His words of encouragement inspired me to change my focus from molecular biology to bioinformatics and computational biology (my previous aversion to which has been documented). He encouraged and guided me to work on projects with different datatypes and complexities. He helped me develop my skills and was always available to guide me when I was stuck. And for all that and more, Manu, I thank you.

I would also like to express gratitude to the various collaborators who I've worked with during my PhD, especially to Dr. Anita Narwani in Switzerland and Dr. Javier Florenza and Prof. Stefan Bertilsson in Sweden. I would also like to thank Niina, Jonna, Elsi, and Santeri for being wonderful officemates and colleagues and creating a great work atmosphere.

I thank the University of Turku Graduate School (UTUGS) for the Finland Fellowship (funded by the Finnish Ministry of Education and Culture) and the Doctoral Programme in Biology, Geography, and Geology (BGG) for continuous funding that allowed me to focus uninterrupted on my research. I thank this thesis' preliminary examiners, Dr. Conny Sjöqvist and Dr. Maliheh Mehrshad, for examining and critiquing my thesis and giving me constructive feedback to improve it. I also thank Prof. Micah Dunthorn for accepting the role of the opponent and for what I'm sure will be a thought-provoking discussion.

I am extremely grateful to all the friends I've made in Turku over the last four years. Prakhar, Oceane, Ludo, Veera, Venni, Purabi, Julius, Heloise, Charlotte, Axelle, Jesse, Aida, Leticia, Constantina, Kjell, Veera, Mark, Nikos, Luigi, Farshad, and more in the department, your company has kept me sane throughout the stressful times. Prakhar, all the time I've spent playing video games with you have been times I've not worried about anything else in the world and for that I thank you. Oceane and Ludo, our weekly sessions of flying on broomsticks, arguing about the patheticness of wizards who can only make boxes, and watching trashy television have been oases of tranquility in times of tumult and turmoil, and for that I thank you. Veera and Venni, the sparse and short occasions we have convened at have always been sources of great fun and digressions into daft yet deft discourse, and for that I thank

you. Purabi, I couldn't have asked for a better partner to organise quizzes with. Also, special shout out to the Bio Bitches!

I would also like to thank the friends I made during my studies in India (most of whom are now everywhere else in the world). Sneha, Burhan, Sarvesh, Srinath, Neeraj, Karthik, Faidh, Aditya, thank you for being there through times of exuberance and exile. Special shout out to Tennis and the Time Zone Gang!

Pappa, mummy, and Anisha, I thank you for raising and caring for me and being the unchanging rock-solid foundation I could depend on.

Finally, I would like to thank my partner, Marie-Pier, for her love and kindness as well as for setting a very high standard for hard-work and resilience. You make me a better person.

December 2025
Aditya Jeevannavar

ADITYA JEEVANNAVAR

I'm an inquisitive researcher working at the intersection of microbial ecology and bioinformatics. I enjoy learning new tools and methods, and the extensive troubleshooting that occurs amidst this process. I believe poly-nucleotide and poly-peptide sequences can tell us a lot about the microbial world we neither see nor know much about. I am fascinated by how many file-types one has to go through to hear what these sequences say.

Table of Contents

Abbreviations

BLAST	Basic Local Alignment Search Tool
CO1/COI	Cytochrome c Oxidase Subunit I
DAA	Differential Abundance Analysis
dbRDA	Distance-based Redundancy Analysis
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DEA	Differential Expression Analysis
DNA	Deoxyribonucleic Acid
eDNA	Environmental DNA
dNTP	Deoxy-Ribonucleotide Triphosphates
ESV	Exact Sequence Variant
ASV	Amplicon Sequence Variant (obtained from DADA2)
sOTU	Sub-OTU (obtained from Deblur)
zOTU	Zero-radius OTU (obtained from UNOISE2)
FACS	Fluorescence-Activated Cell Sorting
FISH	Fluorescence In Situ Hybridization
GO	Gene Ontology
GO:BP	Gene Ontology: Biological Process
GO:CC	Gene Ontology: Cellular Component
GO:MF	Gene Ontology: Molecular Function
GSEA	Gene Set Enrichment Analysis
ITS	Internal Transcribed Spacer between SSU rRNA and LSU rRNA genes
KEGG	Kyoto Encyclopaedia of Genes and Genomes
LSU	Large subunit (of the rRNA operon)
NGS	Next Generation Sequencing
OTU	Operational Taxonomic Unit
PCR	Polymerase Chain Reaction
QC	Quality Check
RCC	Roscoff Culture Collection
RFU	Relative Fluorescence Unit
RNA	Ribonucleic Acid
mRNA	Messenger RNA
rRNA	Ribosomal RNA

tRNA	Transfer RNA
scRNAseq	Single-Cell RNA Sequencing
SEM	Structural Equation Model(ling)
SSU	Small subunit (of the rRNA operon)
TPM	Transcripts Per (kilobase) Million
t-SNE	T-distributed Stochastic Neighbour Embedding
UMAP	Uniform Manifold Approximation and Projection
UMI	Unique Molecular Identifier

List of Original Publications

This dissertation is based on the following original publications, which are referred to in the text by their Roman numerals:

I Aditya Jeevannavar, Anita Narwani, Blake Matthews, Piet Spaak, Jeanine Brantschen, Elvira Mächler, Florian Altermatt, Manu Tamminen. Foundation species stabilize an alternative eutrophic state in nutrient-disturbed ponds via selection on microbial community. *Frontiers in microbiology*, 2024; 15: 1310374.

II Aditya Jeevannavar, Javier Florenza, Anna-Maria Divne, Manu Tamminen, Stefan Bertilsson. Cellular heterogeneity in metabolism and associated microbiome of a non-model phytoflagellate. *The ISME Journal*, 2025; 19(1): wraf046.

III Aditya Jeevannavar, Javier Florenza, Stefan Bertilsson, Manu Tamminen. Microeukaryote-prokaryote associations persist through changing light and expression conditions. *Manuscript*.

1 Introduction

Four and a half billion years ago, the Earth was formed [1; 2]. Over the first half a billion years, the molten surface of the planet Earth cooled down to form a solid crust and oceans [3] and was enveloped by a strongly reducing (CH_4 + N_2, NH_3 + H_2O, or CO_2 + H_2 + N_2) to neutral (CO_2 + N_2 + H_2O) atmosphere [4]. Over the next half billion years, highly diverse and dynamic environments like reductive atmospheres, tidal pools, dry-wet and heating-cooling cycles, freezing temperatures, high energy discharges from lightning and submarine hydrothermal vents, extraterrestrial inputs, and the vast oceans set the stage for the prebiotic synthesis of the building blocks of life–amino acids, lipids, ribose sugar, nucleobases, fatty acids, nucleotides, and oligonucleotides–in the primordial soup on Earth [5; 6; 7].

Over the next few hundred million years, these building blocks of life came together and/or polymerised into phospholipids, nucleic acids, and proteins, forming the basis of life: compartmentalization (through lipid bilayers), replication (through nucleic acids), and metabolism (through proteins and other organic molecules) [5]. Thus, in the form of the first cells that could replicate themselves, life on Earth began, albeit at the microbial level. One of the earliest evidences of life has been seen in the form of Stromatolite colonies of bacteria from 3.4 billion years ago [8]. In all this time and all the time till now, the conditions on Earth changed and these microorganisms evolved with it. Over the aeons, they began to execute various functions to adjust to their environment as well as adjust the environment to themselves. Some performed photosynthesis, some fixed atmospheric inorganic nitrogen, and some developed the ability to make cell walls and cysts. Over the next billion years, these first microorganisms evolved and diverged to form the three domains of life–Bacteria, Archaea, and Eukarya.[1] From nearly four billion years ago, microorganisms have been the most dominant forms of life in Earth's history and continue to form the majority of biomass and biodiversity on the planet [10; 11].

A brief understanding of the origin (as above) and natural taxonomic classification of life is necessary to understand the different forms of life and their interactions. Until the development of high quality optics and the microscope, natural taxonomy was based on complex morphology and fossil records, and life was primarily categorised into animals and plants. In 1866, Haeckel introduced a third

[1] This was first proposed as a natural system of organisms by Woese in 1990 [9].

kingdom, protista, to include microscopic single-celled organisms as distinct from animals and plants [12]. In 1938, Copeland split out bacteria from protista to create a fourth kingdom [13]. In 1959, Whittaker added a fifth kingdom for fungi [14]. Following the discovery of the structure of the DNA, molecular techniques to decipher DNA sequences, and the understanding that DNA forms the genetic material in nearly all forms of life, it became apparent that molecular structures and sequences of the DNA and resulting transcribed RNA contain more information for inferring taxonomic classification and evolutionary relationships than morphology and fossil records. In 1990, using molecular sequence comparisons, Woese proposed that living organisms are more primarily divided in three domains: Bacteria, Archaea (which used to be considered part of the bacterial kingdom previously), and Eukarya or Eukaryota (which contains within it the kingdoms protista, fungi, plantae, and animalia) [9].

In the three decades since Woese's three domain proposal, culture and culture-free sampling and sequencing of microorganisms from diverse environments, from lakes to deep-sea thermal vents and from animal guts to the deep terrestrial subsurface, have enabled the discovery of tens of supergroup and kingdom level branches and hundreds of phyla level branches in the tree of life, the phylogenetic tree that connects all organisms based on shared evolutionary history. Recent discoveries have led to better resolution of branches especially in Archaea (Asgard archaea, *Euryarchaeota*, DPANN–*Diapherotites*, *Parvarchaeota*, *Aenigmarchaeota*, *Nanohaloarchaeota*, and *Nanoarchaeota*, TACK–*Thaumarchaeota*, *Aigarchaeota*, *Crenarchaeota*, and *Korarchaeota* etc.) [15] and Eukarya (*Haptista*, *Cryptista*, *Archaeplastida*, *Amorphea*, CRuMs, *Discoba*, *Metamonada*, TSAR–*Telonemids*, *Stramenopiles*, *Alveolates*, and *Rhizaria* etc.) [16].

Prokaryota or prokaryotes, from *pro* meaning without or before and *karyon* meaning kernel, includes all organisms that do not have a nucleus, i.e., bacteria and archaea, and considering how understudied archaea are compared to bacteria, often refers to just bacteria. Eukarya, eukaryota, or eukaryotes, from *eu* meaning true and *karyon* meaning kernel, includes all organisms with membrane bound nuclei. Aside from the membrane-bound nucleus, eukaryotes are distinct from bacteria in other ways as well: they possess complex morphology like cytoskeletons, mitochondria, endoplasmic reticulum, Golgi aparatus, flagella etc. as well as chimeric genomes, meiosis, mitosis, epigenetic phenomena, and cell cycles [17].

Symbiosis is the spatially close and temporally long-term interaction between organisms of different species [18]. Endosymbionts, smaller symbionts living inside larger symbiotic partners, live either inside the cells of their partners or outside the cell but inside larger multicellular animals. Ectosymbionts or epibionts live outside of and are usually attached to their larger symbiotic partners. From the interaction of prokaryotes in shallow marine environments 3.43 billion years ago to form reef-like build-ups of stromatolite colonies [8] to probiotic interactions in the hu-

man gut, symbiosis is everywhere. Intra- and inter-domain interactions or symbioses have been observed across the tree of life and throughout its multi-billion year long history [18]. Perhaps, the most important of these is the endosymbiotic origin of eukaryotes.

The origin of the first eukaryotes, nearly two billion years ago, is thought to be from the endosymbiosis of an α-proteobacteria in an Asgard archaean host [17]. The α-proteobacterium endosymbiont or parasite lost many of its genes to the archaeal host's nuclear genome and eventually lost its ability to survive independently, thus forming the mitochondria, an organelle found in nearly all eukaryotic cells. The mitochondria provided the bioenergetic means for an increase in cell sizes from 10x to 100x that of bacteria and archaea and the evolution of eukaryotic cell complexity [19]. Another later endosymbiotic event involving a cyanobacteria led to the origin of photosynthesis in eukaryotes via plastids, an organelle found in all plants (including red algae and green algae). The presence of photosynthesis in other eukaryotes like diatoms, cryptomonads, and dinoflagellates is from a secondary endosymbiosis of a green or red alga [19].

Microeukaryote hosts of symbiotic interactions span across all the eukaryote supergroups, particularly well-studied large protists in culture [18]. The diversity of prokaryotic symbionts involved is also vast, including *α-proteobacteria*, *Bacteriodetes*, and *Chlamydiae*, but might have some sampling bias as well [18]. Some microeukaryotes might host communities of many co-habiting symbiotic prokaryotes [20; 21; 22]. These symbionts could be spatially located in the cytoplasm, co-localized with certain organelles (like hydrogenosomes), in the nucleus, or outside the cell. The microeukaryote-prokaryote interaction may be mutualistic (beneficial to both), commensal (beneficial to one), or even parasitic (beneficial to one, harmful to the other). On extreme ends of the benefit scale exist syntrophy–where each partner requires metabolites or toxic waste removal provided by the other, and microbial farming–where the microeukaryote cultivates prokaryotes extracellularly for later consumption. Over evolutionary timescales, the symbiont can have one of three outcomes: to go extinct, adapt to an obligate, but host-species-independent, symbiont lifestyle, or to integrate into the host as an organelle [18].

Observations of microeukaryote-prokaryote interactions in aquatic ecosystems over the last century have enabled our understanding of the variety and wide-spread existence of the interactions as seen above, but most of them rely on the cultivability of the microeukaryotes in the lab and other biases associated with ease of study [23]. A significant majority of the observations involve direct observation of the interaction using microscopy or fluorescence *in situ* hybridization (FISH). This has two main disadvantages. One, microeukaryotes, especially free-living heterotrophs, that are smaller, less-studied, and more difficult to cultivate are scarcely investigated for functional activity and interactions [18]. Two, the fitness effects of microeukaryote-prokaryote interactions under laboratory conditions in nutrient-rich media and ab-

sence of other organisms is likely not representative of the more stressful, diverse, and ephemeral natural conditions [18; 23]. Knowledge of these interactions in the heterogeneous conditions enshrouding natural microbial diversity is lacking and is the focus of this thesis.

1.1 Natural microbial diversity

1.1.1 What is species?

Before we talk about diversity, we must first understand what is a species. According to the classical biological species concept, species represents a population of organisms that can interbreed among themselves to produce fertile offspring and are reproductively isolated from other such populations [24; 25]. This works well for separating humans and horses and helmeted hornbills as distinct species but not for a variety of closely related animal and plant species that can hybridize to produce fertile offspring or bacteria that reproduce asexually. One solution to this limitation is the morphological species concept according to which a species represents a population of organisms that are morphologically and anatomically alike [25; 24; 26]. Another solution is the ecological species concept according to which species represents a population of organisms occupying a unique ecological niche [27]. Both these concepts still present limitations when faced with populations that exhibit high morphological variance or those that change their roles based on available resources, and with evolutionarily distinct populations exhibiting similar morphologies and ecosystem functions. Mitigating the limitations of the other species concepts is the phylogenetic species concept according to which species is a population of organisms that share a common pattern of ancestry and descent and that form a single branch on the tree of life [28; 29]. Species, in this paradigm, are distinguished using unique genetic markers. The limitation of this species concept lies in the ambiguity associated with how recent a common ancestor or how unique a genetic marker demarcates a species. Establishment and use of a singular concept of species appears unlikely with at least 22 different concepts formulated already [30; 31].

For microorganisms, in this era of high throughput sequencing, the phylogenetic species concept appears to be used most often to demarcate species and conduct phylogenetic analyses [32; 33]. Assuming (i) that each cell contains DNA as the genetic material and (ii) that there are certain genes, such as those coding for ribosomal RNAs essential for making proteins, conserved across the genomes of all organisms [34], (iii) that these certain marker genes are sufficiently constrained that they do not change rapidly in time and are resistant to horizontal gene transfer [35; 36; 37], and (iv) that all copies of the marker gene in the organism are identical [34; 38], we can say that populations with more similar gene sequences share a more recent ancestor [39; 40]. In the identification of microbial species and corresponding phy-

logenetic analysis, the 16S (for prokaryotes) and 18S (for eukaryotes) ribosomal subunit's rRNA genes are used most commonly as the representative of these conserved genes with sequence variation across species [39; 41; 42]. In the 16S and 18S genes, the presence of multiple hyper-variable regions aids in identifying species and the presence of conserved regions aids in primer design for amplifying these sequences [43; 44].

To reduce the ambiguity associated with how recent a common ancestor or how similar a sequence to determine a distinct species, the concept of an Operational Taxonomic Unit is used. An Operational Taxonomic Unit or OTU is defined, in a population of different species, as representative of all the sequences that cluster together at (usually) 97% sequence similarity [45]. Defining the concept within the limits of individual populations meant that addition of any data from new species or new populations required the re-clustering and re-analysis of all data, without which populations, in terms of their constituents, could not be compared [46]. Thus, similar to standard field guides that exist to identify specific morphological markers in different animal and plant species, reference sequence databases were prepared after careful sampling and sequence analysis from different ecosystems across thousands of geographic locations. Identifying the constituent species in a new population consequently involves the clustering of marker gene sequences obtained from all the organisms in the population to determine the OTUs and mapping the representative sequences from these OTUs to the reference database to identify the sequence-based taxonomy. These reference databases are regularly updated to include new sequences and species and correct the branches in the tree of life based on new findings. Reference databases for the 16S rRNA gene in prokaryotes include SILVA [47], Genome Taxonomy Database (GTDB) [48], Ribosomal Database Project (RDP) [49], and Greengenes2 [50]. Reference databases for 18S, ITS, and CO1 genes in eukaryotes include Protist Ribosomal Reference (PR^2) [51], UNITE [52], coidb [53], PhytoREF [54], and EUKARYOME [55].

Clustering the sequences at a similarity threshold of 97% sequence identity runs the risk of grouping multiple similar species into a single OTU and consequently losing diversity [56]. OTUs clustered at higher thresholds like 100% sequence identity run the risk of identifying sequencing errors as new species and false diversity [57]. Thus, Exact Sequence Variants or ESVs were developed that omit clustering altogether and use an error model to correct for sequencing errors.[2] This results in data with single-nucleotide resolution that is more easily comparable across studies [61]. Thus, the ESV (used interchangeably with ASV) approach is purported to be the future of the field [62]. However, ASVs could lead to falsely splitting the same or-

[2]ESVs generated from different tools using different error correction or de-noising methods are called different names: Amplicon Sequence Variants (ASVs) when obtained using DADA2 [58], sub-OTUs (sOTUs) when obtained using Deblur [59], and zero-radius OTUs (zOTUs) when obtained from UNOISE2 [60].

ganism into multiple taxa calls, particularly considering some organisms like fungi and apicomplexans can have multiple non-identical target gene copies [63; 64; 65].

Determining the species constituents of a microbial community using singular marker genes such as the 16S rRNA gene can sometimes lead to systematic false negatives or non-identification of certain groups of related species due to technical reasons such as primer incompatibility [66; 67]. Primers are used to amplify the marker gene sequences from the microbial community's DNA pool. Bias in the primers' affinity or compatibility for the marker genes' flanking sequences in certain taxa groups can lead to misleading representations of the community's constituents [68]. Appropriate choices of primers and marker gene targets based on the expected species in a community can alleviate some of the bias. However, shotgun metagenomics, especially shallow shotgun metagenomics which is comparative in cost to marker gene amplicon sequencing, offers alternative solutions for identifying constituent species in a microbial community since they do not involve any sequence-specific amplification (and consequently no primer bias) and provide additional information in the form of whole genome sequences and consequently functional potential [69]. Their limitation lies in the lack of reference genomic sequence information for microbes from all environments. Thus, shotgun metagenomics is more useful in well-studied systems like the human gut microbiome whereas marker gene amplicon sequencing is more useful in other environmental ecosystems [69].

Marker gene amplicon sequencing can also lead to a lack of species-level resolution [70]. This can be alleviated by capturing larger or more sensitive regions of the hyper-variable regions of the marker genes [71] or sequencing multiple marker/conserved genes for phylogenomic analyses [72].

1.1.2 What is diversity?

Diversity refers to the different species present in the study system, from a petri dish to the ocean, and their relative abundances [73]. Estimates of the number of species on Earth have ranged from two million to three trillion, mostly comprised of diversity among insects, fungi, protists, and bacteria [74; 75; 76]. Of the estimated 550 gigatons of carbon (Gt C) in the biosphere, 450 Gt C is thought to be composed of plants followed by 70 Gt C of bacteria, 12 Gt C of fungi, 7 Gt C of archaea, 4 Gt C of protists, and only 2 Gt C of animals [77]. Considering their small sizes and masses, there are an estimated 2×10^{30} prokaryotes and approximately 10^{27} protists on the Earth across marine, freshwater, terrestrial, and deep biosphere ecosystems [77]. Microbial diversity, biomass, and abundance, in the form of prokaryotes and microeukaryotes, are immense and present in parts of the biosphere from inside our guts to the surface of this book to high up in the atmosphere and deep in the ocean. But due to habitat destruction and overexploitation and climate change, biodiversity loss is rampant [78; 79]. It is essential to identify how many species are

present, what they are, where they are, and what they are doing before they vanish in the sixth great extinction.

Most of our understanding of biology comes from the deep study of a few organisms, usually in controlled laboratory conditions, called model organisms. Thus, most of the estimated species on Earth have never been identified or studied or grown in a lab [80]. While the study of model organisms for the understanding of biology conserved across the tree of life is essential and fruitful, it still leaves behind a large gap in understanding those organisms that cannot, yet, be cultivated in the lab [80] or those that behave or interact differently in laboratory conditions [18]. Marker gene based amplicon sequencing, and more recently metagenomic sequencing, has aided tremendously in identifying and counting the different microbial species present in the environment [81; 82]. However, more needs to be done to decipher what they do there.

1.1.3 What is functional diversity?

The millions of different species, in their ecosystems and niches, perform various functions from essential processing and cycling of carbon, energy, nitrogen, phosphorus, and other nutrients, to developing special traits to consume prey (like chemotaxis or flagella) or avoid predators (like toxic metabolites and sharp spikes) [83]. The field of functional ecology investigates the different roles that species play in their environment in a bottom-up manner, by focussing on the functions of individual species or strains [84]. The field of ecosystem function investigates these different functions in a top-down manner by focussing on the ecosystem and deciphering all the active processes in the ecosystem arising as a function of different abiotic and biotic components present in the environment and their interactions [85]. Often, strong associations exist between the physicochemical composition of the environment, the composition of microbial communities, and their functional potential and activity [86]. Deciphering these associations, especially the role of microbial diversity in ecosystem functioning, is essential and requires sufficient understanding of interactions between species [86].

One of the clearest forms of functional diversity seen in the environment is trophic diversity. Different organisms, including microorganisms, have developed different strategies for acquiring the carbon, energy, and nutrients for their survival. Autotrophy is the ability for organisms to convert inorganic forms of energy like light and electrochemical potential to organic forms like glucose and to fix carbon. Autotrophs are the primary producers in the food web. Heterotrophy is the mode of energy metabolism in organisms that cannot produce their own food or energy from inorganic sources and rely on the consumption of other organisms or organic carbon fixed by other organisms. Heterotrophs form the other trophic levels in the food web. Both autotrophs and heterotrophs depend on the environment, other organisms,

and their metabolites for procuring essential nutrients other than carbon and energy. Mixotrophs represent a unique form of trophic activity which enables them to survive in both autotrophic and heterotrophic modes [87].

Life exists rarely in communities comprised entirely of a single species. Multiple species come together, act and interact, to perform different functions for the survival of communities. One of the fundamental goals of microbial ecology is to decipher the link between the diversity of the microbial communities and the ecosystem processes and stability [86]. One of the most well-studied such links between diversity and function involves the nitrogen cycle via nitrification and denitrification. Nitrification, for example, involves two completely distinct sets of prokaryotes: aerobic ammonia oxidizers convert ammonia (NH_3) to nitrite (NO_2^-) and nitrite oxidizers convert nitrite (NO_2^-) to nitrate (NO_3^-) [88]. Survival in some environments and some ecosystem functions might even necessitate syntrophic interactions where metabolites produced by one species in the community need to be consumed and metabolised by another for the survival of both [89].

Not all functions performed by microbes are beneficial. Some outcomes of the functional activity in an ecosystem can be undesired and potentially harmful long-term. One such harmful outcome is caused by eutrophication or the overenrichment of nutrients like phosphorus and nitrogen in aquatic environments [90]. Eutrophication can affect the microbial diversity, including functional diversity, and cause rapid increases in the accumulation of algae (photosynthetic prokaryotes and microeukaryotes) in aquatic ecosystems called algal blooms [90]. These can potentially cause secretion of toxins in the water, depletion of oxygen levels in the water due to blocking of sunlight, ”dead zones” due to anoxia from decomposition of dead algae, and unavailability of water for human and other animals' consumption. The occurrence and persistence of such harmful phenomena in aquatic ecosystems are also mediated by microeukaryote-prokaryote interactions and presence or absence of foundation species [91].

1.1.4 Do all species contribute equally to the community?

Microbial communities, like all biological communities, are composed of few abundant and many rare, sometimes dormant, community members [11]. While some rare microbial taxa maintain high metabolic activity and perform essential ecosystem functions, others may remain dormant or inactive and grow in abundance when the ecosystem changes to fit their optimal conditions [92; 93]. These rare biosphere members host a nearly unlimited repository of genomic innovation and maintain functional redundancies in the community [94]. The functional redundancies may be essential to ecosystem stability in case there is a sudden reduction or extinction of abundant taxa [11].

Often some species perform certain functions that have a disproportionate effect

on the survival of the communities and ecosystem functioning [95]. *Foundation species* are abundant (in number and biomass) members of an ecosystem that are involved in numerous non-trophic mutualistic interactions, some of which provide structural support for other species in the community and modulate fluxes of energy and nutrient flow through the ecosystem [95]. For example, aquatic macrophytes increase the habitat complexity and affect nutrient cycling and sedimentation and sediment-nutrient levels by attenuating currents and waves, and providing physical structure for shelter and attachment [96; 97; 98]. They are distinct from *keystone species* (herbivores or predators) that increase local species richness via their trophic interactions and creating physical space for subordinate competitors, and *ecosystem engineers* that physically change materials in the environment and mediate energy and nutrient fluxes in their ecosystem [95]. Although organisms of all species modify their environment, *foundation species* are distinguished due to the disproportionate effect of their abundance or biomass [95].

1.2 Molecular methods and sequencing techniques

1.2.1 Information content of the cells

All prokaryotes and eukaryotes contain DNA as the genetic material (the means for passing information from one generation to the next) which is transcribed into RNA. RNAs act as the template for the production of proteins (mRNA) or are in other ways involved in the production of proteins (rRNA and tRNA) or catalyse some reactions. RNAs mediate the expression of the genes encoded by the DNA in the genetic material. Proteins synthesised by the translation of RNAs are involved in the catalytic functions in all cells.

Genes, subsequences of DNA present in the cell, dictate what functions can be potentially performed by the organism. Gene transcripts, particularly mRNAs, dictate what functions are being performed actively by the organism. This follows the assumption that each mRNA transcript leads to the synthesis of the same number of proteins and that each protein has similar catalytic activity in the cell. The assumptions are not egregious since studies of functional activity usually involve comparisons of the same function across cells, where they hold true. Nonetheless, for a more accurate and precise picture of functional activity in the cell, the proteins and metabolites in the cell can be studied.

Eukaryotes, especially microeukaryotes, however, add many additional complexities in the process of transcription. Transcripts (pre-mRNA) transcribed from the gene generally contain introns (or intragenic regions that are removed) and exons (or expressed regions that are spliced together to make mature mRNAs). In most eukaryotes, alternative splicing is used to join exons in different permutations to make different mature mRNAs and proteins from the same gene. Some microeukaryotes pos-

sess even more complex transcriptional mechanisms like twintrons (introns within introns that need to be removed sequentially) [99], noncanonical genetic codes (for example UGA encoding for tryptophan instead of the usual stop codon) [100], and 4- or 5-letter codons (due to translational slippage or shifting of reading frame every time certain codons appear in-frame) [101]. Thus, the inference of the mature mRNA transcript sequence from the gene sequence is obscured, which consequently obscures the inference of the translated protein sequence and function [102].

Phylogenomic analyses, which include the identification of taxonomy of the community components and the measurement of their abundances, are based on certain genes conserved across the tree of life (as mentioned in section 1.1.2). DNA sequences of these genes identified in the community can be used to infer who is present. Additionally, RNA sequences of these genes' transcripts can be used to infer who in the community is functionally active [11].

1.2.2 Broad sequencing methods

Amplicon sequencing (also called metabarcoding, marker gene sequencing, or metataxonomics) is the PCR amplification and sequencing of highly conserved genes, such as the rRNA gene. This is used to identify and calculate relative abundance of the prokaryotes and eukaryotes (based on the marker gene amplified) present in the community. The most common target for identifying prokaryotes is the 16S rRNA gene and eukaryotes is the 18S rRNA gene. Tens of thousands of studies have used these marker gene targets to identify prokaryotes and eukaryotes. Amplicon sequencing is fast and cost-effective since it captures sequences for a single gene, but can encounter amplification bias, i.e., the primers used to amplify the marker genes might systematically amplify sequences from certain species more than from others [66; 67]. Well-established and publicly available workflows and tools exist for the analysis of amplicon sequencing data [103].

Genomics (or genomic sequencing) is the sequencing and analysis of all DNA content in cells, tissues, and communities, without selecting for individual genes. Sequencing can be done after amplifying the genetic content using random nucleotide primers or with no amplification by cutting the DNA sequences at random places (like a shotgun would). The DNA can be sourced from single isolated cells or from environmental samples or from cultures.[3] Shotgun metagenomics refers to the sequencing, without amplifying, and analysis of DNA sequences from all organisms in an environment, thus yielding information on community composition, phylogeny, and functional potential of individual species in the community. After assembly of metagenomic reads into contigs and chromosomes, the sequences are binned to yield

[3] The cultures can be axenic, i.e., of a single organism or strain, or non-axenic, i.e., of more than one organism. Heterotrophic microeukaryotes are often not cultivable in axenic cultures due to the necessity of prey and symbionts for their survival.

metagenome assembled genomes (MAGs) for individual species in the community. Metagenomics, however, ignores cellular heterogeneity and involves complex computational analyses related to binning. To mitigate these shortcomings, single-cell genomics, at single-cell resolution, amplifies the small quantity of genetic material present in individual cells and yields single amplified genomes (SAGs). In the case of microeukaryotes that can host prokaryotic symbionts intracellularly or on the surface of their cells, the amplified genetic content could contain prokaryotic sequences as well, which can be filtered out as contamination or separated and used to investigate the potential roles of the symbionts.

Transcriptomics (or transcriptomic sequencing), analogous to genomics, is the sequencing and analysis of all RNA content in cells, tissues, and communities, without selecting for individual transcripts. Sequence information from the isolated RNA is first converted to DNA in the form of complementary DNA (cDNA) which is consequently amplified and/or sequenced. Since mRNA forms a small percentage of the total RNA present in a cell, mRNA enrichment (using poly-dT primers in eukaryotes) is carried out prior to sequencing. Analogous to genomics, metatranscriptomics and single-cell transcriptomics can be performed. For organisms with well-curated annotated reference genomes, such as humans and mice, only short sequences from either end of the transcripts can be amplified or isolated and sequenced to infer the functional activity of the cells by mapping the transcript ends to the reference genomes. For environmental microeukaryotes, however, reference genomes are rarely available and thus full-length transcripts are sequenced to enable *de novo* transcriptome assemblies.

Analogous to genomics and transcriptomics, proteomics and metabolomics are used to characterise the profiles of proteins and metabolites in the cells, but with methods more expensive and experimentally challenging than sequencing, such as liquid chromatography-coupled tandem mass spectrometry. These approaches also rely on the presence of high quality reference genomes and/or reference transcriptomes. They are used very rarely in the study of environmental microorganisms without prior characterization of the communities using metagenomics or metatranscriptomics.

The presence of complex transcriptional mechanisms in microeukaryotes complicates the inference of functional potential from genomic sequences alone due to the lack of confidence in directly predicting transcript and protein sequences from gene sequences [102]. Transcriptomics offers a good alternative to gain similar inferences and sidestep the transcriptional complexity, but it has its own trade-offs with respect to the ability to maintain laboratory cultures and the representativeness of the assembled transcriptomes. Culture-based transcriptomics yields transcriptomes representing nearly the complete functional capacity, but require that the microeukaryotes be maintained in culture. Culture-free transcriptomics yields partial transcriptomes that cover the functions active at the time of sampling, but can be used to study

environmental microeukaryotes in natural conditions.

1.2.3 Specific methods for inferring interactions

Metataxonomics, metagenomics, and metatranscriptomics can be used to identify and quantify the prokaryotes and eukaryotes present in an ecosystem and consequently infer potential interactions between microeukaryotes and prokaryotes based on their co-occurrence. However, definitive evidence of the interactions can only be obtained by studying them at single-cell resolution [104]. The earliest studies of microeukaryote-prokaryote interactions involved the manual isolation of microeukaryote single-cells and identifying the interactions based on morphology. Morphology-based identification is expensive, time-consuming, and prone to misassignments among morphologically similar but phylogenetically distant taxa.

Metataxonomics, metagenomics, and metatranscriptomics can provide more definitive evidence of interactions when performed on non-axenic but unialgal cultures of microeukaryotes, i.e., cultures containing a singular phylogenetically unique microeukaryote but with all associated prokaryotes [105]. However, this approach has many challenges: (i) the microeukaryotes should be amenable to maintaining in culture, (ii) it is expensive and time-consuming to maintain cultures and potentially impossible for all members of the natural microeukaryote diversity, and (iii) microeukaryotes potentially behave and interact differently in laboratory cultures as compared to natural conditions.

Single-cell interactomics or microbiomics, by isolating singular microeukaryote cells and then performing 16S rRNA amplicon sequencing, aims to mitigate some of the challenges above. The cells can be isolated manually or using less time-intensive methods like fluorescence-activated cell sorting (FACS). 18S rRNA can also be amplified in the same single-cell pool to enable identification of microeukaryote and extend the method to natural diversity and beyond lab cultures of well-studied microeukaryotes [106; 21; 22]. A recent study even depleted the microeukaryote host's chloroplast 16S genes using CRISPR-Cas9 to reduce host sequence contamination [107]. However, challenges associated with amplification bias, lack of distinction between prey and symbionts, and unavailability of information about functional potential or activity persist.

A different way to isolate single cells is by capturing them in emulsion droplets or capsules, such as in epicPCR [108]. Each droplet or capsule can then be used as a separate reaction mixture for amplification of multiple target genes and physically linking them. This enables the identification of linkages, here rRNA gene for taxonomic identity and another target gene in the same prokaryote. By considering symbiotic interaction as the linkage, epicPCR can be used to identify interactions between bacteria and viruses [109] or bacteria and smaller bacteria [110], or potentially even microeukaryote and bacteria [104]. However, microeukaryotes have been

rarely captured in emulsion droplets or capsules, particularly due to the difficulty in breaking microeukaryote cell walls. Capturing microeukaryotes and their associates in the same droplet faces the challenge of massive size differences between the interaction partners jeopardizing capture of exactly two partners at a time. Additionally, the challenges associated with single-cell interactomics mentioned above exist here as well.

Single-cell based metagenome sequencing, by isolating singular microeukaryote cells and then performing genomic (technically metagenomic due to the presence of prokaryotic symbionts) sequencing, foregoes issues associated with amplification bias and missing data on functions. Such studies have begun identifying extensive collections of bacteria linked to microeukaryotes and observed that individual microeukaryotes simultaneously host multiple bacteria in symbiotic relationships [111; 112; 113]. Current studies have mostly restricted themselves to certain branches of the tree of life such as the ciliates, but their potential to uncover the diversity of symbiotic associations in nature is high.

Single-cell transcriptomics, similar to single-cell based metagenome sequencing, can yield high resolution data on the microeukaryotes' functional activity, as well as their taxonomic identity and the identities of their prokaryotic associates from the rRNA sequence reads and transcripts. Transcriptomics, compared to genomics, skips the complications associated with complex transcriptional mechanisms in microeukaryotes described in section 1.2.2. Early single-cell transcriptomics studies focussed on microeukaryotes in culture [114; 115], model parasites [116; 117] and yeasts [118; 119], or on manual isolation of single cells [120; 121; 122; 123; 124; 125; 126]. The isolation of cells using FACS and sequencing of full-length transcripts using tools like Smart-seq2 and Smart-seq3 enable high-throughput and high-resolution gene expression and associate identification among microeukaryotes [127]. Morphological information can also be linked to the sequencing information by using tools like MASC-seq [128]. Previous studies using these methods have tested them on microeukaryotes with reference genomes/transcriptomes, but their potential for reference-free and culture-free study of microeukaryotes' expression and their associates is high.

1.3 Aims of the study

This work focuses on the development and use of different methods for the study of microeukaryote-prokaryote interactions in aquatic ecosystems. Microeukaryote-prokaryote interactions have been prevalent in aquatic ecosystems for over half of Earth's history and are crucial to ecosystem functioning and stability. The interactions have been studied but using methods with insufficient throughput to characterize them in natural populations. The aim of this thesis is to add a useful high-throughput tool to the microbial ecology toolbox by characterizing microeukary-

ote function and prokaryotic interactions at single-cell resolution in a reference-free culture-free manner. An additional goal of the thesis is to compare this single-cell transcriptomic approach to the widely used amplicon sequencing approach.

The specific aims are:

1. To identify the microeukaryotes and prokaryotes mediating the unexpected ecological outcomes of co-occurring foundation species in eutrophic aquatic ecosystems.

2. To measure gene expression changes and associations with prokaryotes of a non-model microeukaryote in culture.

3. To identify environmental microeukaryotes and their associated prokaryotes as well as measure the microeukaryote's gene expression changes in a natural sample without cultures and references.

2 Materials and Methods

Microeukaryote-prokaryote interactions were studied in different systems and using different sequencing technologies and analysis techniques. Broadly, chapter **I** utilizes 16S SSU rRNA, 18S SSU rRNA, and CO1 gene amplicon sequencing to identify the prokaryotes and eukaryotes in the microbial communities in large artificial ponds and tracks changes in the community structure as nutrients are added in. The artificial ponds also contain foundation species which affect the physicochemical characteristics of the pond via the microbial communities. Chapters **II** and **III** both utilize single-cell transcriptomics but with two slightly different library preparation protocols and completely different systems. In short, chapter **II** utilizes Smart-seq2 on a non-axenic culture of *Ochromonas triangulata* to characterize the gene expression of the microeukaryote as well as identify its prokaryotic associates. Expanding on the potential of single-cell transcriptomics identified in chapter **II**, chapter **III** utilizes Smart-seq3xpress on the mixotrophic subset of a natural freshwater sample to identify the microeukaryotes, characterize their gene expression, as well as identify each one's prokaryotic associates. The different methods used in the three chapters are summarised in table 1.

2.1 Experimental setup and sampling

For Chapter **I**, 20 large ($15m^2$) artificial ponds, situated at the Swiss Federal Institute of Aquatic Sciences (Eawag) research facility in Dübendorf, Switzerland were used for a full-factorial manipulation of the presence and absence of two foundational species: *Dreissena polymorpha* (referred to henceforth as *Dreissena*), a freshwater mussel, and *Myriophyllum spicatum* (referred to henceforth as *Myriophyllum*), a freshwater plant, with an oligotrophic control and four replicates each. The ponds were identical, fiberglass-lined, in-ground, and outdoors (Figure 1). Each had a shallow end (0.5 m) and a deep end (2.0 m). The ponds with *Dreissena* received 100 live mussels (mean dry biomass without shells = 632.67 mg) and those with *Myriophyllum* received 100 live macrophytes (mean dry biomass = 19.84 g). The ponds were set up and the experiment begun in June 2016. The experiment ended in June 2017.

Stated simply, there were four ponds each with *Dreissena* and *Myriophyllum* (henceforth *Myriophyllum+Dreissena* or MD), only *Dreissena* (henceforth *Dreissena* or D), only *Myriophyllum* (henceforth *Myriophyllum* or M), and neither (hence-

Method	Study
Amplicon Sequencing	I, III
Single-cell transcriptome sequencing	II, III
Adapter trimming & QC	I, II, III
OTU/ASV/ESV calling	I, II, III
Contaminant removal	II, III
De novo transcriptome assembly	II, III
Read mapping	II, III
Functional annotation	II, III
Diversity analyses	I, II, III
Dimensionality reduction & Clustering	I, II, III
Differential expression analysis	II, III
Differential abundance analysis	I, II, III
Metabolic mapping	II, III
Structural equation modelling	I

Table 1. Summary of methods used

forth Control or C), each of which received five doses of nutrients in the form of nitrates and phosphates. The last remaining four ponds, henceforth oligotrophic controls or O, did not have either foundation species and did not receive nutrient additions. The effect of the foundation species on microbial community structure and ecosystem characteristics were inferred by comparing against the controls. The effect of the nutrient additions were inferred by comparing against the oligotrophic controls.

Nutrient disturbance or additions of nitrogen and phosphorus were made in the form of KNO_3 and K_2HPO_4 respectively to the first 16 ponds. No nutrients were added to the oligotrophic control ponds. Five nutrient additions were made in increasing concentrations at 2-2.5 week intervals. The phosphorus concentrations were 10, 20, 30, 40, and 50 μgl^{-1}. Since nitrogen can escape the ponds via denitrification, nitrogen was added at double the Redfield ratio or 32 times the molar concentration of phosphorus.

Full-depth samples were collected from each pond in the deep end using a custom-made Leibold-sampler consisting of a 180 cm long PVC tube (with a diameter of 5 cm) and a rubber ball attached to a rope passing through it. Samples were taken in intervals of weekly to monthly over a period of 13 months. Two different sampling schedules existed, one for the 16S and 18S rDNA amplicon sequencing, and another for the CO1 gene amplicon eDNA sequencing. After sampling, the water was filtered through a GF/F filter (0.7 µ m pore size). DNA was extracted from the filter using a xanthogenate nucleic acid isolation [129], phenol:chloroform:isoamyl alcohol extractions, isopropanol precipitation, and ethanol precipitation. The resulting pellet

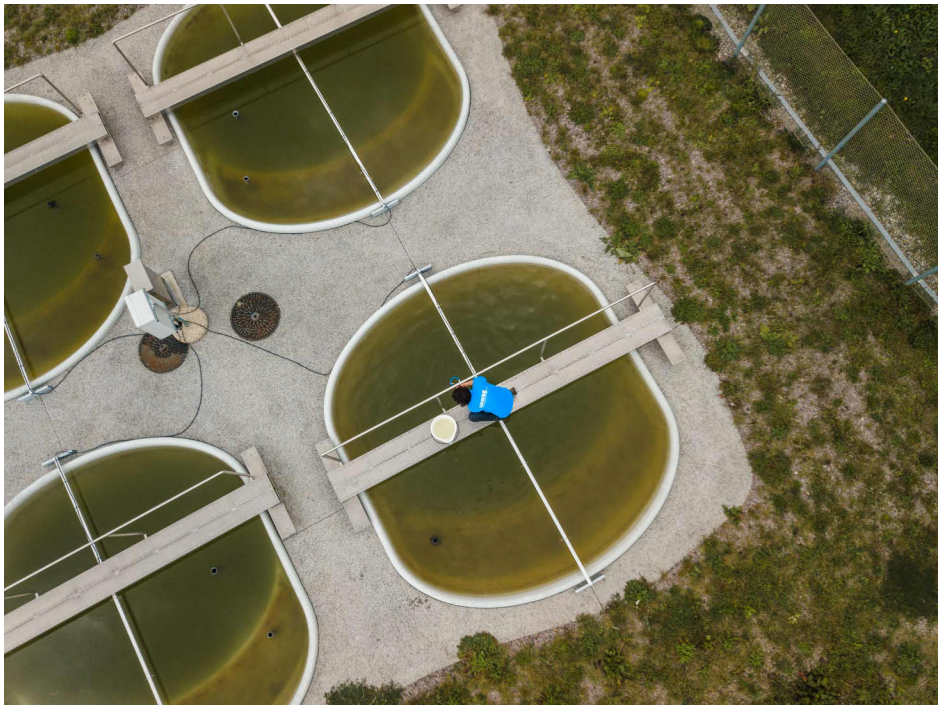

Figure 1. Overhead view of the experimental ponds, with Dr. Anita Narwani for scale.

was dissolved in pure water and stored at -20° C. DNA from the eDNA samples, which were collected similarly, were extracted using a DNeasy Blood & Tissue Kit (Qiagen, Germany) with small modifications [130].

For Chapter **II**, *Ochromonas triangulata*, a mixotrophic microeukaryote, i.e, one that performs photosynthesis as well as preys on bacteria in its environment, was studied. To maintain its access to prey as well as other bacteria it interacts with, *Ochromonas triangulata* cells (strain RCC21) were maintained as unialgal but non-axenic (i.e., not single species) cultures in filter-sterilized K/2 medium [131; 132]. The cultures were kept serially at 18° C and 12 hour photoperiod by diluting 10X every 10 days.

Samples were taken at two time points; one, two days after refreshing (or diluting) in a fast-growth phase, and two, eleven days after refreshing (or diluting) in a slow-growth phase. Samples from both growth phases were independently stained with LysoSensor Blue DND-167 (Invitrogen) and sorted using a MoFlo Astrios EQ cell sorter (Beckman Coulter). Cells were selected based on chlorophyll *a* signal, indicative of photosynthesis, and LysoSensor signal, indicative of viable cells with food vacuoles. These cells were deposited into 384-well PCR plates (kept at 4° C and containing 2 µl of lysis buffer), immediately spun down, and then stored at -80° C.

For Chapter **III**, natural water samples were collected from lake Långsjön (Stock-

holm, Sweden) in oligotrophic and clear water conditions. The water was filtered on-site through a 42 μm net and acclimatized in five litre glass bottles for two days under a 14 hour photoperiod. After acclimation, half the sample was subjected to a shading treatment (with only 7.5% of acclimation irradiance) while the other half continued in acclimation conditions. Samples (250 mL) were taken at four time intervals, one just before the onset of shading, and the others six hours, 24 hours, and 48 hours after. 190 cells per treatment per time-point were sorted based on simultaneous signals for SYBR Green II, chlorophyll autofluorescence, and LysoSensor Blue, to select for putative mixotrophic microeukaryotes.

2.2 Library preparation and sequencing

For chapter **I**, the 16S and 18S regions were PCR-amplified using primer sequences listed in supplementary table 1 of chapter **I** resulting in fragments of 464 bp and 576 bp respectively [133; 134]. The primers include random nucleotide frameshifts to enhance the sequencing performance. PCR for 16S amplicons was done using a QIAGEN Multiplex PCR Master Mix (Catalogue No. 206145). PCR for 18S amplicons was done using the Phusion polymerase (Catalogue No. M0535L). The samples were indexed using an Illumina Nextera XT index kit and sequenced on an Illumina MiSeq machine in 300 bp paired-end mode.

The CO1 gene from the eDNA sampling was amplified using established degenerate primers to yield a 313 bp fragment [135; 136]. PCR was done using the AmpliTaq Gold polymerase (Thermo Fisher Scientific, USA). The samples were indexed using an Illumina Nextera XT index kit and sequenced on an Illumina MiSeq machine in 300 bp paired-end mode.

For chapter **II**, the Smart-seq2 protocol was used to prepare single-cell transcriptomic libraries for the 768 single cells in two 384-well PCR plates [137]. Smart-seq2 enables the full-length cDNA generation from single-cells with high sensitivity and accuracy, all using standard reagents [137]. The lysis agent was 0.2% Triton-X100 (Sigma) and the reverse transcriptase was 5U/µl SuperScript II Reverse Transcriptase (Invitrogen). Following reverse transcription, the resulting cDNA generated were pre-amplified (using PCR), here using a KAPA HiFi HotStart Ready Mix (Roche). Next, tagmentation (or the simultaneous DNA fragmentation at random sites and adapter ligation) was done with the enzyme Tn5 transposase in the Nextera XT kit (Illumina) to prepare the sequencing libraries. The two pooled libraries, one from each 384-well PCR plate, were sequenced on separate lanes of an Illumina NovaSeq 6000 SP v1.5 machine in 150 bp paired-end mode.

For chapter **III**, the Smart-seq3xpress protocol was used to prepare single-cell transcriptomic libraries for the 1,520 cells in four 384-well PCR plates [138]. Smart-seq3 improved upon the Smart-seq2 protocol by incorporating 5' Unique Molecular Identifiers (UMI) and improving sensitivity [139]. Smart-seq3xpress improved

upon Smart-seq3 by shrinking the protocol so smaller quantities of the reagents are needed [138]. The molecular steps in the protocol are similar to Smart-seq2 except for the addition of UMIs in the reverse transcription step. The single-cell transcriptomic library prep for this study were outsourced to Xpress Genomics (Stockholm, Sweden) who also did the sequencing on separate lanes of DNBSEQ-G400 sequencing machine (MGI Tech, Shenzhen, China) in 150 bp paired-end mode to obtain approximately 1 million sequence reads on average.

2.3 Computational analyses

Amplicon sequencing involves the PCR amplification of a small target sequence of DNA. In chapter **I**, it was used to amplify fragments of the 16S SSU rRNA, 18S SSU rRNA, and CO1 genes from microbial communities. Due to the hypervariable regions present in these genes, amplicon sequences can be used to taxonomically identify the microeukaryotes and the prokaryotes. Assuming no PCR bias, i.e. the polymerase did not selectively amplify some sequences more than others, the number of amplicon sequence variants (ASVs) or sequences of the same kind can be used as a proxy for the number of that particular microbe in the community. Since the whole sample is processed, the data obtained this way contains the information for all the eukaryotes and prokaryotes present in the community (or at least those that were filtered and lysed). However, information about individual microeukaryote-prokaryote interactions is lost in the process.

nf-core is a community effort to collect and share analysis pipelines built using Nextflow, a workflow tool for data-driven scientific workflows [140; 141]. The pipelines in nf-core are useful for well-documented replicable end-to-end analysis of sequencing data. One such analysis pipeline, relevant to amplicon sequencing data, is *ampliseq* [103]. The *ampliseq* pipeline includes primer trimming using cutadapt, quality evaluation using FastQC, ASV inference using DADA2, and reporting using MultiQC [142; 143; 58; 144]. Modules for downstream analyses like diversity analysis, differential abundance, and function prediction are available, but these analyses were done separately and not as part of the pipeline in this study.

Single-cell transcriptomics using full-transcript sequencing (like the Smart-seq2 and Smart-seq3xpress protocols) yields primarily sequence reads corresponding to random fragments of the full-length of the cDNA representing all the mRNA in the individual cells. These reads can be assembled together to obtain the full-length mRNA sequences. The number of reads (or the UMIs for more precision) so obtained are used as a proxy for the number of those mRNA sequences present in the cell, and consequently as a proxy for the relative number of the protein encoded by that mRNA and the activity of the function performed by the resulting protein. Read counts are adjusted for the length of the mRNA sequence, since long mRNAs can produce more fragments. It is assumed that each mRNA produces relatively the same number of

proteins before degrading.

rRNA constitutes the overwhelming majority of the RNA present in the cells. The scRNAseq protocols mentioned above target only the mRNA in the cells using oligo-dT primers which attach to the poly-A tails of the mRNA. Despite this mRNA targeting, the sheer amount of the rRNA in a cell results in many fragments of the rRNA accumulating in the sequencing library. Although rRNA contains little information about cellular function beyond relative levels of transcriptomic activity, parts of this rRNA, as described above, can be used to identify the microbe bearing the rRNA. In chapters **II** and **III**, this carryover rRNA was used to identify the microeukaryote as well as its individual-level prokaryotic associates.

2.3.1 Adapter trimming and quality check

After base calling by the sequencers, the sequencing data is provided to the end user almost universally in the FastQ format which consists of the DNA read sequence and corresponding to each of the nucleotide base call in the read sequence, a quality score. The per base quality score indicates the confidence in the base call based on a sequencing technology dependent statistical model. These sequence qualities can be summarised and visualized across all the reads of the sample using FastQC [143]. Summary reports like this can be used to identify and rectify systematic errors in sequencing. Reports like this can be summarised across all samples, and even across multiple analyses and checks, using MultiQC [144].

In library preps for next-generation sequencing (NGS) experiments, like those described here, adapters are key components of the workflow. Adapters are short pieces of DNA which attach to the DNA fragments, that are to be sequenced, on either side. Adapters enable them to bind to the sequencer's flow cell's solid surface. They also contain primer sites for the PCR primers or sequencing primers to bind to as well as indexes or barcodes that are unique to samples/plate wells so multiple samples can be pooled and sequenced together. While adapters are essential for sequencing, they are unnecessary and potentially error-inducing in downstream analyses. Therefore, after demultiplexing (or separating sequence reads based on the sample-specific indexes) the data, the adapters are trimmed from the reads. Many tools exist for trimming adapters and low quality bases from the ends of the reads such as Cutadapt [145], Trimmomatic [146], fastp [147], Trim Galore! [148], BBmap BBduk [149], AdapterRemoval [150], chopper [151] and dozens more. In chapter **I**, adapters and low quality bases were trimmed using Cutadapt (as part of the nfcore AmpliSeq pipeline), usearch, and prinseq-lite [145; 152; 153]. In chapters **II** and **III**, adapters and low quality bases were trimmed using Trim Galore! and Cutadapt.

2.3.2 OTU/ASV calling

OTU/ASV calling is the process of determining the taxonomic origin of a sequence and is based on the assumption that the target sequence (like 16S, 18S, CO1) contains variable regions that are specific to individual species. In essence, one might take the specific DNA sequence of the 16S gene in a microbe, look it up in a database of 16S gene sequences, and identify that this sequence belongs to, for example, the *E. coli* bacteria. In microbiome research, two main strategies have been used to do this: Operational Taxonomic Units (OTUs) and Exact Sequence Variants (ESVs), most commonly used in the form of Amplicon Sequence Variants (ASVs) obtained from the DADA2 tool. Tools for generating OTUs include usearch [152; 60] and vsearch [154]. Tools for generating ESVs include DADA2 (ASV) [58], Deblur (sOTU) [59], and UNOISE2 (zOTU) [60].

The OTUs and ASVs generated all need reference databases to look up sequences in to assign the taxonomy. Reference databases for the 16S rRNA gene in prokaryotes include SILVA [47], Genome Taxonomy Database (GTDB) [48], Ribosomal Database Project (RDP) [49], and Greengenes2 [50]. Reference databases for 18S, ITS, and CO1 genes in eukaryotes include Protist Ribosomal Reference (PR^2) [51], UNITE [52], coidb [53], PhytoREF [54], and EUKARYOME [55]. Aside from taxa calling for amplicon sequencing data which relies on taxonomic databases mentioned above, there exist tools like Kraken 2, Diamond, and FastQ Screen that can be used to search sequences against large repositories of genomes to determine the taxonomic origin of the sequences [155; 156; 157]. These can use as reference databases the complete genomes in RefSeq or the BLAST nt and nr databases.

In chapter **I**, ASV calling was done for 16S, 18S, and CO1 genes using DADA2, DADA2, and UNOISE2 respectively, using as references SILVA, PR^2, and Barcode of Life Data Systems (BOLD) databases respectively. In chapters **II** and **III**, the rRNA reads were first separated out from the rest using RiboDetector and then used for ASV calling using vsearch against the SILVA and PR^2 databases [158]. The RiboDetector and vsearch workflow for taxa calling in scRNAseq data was validated in chapter **II** by the accurate identification of *Ochromonas triangulata*. Trimmed reads were also filtered using Kraken 2 to keep only unclassified reads, since most microeukaryote sequences are absent from standard databases.

2.3.3 Contaminant removal

Independent studies have shown that sample handling, especially DNA handling, in the laboratory can incur human-associated and laboratory contaminants from ultrapure water, culture media, PCR reagents, and DNA extraction kits themselves [159; 160; 161; 162; 163; 164; 165; 166]. It is important to remove contaminating sequences from the data and the contaminants from the analysis results in order to cor-

rectly identify the microbes under study and infer the correct associations between microeukaryotes and prokaryotes. To this end, a table of common human-associated and laboratory contaminant genera and families of bacteria was prepared. Taxonomy tables of the samples under study were filtered using this contaminant or "kitome" table to rid contaminants from downstream analyses [159].

Additionally, the non-rRNA reads in the single-cell transcriptomics data were put through Kraken 2 to identify contaminants. The reference database consisted of NCBI taxonomic information, as well as the complete genomes in RefSeq for the bacterial, archaeal, and viral domains, along with the human genome, and a collection of known vectors (UniVec Core). A confidence score threshold of 0.60 was used [167]. This enabled the filtering out of potentially contaminating DNA from human handling of samples in laboratory as well as ambient bacterial, archaeal, and viral contaminants.

2.3.4 *de novo* transcriptome assembly and read mapping

Adapter and quality trimming of the raw data from short-read single-cell transcriptomics yields millions of short sequence reads of length 150 bp. Short reads are a consequence of the choice of sequencing technology, for example Illumina NovaSeq short reads vs PacBio long reads. The aim of transcriptome assembly is to decipher the origin of the short reads and rearrange them together to form the original transcript sequence [168]. This is usually achieved by identifying overlaps between reads[1] and arranging them accurately to form contiguous sequences or contigs. Transcriptome assembly is most commonly done using k-mers[2] and De Bruijn graphs (directed acyclic graphs representing overlaps between sequences of symbols like k-mers). All the k-mer subsequences (with different choices of k) from each read are identified, tabulated, and stored as nodes in the De Bruijn graph. Edges are drawn between nodes with exactly *k-1* overlap. Long paths through such a De Bruijn graph represent contigs. This, in addition to handling edge cases of isoforms etc., is done by numerous transcriptome assemblers like Trinity [169], rnaSPAdes [170], SOAPdenovo-Trans [171], Oases [172], Trans-ABySS [173], IDBA-tran [174], and RNA-bloom [175]. However, long-read sequencing, already in 2011, was thought to generate reads longer than transcripts and obviate the need for transcriptome assembly in the near future [176]. With improvements in the quality of long-read se-

[1] Overlaps between reads from the same transcript in a cell are possible due to the cDNA amplification and tagmentation done in library prep. cDNA amplification ensures multiple full copies of the transcript are available for tagmentation where random cuts are made in these copies. The random cuts ensure that the same transcript sequence can yield different reads with partial overlaps.

[2] k-mers are oligonucleotides of length k. Think of it as a polymer of nucleotides. Polymer of one nucleotide would be a monomer, of two nucleotides would be a dimer, and of k nucleotides would be a k-mer.

quencing technologies from PacBio and Oxford Nanopore as well as new long-read sequencing technologies from Roche and MGI Tech, long-read scRNAseq might actually obviate transcriptome assembly [177].

The quality of the assembly can be assessed in various ways to ascertain different aspects of it. The fragmentation, or lack thereof, of the assembly can be assessed using the N50 and Ex90N50 statistics [178]. The N50 length statistic implies that 50% of the assembled transcript nucleotides are found in contigs that are of at least N50 length. The Ex90N50 statistic is the N50 statistic limited to the most expressed transcripts representing 90% of expression. The read representativeness of the assembly can be assessed by aligning the reads against the assembly using an aligner like bowtie2 [179]. The completeness or near-completeness of the the transcripts assembled can be assessed by BLAST-ing the sequences against all known proteins and determining the number of unique top matching proteins that align across more than a given proportion of their length. Finally, the completeness of the assembled transcriptome can be assessed by how many conserved orthologous genes are are present in the assembly, using tools like BUSCO and compleasm and databases of conserved genes from orthoDB [180; 181; 182]. The contiguous sequences assembled do not necessarily represent mRNA transcripts in the cell since they could contain isoforms, splice variants, non-coding regions etc. Thus, the contigs were processed through TransDecoder to get the longest isoforms and open reading frames and clustered using CD-HIT to keep only the representative sequences [183; 184].

Read mapping here refers to the alignment of reads to the assembled transcripts and then using this alignment to estimate the abundance of each transcript in the cell, or in other words to quantify the gene expression. Read alignment can be done using tools like bowtie [185], bowtie2 [179], and STAR [186] and then counted using RSEM [187] (RNA-seq by Expectation-Maximization). This can also be done in a single step by performing pseudoalignment using k-mers, instead of comparing every read against every transcript, with tools like Kallisto [188] and Salmon [189]. The UMIs or Unique Molecular Identifiers can be used to more accurately quantify the gene expression [190]. Read counts are normalized by transcript length and library size or the total reads per cell to yield metrics like transcript per (kilobase) million (TPM) or fragments per kilobase million (FPKM). RNAseq data is compositional, i.e, measures of expression of genes are relative to the whole sample of gene expression. Thus, to adjust for some cells expressing some genes that are not expressed in others and still be able to compare expression of individual genes across samples, a normalization of weighted trimmed mean of the log expression ratios (trimmed mean of M values (TMM)) can be used [191].

2.3.5 Functional annotation of transcripts

de novo transcriptome assembly yields us sequences representing transcripts in the cell, but these still contain no information on what genes these transcripts represent or what proteins they encode. The most common strategy for gene/protein identity assignment and functional annotation of sequences is homology search. This usually means sequence homology search using tools like BLAST [192] or Diamond [156] against databases like the UniProt/Swiss-Prot [193] or NCBI Refseq [194] or UniRef [195]. Structure homology is a newer strategy enabled by the recent advances in deep learning models that can predict secondary and tertiary protein structures from just the protein sequences. By utilizing the tertiary structures of the proteins encoded by the assembled transcripts, predicted using models like AlphaFold2 [196], RoseTTAFold [197], and ESMFold [198], and performing structural search using tools like Foldseek [199] against databases of protein 3D structures obtained empirically like in the Protein Data Bank (PDB) [200] or predicted using state-of-the-art models like in AlphaFold Protein Structure Database [201] (which contains high-confidence predicted structures of all the proteins in UniProt/Swiss-Prot), we can assign protein identity using structural homology.

Following protein identity assignment using sequence or structure homology, the process of annotating transcripts continues with ontology[3], orthology[4], and protein family assignment, all of which associate some form of biological function or process to the transcript or protein. These functions can be obtained from sequence identity, assigned above, using various publicly available databases or via specific tools like Blast2GO for Gene Ontology [205], KofamScan (github.com/takaram/kofam_scan) or KofamKOALA for KEGG Orthology [206], and InterProScan for domains, functional sites, and protein family annotation [207; 208]. Functional annotation like Gene Ontology and Enzyme Commission (EC) numbers can also be assigned to proteins based on their 3D structures using tools like DeepFRI [209]. Other tools for functional annotation or sequence feature detection include SignalP for signal peptide identification [210], DeepTMHMM for predicting transmembrane domains [211], Infernal for RNA secondary structures [212], and eggNOG-mapper for functional annotation across multiple databases [213]. Functional annotation using multiple tools and databases mentioned above are combined in annotation suites like Trinotate [178].

[3] Gene Ontology, in the three key biological domains of Cellular Component, Molecular Function, and Biological Process, provides a biologically meaningful annotation of genes and gene products in a large number of organisms [202; 203].

[4] KEGG O or Kyoto Encyclopaedia of Genes and Genomes Orthology is a set of databases and a mechanism to link gene and protein orthologs to cellular metabolic enzymatic pathways, independent of organism [204].

2.3.6 Diversity analyses

As a result of ASV calling, we obtain, one, an ASV counts table containing samples on one axis and ASVs on the other, two, a taxonomy table containing ASVs on one axis and the different taxa[5] on the other, and optionally three, a phylogenetic tree. Combining this with a third table of metadata, containing variables/covariates like sampling time, treatment condition, phosphate levels etc., nets us all the data in neat tabular forms for all downstream analyses like the measurement and comparison of diversity across treatments and time. Most of the analyses preceding this were done on supercomputers. Most of the analyses downstream, including diversity analyses, were done on desktop computers with the R programming language [214]. Diversity analyses were done using the vegan package in R [215].

Alpha diversity is the diversity of species in a particular locality or sample [73]. It covers two important diversity characteristics of a sample–species numbers and relative importances–and therefore can be measured either as species richness of the community in terms of number of species or as evenness or concentration of dominance of species in the community in terms of diversity indices like the Shannon index or the Simpson index. The different measures or indices of alpha diversity can be thought of as special cases of a unified diversity equation with different Hill numbers [216; 217]. In analyses of interventional or experimental microbial communities, one is generally interested in the difference in alpha diversity of communities under different treatment conditions or over time.

Beta diversity or between-habitat diversity is the change of (number of) species along environmental gradients. Depending on whether the abundance of microbes (as opposed to only presence/absence) and the phylogenetic relationships between the microbes are considered, there are many different measures of distance or beta diversity between samples that can be used, for example Euclidean distance, Jaccard distance, Bray-Curtis dissimilarity, UniFrac and Weighted UniFrac distances, and Aitchson distance.

Gamma diversity is the total diversity in all samples combined. This is rarely used in microbiome analyses.

Redundancy analysis (RDA) is a multivariate statistical technique that is, here, used to understand the relationships between the microbial communities and the treatment or environmental conditions. RDA can be used to infer how well and how much individual relevant covariates explain the variance in the ASV table. Distance-based RDA (dbRDA) and permutational multivariate analysis of variance (PERMANOVA) can be used to compare groups of samples and to assess whether a treatment or environmental condition explains differences in the species abundances.

[5] namely Kingdom, Phylum, Class, Order, Family, Genus, and Species for the prokaryotes [47] and Domain, Supergroup, Division, Subdivision, Class, Order, Family, Genus, and Species for the eukaryotes [51].

2.3.7 Dimensionality Reduction and Clustering

The result of ASV calling is a normalized ASV counts table that has samples on one axis and ASVs on the other. Considering we had 20 sampling ponds sampled across 20 sampling time points, ponds with hundreds of different microorganisms, such a table can be very high dimensional. The most common strategy to visualize such a high dimensional dataset is a dimensionality reduction technique like Principal Component Analysis (PCA) and plotting the first two principal components as scatter plots. Other such techniques include Canonical Correlation Analysis (CCA), Non-metric Multi-Dimensional Scaling (NMDS) and distance-based Redundancy Analysis (dbRDA). Such analyses and statistical testing using PERMANOVA can be done using the vegan package in R [215; 214].

Similarly, the result of read mapping is a normalized read counts table that has samples on one axis and transcripts on the other. Considering we had 644 and 1,520 samples here and hundreds of thousands of transcripts, many thousands of which were functionally annotated, such a read counts table can be very high dimensional, more so than ASV tables. Dimensionality reduction of data is done using non-linear techniques like t-distributed Stochastic Neighbor Embedding (t-SNE) or uniform manifold approximation and projection (UMAP) [218; 219].

Analogous to how mammalian single-cell transcriptomic studies produce dimensionally reduced visualizations of the gene expression of different cell types and use unsupervised clustering to identify cell types and subtypes, in environmental single-cell transcriptomic studies, we can visualize the gene expression of different species populations and use unsupervised clustering to identify populations and sub-populations. In chapter **II**, since we had only one species and expected only two transcriptomic sub-populations, we used a simple DBSCAN to cluster the cells. In chapter **III**, due to the increased number of populations and sub-populations, we clustered the samples using shared nearest-neighbour (SNN) graphs or hierarchical clustering.

2.3.8 Differential Expression and Abundance Analyses

Differential expression analysis (DEA) and differential abundance analysis (DAA), which refer to the measurement of change in gene expression and species (or any other taxonomic level) abundance respectively as a result of changing environmental conditions, are the primary goals of hypothesis-based transcriptomic and amplicon sequencing studies respectively. In both *de novo* transcriptome assembly and ASV calling, it is essential to process the samples together, i.e., pool reads from the same species and assemble *de novo* together and ASV call and cluster amplicon reads together respectively, in order to make differential expression and abundance analyses possible. The two analysis methods are described together here because they share

many similarities. Read mapping/counting transcriptomic data and ASV calling amplicon data both yield compositional (i.e. relative measures of gene expression and species counts as compared to the whole samples) and sparse (i.e. high number of zeroes in the table because not all genes are expressed in all conditions and not all species are present in all samples) data that have samples on one axis and genes or ASVs on the other. The gene expression or ASV count tables together with sample metadata like treatment conditions and environmental covariates like temperature and turbidity are data necessary for the two analyses.

There are dozens of R packages available to perform differential expression and abundance analyses which differ in the way they handle the compositional and sparse nature of the data and what they assume the underlying distribution of the expression or abundance to be. Sparse or zero-inflated nature of the data is mitigated by the use of over-dispersed count models or zero-inflated count models or by zero imputation using Bayesian models to estimate values or adding a small pseudo-count to all zeroes. Compositional nature of the data is mitigated by normalizations like Trimmed mean of M-values (TMM) normalization, Relative log expression (RLE) normalization, Cumulative sum scaling (CSS), Centered log-ratio transformation (CLR) normalization, or Geometric mean of pairwise ratios (GMPR) normalization. The underlying distribution of the data is generally assumed to be negative binomial, beta-binomial, normal, or log-normal. Packages in R to perform these analyses include DESeq2 [220], LinDA [221], ANCOM-BC [222], ALDEx2 [223], MaAsLin 2 [224], edgeR [225], limma [226], corncob [227], eBay [228], sccomp [229], Cuffdiff2 [230], NOIseq [231], and sleuth [232]. Any single tool may not be perfect for identifying differentially expressed genes or differentially abundant taxa due to imperfect fitting of the data to the tool's assumed distribution or normalization and zero-handling strategy. So, it is suggested to use the consensus among five tools/methods to ensure robust biological interpretations [233; 234]. A recent study also found that elementary methods like Wilcoxon tests or linear regression/t-tests were more consistent and sensitive than more widely-used packages like those listed above [235].

2.3.9 Metabolic Mapping

Information on the expression of individual genes is obtained from the combination of functional annotation of transcripts and read mapping/counting. Information on which of the above genes (in the form of a list) are differentially expressed is obtained from differential expression analysis. But, practical insights into what this means inside the microeukaryote cells biochemically, enzymatically, or metabolically are missing. In chapters **II** and **III**, such insights were gained by mapping the expressed genes, especially the differentially expressed genes, onto KEGG Orthology maps using an R package called Pathview [204; 236]. Pathway maps from KEGG Orthology database were chosen because microeukaryotes are often not well-studied and

consequently do not often have species specific pathways available. By mapping differentially expressed genes between, for example, different treatment conditions, one can quickly and easily visualize which genes in which metabolic pathways are upregulated in which conditions. After inferences are made, the maps can then be rendered for more aesthetic presentation with tools like GraphViz [237] or vector image editors like InkScape.

Gene Set Enrichment Analysis (GSEA) is an analysis with a similar outcome of identifying metabolic functions that are upregulated or downregulated in one condition compared to another. GSEA with a pre-ranked list involves pathways or metabolic functions being represented as gene sets and the gene expression as ranked list of genes ordered based on the expression level change between conditions. GSEA then checks if a given gene set is overrepresented at either end of the ranked list of genes. This analysis is particularly useful to examine if any particular gene ontologies are enriched in particular treatment conditions. In chapters **II** and **III**, this was performed using the R package fgsea [238].

2.3.10 Structural Equation Modelling

Results from most of the analyses described here are based on correlations and associations. In order to perform causal hypothesis testing, in chapter **I**, we used a form of confirmatory path analysis called a structural equation model (SEM) to resolve complex multivariate relationships between treatments, environmental variables, prokaryotic and microeukaryotic community compositions, and biomass proxies. The R package piecewise SEM (or pSEM) enables using linear mixed effects models to explain each endogenous biotic and abiotic variable and accounting for random effects and temporal auto-correlation of dependent variables. Multiple distinct attempts were made to understand the effects of nutrient additions as well as the mediation of the foundation species in the realisation of these effects. Models with endogenous and exogenous variables represented as measured values, change in measured values from the start of the experiment, and changes or rates of change from previous sampling state, were tested. Changes in community composition from previous sampling point best explained the hypothesised model.

3 Results

3.1 Selection on prokaryotic community stabilizes alternative eutrophic state in nutrient-disturbed ponds

Eutrophication due to nutrient disturbance can cause major changes to aquatic ecosystems. Invasive foundation species, even if used as ecosystem management instruments, can, individually and interactively, have unexpected effects like stabilizing algal blooms. In this study, artificial ponds were sampled to test the interactive effect of two foundation species on a freshwater ecosystem's response to nutrient disturbances. Longitudinal samples of water were taken to filter out microbes for 16S and 18S rRNA gene amplicon sequencing, and to filter out eDNA for CO1 gene amplicon sequencing. Ecosystem dynamics were measured in terms of alkalinity, conductivity, dissolved nitrogen, dissolved oxygen, pH, dissolved phosphorus, and turbidity. A previous study on the same experimental study system measured phytoplankton trait evenness (using flow cytometry and machine learning), algal biomass stoichiometry, and zooplankton abundance to study the same interactive effect, but without species-level resolution of the microbial communities [91].

Only the prokaryotic communities responded systematically to nutrient disturbance and foundation species

Nutrient disturbance, representing eutrophication or the gradual build-up of phosphorus, nitrogen, and nutrients in aquatic ecosystems, led to a strong directional shift in the prokaryotic community composition (Figure 2A). The rate of change of prokaryotic community composition was highest during the nutrient disturbance and was stabilised post-disturbance. In contrast, the eukaryotic community composition shifted stochastically as a result of the nutrient disturbance (Figures 2B 2C). Variance among the replicates' eukaryotic communities and the rate of change of their compositions remained high throughout the sampling schedule: pre-disturbance, during disturbance, post-disturbance, and the subsequent winter. While there existed dominant families of prokaryotes before and after nutrient disturbance, no such dominance was observed among the eukaryotes.

The presence of the two foundation species, individually or in combination, affected the prokaryotic communities' trajectories (Figure 2A). PERMANOVA was used to determine that the final compositions of the prokaryotic communities were

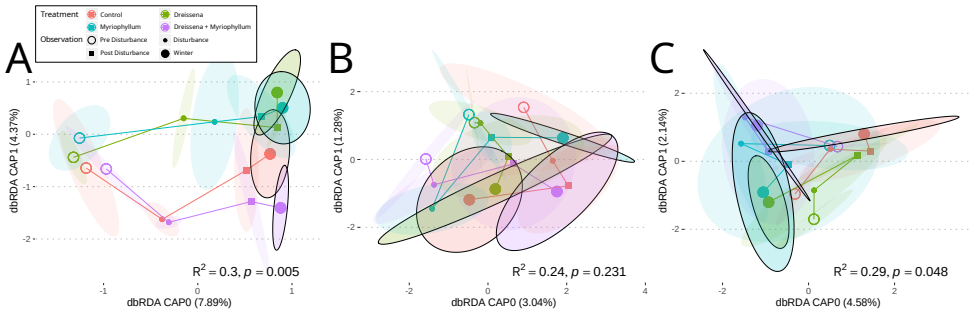

Figure 2. Mean community trajectory visualized after RDA for (A) prokaryotes via 16S, and micro-eukaryotes via (B) 18S and (C) COI amplicon sequencing.

significantly different depending on the combinations of the two foundation species present. However, no such directional effect was seen on the eukaryotic communities' trajectories or their final compositions (Figures 2B 2C). The prokaryotic communities were more stable, had multi-week trends, and appeared to converge to a steady state post-disturbance, whereas the eukaryotic communities were unstable, varied week-to-week, and did not converge.

Prokaryotic community composition covaries with environment

When considering both exogenous (dissolved N_2 and P) and endogenous (dissolved O_2, turbidity, pH, alkalinity, conductivity) factors, changes in ecosystem-level properties correlated highly with changes in prokaryotic community compositions. The correlation was lower for eukaryotic communities. The highest correlations of changes between prokaryotic communities and the environment (as well as the biggest differences between these correlations for prokaryotic and eukaryotic communities) was observed in oligotrophic ponds followed by ponds with both *Dreissena* and *Myriophyllum* or with only *Myriophyllum*.

Presence of both foundation species influences most ecosystem properties post-disturbance

At the beginning of the experiment, all the ponds had clear water, oxygen levels close to saturation, and low phosphorus and nitrogen levels. As the nutrient supplements were added, phosphorus levels increased steadily and remained high even after the supplements stopped, especially for the ponds with both foundation species. The supplemented nitrogen, however, was absorbed quickly and remained scarce throughout. The nutrient supplements led to supersaturation of oxygen and turbid waters. In oligotrophic ponds, phosphorus and nitrogen was scarce throughout re-

sulting in clear water and oxygen levels close to saturation throughout as well.

The nutrient disturbance caused a destabilization of the prokaryotic community composition, i.e., a higher successive change in community composition from its previous state. Post-disturbance, the prokaryotic community composition was stabilized, or the successive change was reduced, especially for the ponds with both *Dreissena* and *Myriophyllum* and less so for ponds with either one present alone.

Hypothesised causal relationships between the foundation species' presence, prokaryotic and eukaryotic communities, and endogenous and exogenous ecosystem properties were investigated using piecewise structural equation modelling. Post-disturbance, after the treatment had time to affect the system, the presence of both *Dreissena* and *Myriophyllum* had a significant negative effect on the successive change in prokaryotic community composition, indicating a stabilizing effect (Figure 3). The prokaryotic community change directly affected the eukaryotic community change. The foundation species and community structure further affected biomass properties like total phytoplankton, chlorophyll *a*, and total cell concentration. All of these, in turn, affected endogenous ecosystem properties like dissolved oxygen levels and pH. The structural equation model best explained the variance of these endogenous ecosystem properties (Figure 3).

In this study, it was possible to determine that the individual and interactive effects of nutrient disturbance and foundation species on aquatic ecosystems are mediated by prokaryotic and eukaryotic communities. However, the mechanistic details of this mediation and the contribution of individual microbial taxa is uncertain. Deciphering these details required a new type of methodology, focussing on cellular heterogeneity and functional behaviour of individual microbial community members, prototyped in chapter **II**.

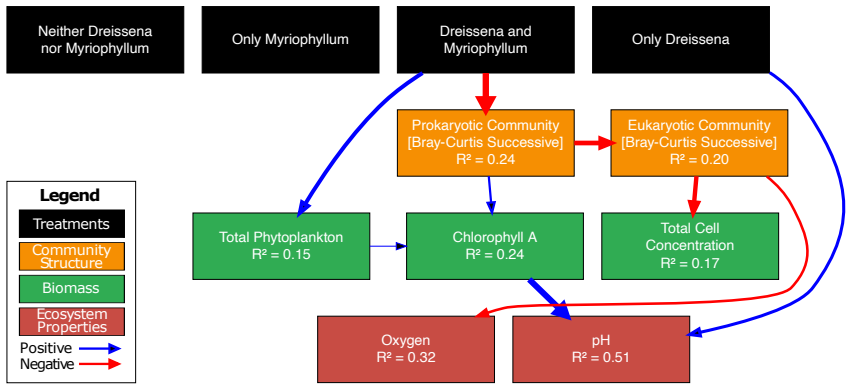

Figure 3. The best fit structural equation model for the time interval post nutrient additions

3.2 Metabolism and prokaryotic associates' composition of *Ochromonas triangulata* varies by growth stage

Microeukaryotes are complex organisms and perform various ecosystem functions under different environmental conditions. Chapter **II** involves our investigation of the metabolism and interactions of microeukaryotes at the single-cell level. Non-axenic cultures of *Ochromonas triangulata* were sampled at an early fast-growth stage and a late slow-growth stage (Figure 4A). 380 cells from each sample were sorted based on chlorophyll *a* and LysoSensor Blue markers. Single-cell transcriptomic libraries were prepared using the Smart-seq2 protocol. Thus, we obtained FACS data as well as sequence reads from full-length transcripts of individual cells. Reference genome for another species from the same genus was available, but very few reads mapped to this reference. Therefore, *de novo* transcriptome assembly and annotation was performed.

Core metabolic activity occurs in three distinct levels among the two sampled stages

Transcriptome assembly by pooling reads from all the *O. triangulata* cells, filtering out contigs that were present in only one cell (indicative of misassembly), and filtering out cells that have too few reads or are not representative of the assembled transcriptome yielded gene expression that clustered clearly into three groups: the two expected expression clusters of fast-growing and slow-growing cells, and a third unexpected expression cluster which we referred to as the uncharacterised cluster (Figure 4B). Comparing the gene expression among the three clusters in a pairwise manner yielded 538 differentially expressed genes, most of which had high expression in the fast-growing cluster, intermediate expression in the slow-growing cluster,

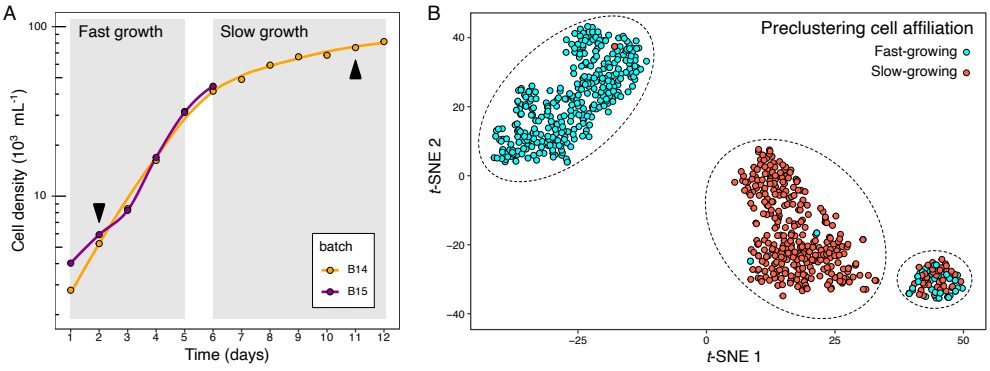

Figure 4. (A) Growth curves of source cultures for the *Ochromonas triangulata* cells. (B) t-SNE visualization of expression clusters.

and low expression in the uncharacterised cluster.

Metabolic mapping against the KEGG Orthology pathways detected the three distinct levels of expression among four main pathway families: transcription and translation, carbohydrate metabolism, photosynthetic activity, and endocytosis and lysosomal activity. Downregulation of genes in transcription and translation associated pathways agrees with the overall lower gene expression seen in the uncharacterised cells. Widespread downregulation of genes involved in carbon catabolism related pathways is indicative of a harsher energy landscape and worse cell maintenance for the uncharacterised cells (Figure 5A). Higher expression, in fast-growing cells, of genes involved in chlorophyll *a* synthesis and carbon fixation is indicative of the early fast-growth in *O. triangulata* being powered by photosynthesis (Figures 5B & 5C). Lastly, higher expression of lysosomal proteases in fast- and slow-growing cells is indicative of active lysosomal digestion in growing cells, but absent in the uncharacterised cells. Altogether, one possible explanation of the uncharacterised cells is that they are close to dormancy or cyst forming/breaking leading to observations of lower transcriptional activity and cellular metabolism.

Structural homology analysis upholds sequence homology-based inferences

Structural homology based annotation of predicted proteins folded using ESMFold yielded a much higher number of annotated transcripts compared to sequence homology, but these transcripts still represented a similar quantity of sequenced reads as the sequence homology based annotated transcripts. Pairwise comparison across clusters yielded 1,397 differentially expressed genes, a higher number than DEA based on only sequence homology based annotation. Metabolic mapping against KEGG Orthology pathways again detected differential expression in a similar set of pathways, upholding previous inferences, and additionally detected two new differential pathway families: cell cycle and oxidative stress. Identification of differential expression in these pathways was also bolstered by enrichment of associated GO gene sets, enabled by direct structure-based gene ontology prediction. Gene set enrichment was also detected in some pathways related to bacterial symbiosis like symbiotic interaction (GO:0044403) and to bacterial invasion like entry into host (GO:0044409).

rRNA read carry-over enables identification of microeukaryote and distinct associated prokaryotes

The non-specific rRNA reads recovered from the cells successfully identified the fast-growing and slow-growing *Ochromonas triangulata* cells. The number of rRNA reads recovered from the uncharacterised cells was much lower and consequently precluded accurate taxonomic identification. The accurate identification of microeukary-

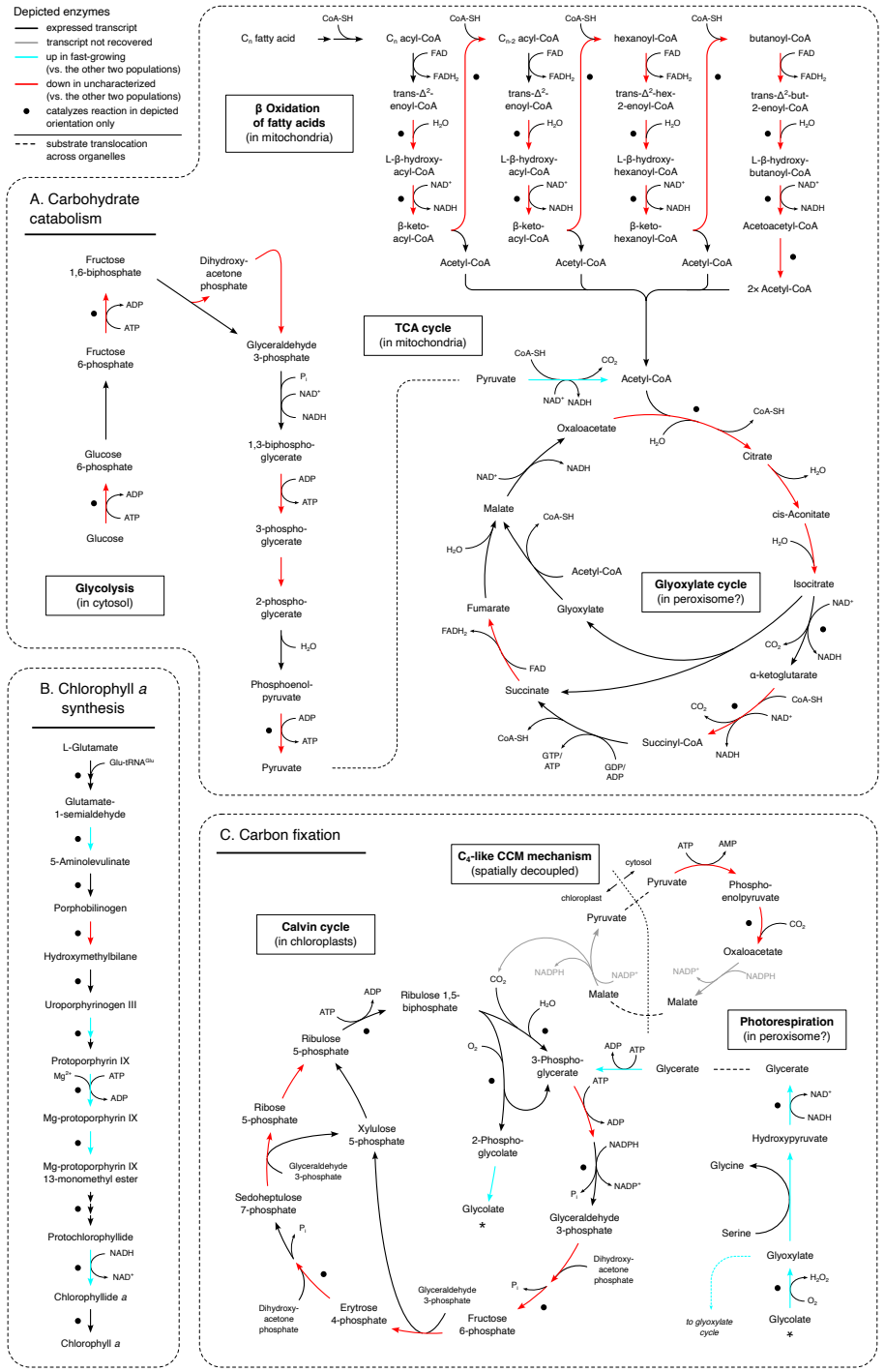

Figure 5. Differential expression in key metabolic pathways.

otes from the carry-over rRNA of single-cell transcriptomic data indicated the possibility of identifying *a priori* unknown microeukaryotes from bulk sampling of microbial communities, as tested in chapter **III**.

Compared to 18S reads, the abundance of 16S reads recovered was much lower, but sufficient to characterize the prokaryotic associates of the individual microeukaryotes. All downstream analyses of the prokaryotic communities were possible. The alpha diversity of the prokaryotic communities associated with cells in the three clusters were different, highest in association with fast-growing cells and lowest in association with the uncharacterised cells. The prokaryotic communities were similar within the clusters and different between each pair of clusters, with most internal uniformity seen in the uncharacterised cluster. The ratio of 16S rRNA reads to all rRNA reads was highest in uncharacterised cluster and *Endomicrobiaceae* was among the differentially abundant prokaryote families that was more abundant in association with the uncharacterised cells compared to cells from the other clusters. Thus, another possible explanation of the uncharacterised cluster is an invasion of the microeukaryote by specific prokaryotes.

3.3 Microeukaryote-prokaryote associations persist through changing light and expression conditions

The observation in chapter **II** that scRNAseq of microeukaryotes allows for their taxonomic identification using the carry-over reads led us to test the efficacy of the method in a natural environment with higher diversity of *a priori* unidentified microeukaryotes in chapter **III**. Natural samples from Lake Långsjön (Stockholm, Sweden) were acclimated in the lab for two days, split into two treatment groups, one receiving illumination and the other shading, and sampled at four time points, 0, 6, 24, and 48 hours after onset of treatment. Samples were taken for single-cell transcriptomic library prep (190 cells per time point per treatment), and 16S and 18S rRNA gene amplicon sequencing (for each treatment and time point separately as well as from initial unacclimatized sample). The amplicon sequencing data was unavailable at the time of reporting these results. They will be used to validate the taxonomic inferences made from the carry-over rRNA reads.

Gene expression clusters by taxonomic origin of the cells

Of the 1,520 single-cell libraries sequenced, 22 distinct taxa (at the genus or family level) were detected based on the non-specific recovery of the non-target carry-over rRNA reads. Thus, 22 *de novo* transcriptomes were assembled. This is a significant expansion in the breadth of sampled biodiversity as compared to the previous study with only *Ochromonas triangulata*. The taxa identified, and transcriptomes assembled for, were the following (with number of cells sequenced in parentheses):

Rhodomonas sp. (490), *Cryptomonadales* (309), *Lecanicillium psalliotae* (148), *Isaria farinosa* (114), *Pedinellales* (97), *Fragilariopsis kerguelensis* (70), *Cylindrotheca sp.* (64), *Paraphysomonas sp.* (49), *Kathablepharidida* (32), *Askenasia sp.* (26), *Stephanodiscus hantzschii* (23), *Paraphysomonas uniformis* (11), *Tetrahymena rostrata* (11), *Olisthodiscs luteus* (8), *Ochromonadales* (7), *Pelagomonas calceolata* (7), *Chlorophyta* (3), *Chrysochromulina sp.* (3), *Pirsoniales* (3), *Haslea sp.* (1), *Paulinella chromatophora* (1), and *Telonemia* (1).

The *de novo* transcriptome assembly and the annotation was done separately for each of these distinct 22 sample subsets. Therefore, to compare gene expression between the different transcriptomes, only those genes that were expressed in common among the different taxa (with the additional impediments of successful detection in the libraries and successful annotation in computational analysis) could be used for comparison. In this study, the 12 biggest annotated transcriptomes, comprising 88.7% of the 1,478 sampled cells, had 98 gene expressed in common. The expression of these 98 genes were used to compare the 12 transcriptomes. Including all the other smaller transcriptomes, with very few annotated and detected genes, led to no gene expressed in common and consequently no comparison. Visualization and clustering of gene expression from these 12 sampled taxa revealed clustering corresponding to the taxonomic origin of the samples (Figure 6).

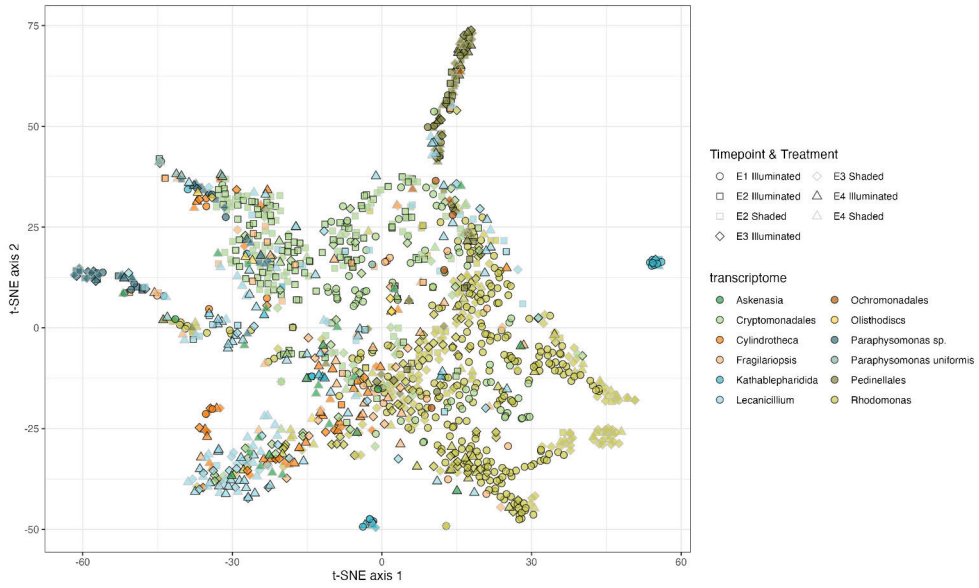

Figure 6. t-SNE visualization of the gene expression based on genes expressed commonly among the 12 largest transcriptomes in the study.

Shading results in downregulation of photosynthesis, lysosome, and carbon metabolism related pathways

Rhodomonas, the most abundant mixotroph among the sequenced cells, saw a dip in transcription six hours after the onset of shading treatment. Among the *Rhodomonas* samples from the shading treatment, compared to the cells sampled right at the onset of the treatment, 275 genes were downregulated and four upregulated six hours later, indicating a dip in transcriptional activity. This transcriptional activity partially recovered eighteen more hours later with 95 fewer genes downregulated and 61 more upregulated. Twenty-four more hours later, more recovery was seen with still 95 fewer genes downregulated.

Metabolic mapping of genes differentially expressed between the illuminated and shaded treatment samples of *Rhodomonas* cells led to identification of metabolic downregulation due to shading in five main KEGG Orthology pathway families: transcription and translation, photosynthetic activity, cell metabolism, cell stress response, and endocytic activity (Figure 7). Except for photosynthetic activity, all the differentially expressed genes involved in these pathway families were downregulated due to shading. Widespread downregulation of transcription and translation related pathways agrees with the dip in the transcriptional activity detected in the shaded samples. Higher expression of protochlorophyllide reductase, light-harvesting complex I proteins, and photosystem II proteins alludes to *Rhodomonas* cells trying to cope with reduced light conditions, but failing to do so as seen by the downregulation of Calvin cycle proteins indicating lower energy capture success. Shading also led to a widespread downregulation of cellular metabolism as seen from the pathways of oxidative phosphorylation, glycolysis, citrate cycle, and oxidation of fatty acids reinforcing the inference of a harsher energy landscape under shaded conditions.

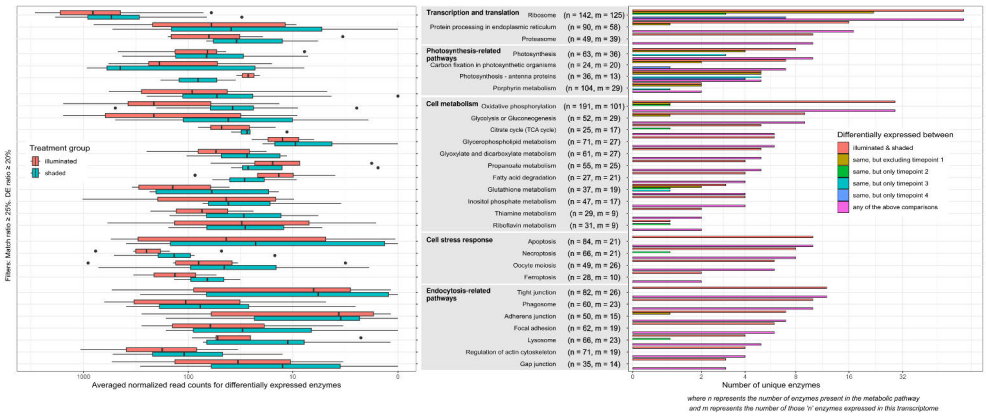

Figure 7. Summary of KEGG pathways associated to the differentially expressed genes.

Microeukaryote-prokaryote association unaffected by light conditions

As with the taxonomic identity of the microeukaryotes in the previous section, the taxonomic identity of their prokaryotic associates were detected based on the non-specific recovery of the non-target carry-over rRNA reads. *Endomicrobiaceae*, *Leptolyngbyaceae*, and *Microcystaceae* were the three most common families of prokaryotic associates (Figure 8). RDA and PERMANOVA showed that the composition of the prokaryotic associates was most affected by the microeukaryote identity and then by sampling time point. Neither the relative abundance of the prokaryotes nor their alpha diversity was affected by the treatment conditions. Many prokaryote families, including the most common ones listed above, were differentially abundant in their associations with the different microeukaryotes, for example, *Leptolyngbyaceae* was associated with *Rhodomonas* and *Cryptomonadales*, *Endomicrobiaceae* with *Askenasia*, *Cylindrotheca*, and *Paraphysomonas*, and *Microcystaceae* with *Lecanicillium*.

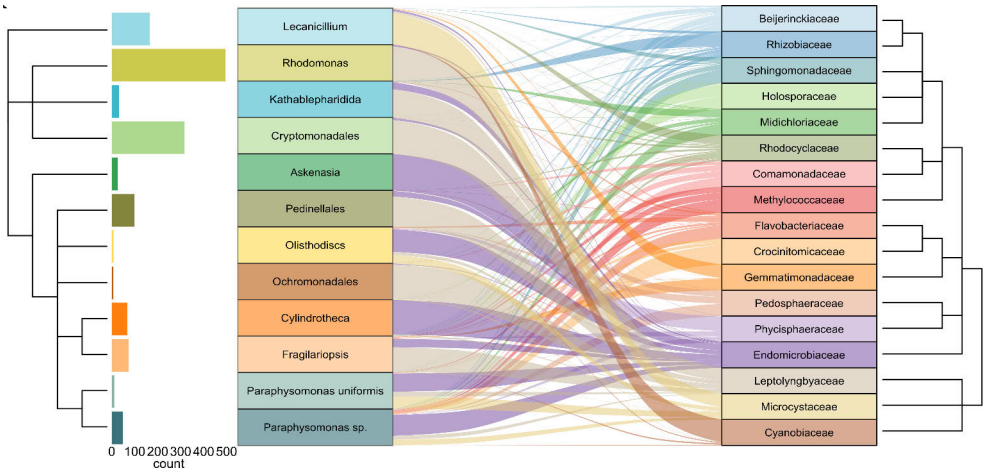

Figure 8. Alluvial plot indicating each microeukaryote's prokaryotic associations.

The initial taxonomic identity assignment was done via rRNA reads detected using RiboDetector. This was validated by rRNA transcript detection from the *de novo* transcriptomes using Infernal. Further validation will be done using 16S and 18S rRNA amplicon sequencing when the data becomes available.

4 Discussion

4.1 Why study microeukaryotes and microeukaryote-prokaryote interactions?

Prokaryotes have existed on Earth for nearly four billion years. Microeukaryotes and microeukaryote-prokaryote interactions have persisted on earth for over two billion years. Yet, we are still unaware of exactly who is present where, what they do, and who interacts with whom. Multiple supergroup level taxa of eukaryotes have been discovered and added to the known tree of life only in the last decade [16]. An overwhelming >90% of known species of bacteria and microeukaryotes have never been isolated in laboratory conditions [80]. Representativeness of the inferences about behaviours and interactions of species grown in laboratory conditions to their behaviours and interactions in nature have not been validated [18]. Thus, it is essential to study microeukaryotes and microeukaryote-prokaryote interactions, especially in culture-free natural conditions, to broaden and deepen our knowledge of natural microbial diversity and fill in the gaps in the tree of life and learn what each of the leaves do.

Microeukaryotes have been of human utility from times beyond recorded history in the form of baker's yeast and brewer's yeast (different strains of the microeukaryote *Saccharomyces cerevisiae*). They are components of all normal human-, animal-, and plant-associated microbiomes and are necessary for normal functioning of bodily functions. Many antibacterials (most famously penicillin), antifungals, and antivirals have been isolated from them. Today, they are used in wastewater plants, industrial fermenters, and in various industrial bioreactors to produce foods (such as yogurts, single cell proteins), organic acids, enzymes, biofuels, pharmaceuticals, synthetic polymers, and more. While many of these molecular discoveries and uses have arisen from cultivable model microeukaryotes, uncultivable environmental microeukaryotes hold the potential for a plethora of more utilitarian discoveries. For example, there is a case developing for more research into the role of environmental microeukaryotes in the mitigation of the ongoing microplastic pollution in all ecosystems [239; 240].

The study of model microeukaryotes in axenic cultures has developed our understanding of their molecular biology in great detail, but their representativeness of heterotrophic and mixotrophic microeukaryotes that cannot be cultured axenically is

debatable. No organism lives naturally in isolation. Many microeukaryotes possibly behave and interact differently in the lab when compared to their natural habitats [18]. Thus, it is necessary to study microeukaryotes and their function in the context of the ecosystem in which they exist, inclusive of symbionts and prey.

Microeukaryote-prokaryote interactions may be essential for ecosystem stability. For instance, consider a macroecological analogy. In the ecosystems that they occur in, wolves occupy the role of apex predators. When wolves are eliminated from the habitat, in just a few years, the population of all the wolves' prey (like moose and deer) will increase immensely which can graze unchecked and reduce biodiversity, cause riverbank erosion, and reduce fish habitats, and consequently affect fish available for human consumption [241; 242]. We now know how essential wolves are for ecosystem stability and human utility because of a clear understanding of what wolves do and how they interact with others in the ecosystem. Similarly, we need to learn what microeukaryotes in aquatic ecosystems do, who they interact with, and how, to learn how essential they are for ecosystem stability as well as human utility [243].

From the bacteria-archaea interaction based endosymbiotic origin of eukaryotes to trophic interactions between photoautotrophs and chemolithoautotrophs and heterotrophs and mixotrophs in all ecosystems to human gut microbiome interactions that influence nutrition and brain activity, cross-domain interactions between organisms is imperative for life on Earth. In many ecosystems, microeukaryotes and prokaryotes interact syntrophically and depend on one another for survival. Identifying these interactions, especially in natural conditions and with functional context, generates new knowledge as the basis for better understanding of ecosystems and for future research and innovations, such as in environmental monitoring and mitigation.

4.2 Microeukaryote composition is affected by prokaryote composition

In chapter **I**, prokaryotic and microeukaryotic community compositions were measured in 20 artificial ponds over several time points before, during, and after nutrient disturbance. Changes in the composition of the microeukaryotic community were influenced by changes in the prokaryotic community. This change, whether measured as the immediate change from the previous time point or as the change from the initial composition in the pond, mediated downstream outcomes like increase in turbidity and decrease in dissolved oxygen.

Foundation species, such as *Dreissena polymorpha* and *Myriophyllum spicatum* used in this study, are known to have direct and indirect impacts on aquatic communities and ecosystem functions. Both are known to increase water clarity, by filtering or stabilizing sediments, and provide physical structure [96; 97; 98]. Critical transition of the ecosystem toward an alternative turbid stable state following nutrient

disturbance was proposed previously based on temporal autocorrelation [244; 245]. The prokaryotic and eukaryotic community dynamics measured by the amplicon sequencing adds further evidence to this. The presence of multiple foundation species, whether as invasive species or as restorative management tools, can have unanticipated interactive effects, such as stabilizing turbid eutrophic states.

Community composition in the ponds, at various time intervals, was measured using 16S, 18S, and COI amplicon sequencing which only provides broad characterization of prokaryotes and eukaryotes present in the ecosystem and an estimate of their abundance. Statistical testing and structural equation modelling was used to test the hypothesis that prokaryotic community stabilization, measured by successive changes in the community composition, led to eukaryotic community stabilization. This hypothesis was based on preexisting knowledge of foundation species affecting microbial communities and trophic interactions from prokaryotes to microeukaryotes to grazers like cladocerans and copepods. However, no symbiotic interactions between the microeukaryotes and prokaryotes could be inferred or tested here.

4.3 Reference-free function and interaction analyses are possible

The inability of amplicon sequencing in chapter **I** to elucidate individual microeukaryote-prokaryote interactions led us to microeukaryote single cell isolation. *Ochromonas triangulata* is a marine mixotrophic microeukaryote, i.e, it can survive by either photosynthesis or preying on bacteria. Although available in non-axenic pure culture (i.e., culture contains *O. triangulata* and associated bacteria) from the Roscoff Culture Collection, it has no available reference genome. Single cell isolation of the microeukaryote and full-length transcriptomic sequencing using Smart-seq2 enabled us to assemble *de novo* partial transcriptomes to infer the functional activity of the microeukaryotes as well as use the carry-over rRNA reads to identify their prokaryotic associates, at single-cell resolution.

Analogous to the microbial dark matter (majority of microorganisms not available in culture) [246; 247], dark matter of the genome (majority of the genome not coding for proteins or regulatory mechanisms) [248], and dark genome regions (regions of the genome that cannot adequately be assembled or aligned to) [249], we observed and were limited by the dark expression or a majority of the assembled transcripts having no known function. Only around 11% of the sequenced reads contributed to our understanding of the active functions in the microeukaryote. Prediction of the tertiary structures of the proteins and their annotation using structural homology yielded more transcripts with known function, but their read representation still remained close to 11%. This could be mainly due to two reasons, one, that many of the sequence reads come from pre-mRNA which are yet to undergo splicing and other transcriptional editing before they represent mature mRNA and

the corresponding function, and two, that gene, transcript, and protein sequences from environmental microeukaryotes are vastly underrepresented in publicly available sequence databases. Despite this, it was possible to perform differential gene expression analysis, metabolic mapping, and gene set enrichment analysis to decipher microeukaryote functions, for example, associated with photosynthetic activity and carbon metabolism.

O. triangulata cells from two distinct growth stages, an early fast-growing stage and a late slow-growing stage, were sampled and sequenced. But gene expression clustering revealed three distinct clusters, two corresponding to the two growth stages and the third comprising of cells from both growth stages. This third uncharacterized cluster contained cells that, compared to those in fast-growing and slow-growing clusters, exhibited lower photosynthetic activity, carbon fixation, and phagocytosis. Based on this functional characterisation and the knowledge that the fluorescence activated cell sorting selected for live cells, we hypothesised that these might be cells transitioning toward or away from dormancy. This would explain their presence among both growth stages, coming out of dormancy among fast-growing cells amid higher abundance of prey and going into dormancy among slow-growing cells amid lower abundance of prey.

Ribosomal RNA carry-over due to the higher abundance of rRNA compared to mRNA in the cells, despite mRNA targeting using poly-dT oligonucleotide primers, enabled the identification of prokaryotic associates of individual *O. triangulata* cells. Each microeukaryote was associated with a diverse assemblage of bacteria, mostly assumed to be prey, but the composition of the associated bacterial community was quite uniform within the three clusters, especially so for the uncharacterised cluster. The difference in the associated bacteria between fast-growing and slow-growing clusters can be explained by higher abundance of bacteria and lower grazing pressure in the early sample with fewer fast-growing microeukaryotes and higher grazing pressure and lower abundance of, potentially grazing-resistant, bacteria in the late sample with more microeukaryotes. The distinct, yet uniform, bacterial community associated with the uncharacterised cells obtained from the two distinct samples proposes another hypothesis: that these might be cells under colonisation or invasion from bacteria. The presence of *Endomicrobiaceae* bacteria, known endosymbionts of protozoa, adds to this hypothesis.

4.4 Culture-free and reference-free study of environmental microeukaryote functions and interactions is nearly possible

In addition to the taxonomic identification of associated prokaryotes, the carryover rRNA in chapter **II** also enabled the accurate identification of the (*a priori* known) microeukaryote from just single-cell transcriptomic data. This showed us the pos-

sibility of identifying microeukaryotes and their individual-level prokaryotic associates along with their functional activity **without** *a priori* knowledge of the microeukaryotes and consequently the potential for characterising natural microeukaryote diversity. However, from chapter **II** and previous studies, we know that single-cell RNA reads from approximately 100 cells are needed to assemble a satisfactory *de novo* transcriptome due to low quantity of RNA in cells [250] and transcription occurring in bursts [251]. Thus, we limited the breadth of natural microeukaryote diversity in our samples by selecting for only mixotrophs, i.e., ones expressing chlorophyll *a* and containing acidic vacuoles representative of active feeding.

Transcriptomes of 1,520 microeukaryotes from light and shaded treatment groups were sequenced, taxonomically identified (using rRNA reads) to belong to 22 families across five supergroups, and assembled *de novo* into 22 transcriptomes. Transcripts were annotated using sequence and protein structure homology. Similar to chapter **II**, we observed dark expression again with functional annotation available for less than 10% of the transcripts in the *Rhodomonas* transcriptome (comprising 491 out of 1,520 sampled cells). The *Rhodomonas* transcriptome contained only half the OrthoDB eukaryota conserved orthologous genes indicating incomplete recovery due to library prep shortcomings, insufficient sequencing depth, or lack of their expression at the sampling time. The *Kathablepharidida* transcriptome, assembled from only 32 cells but $3.25\times$ the number of reads per cell, contained nearly the same number of orthologous genes, indicating that more complete transcriptomes could be assembled by sequencing more deeply. Overall, we believe that 100 million short reads retrieved across multiple single cells is sufficient to assemble a satisfactory *de novo* transcriptome.

Each microeukaryote hosted communities of diverse co-existing prokaryotic associates, similar to observations in other natural microeukaryote populations [20; 21; 18; 22]. Microeukaryotes within the same supergroup hosted the same prokaryotes, unlike previous observations of microeukaryotes of the same species in the same population hosting distinct prokaryotic symbionts [21]. Cryptophytes (*Rhodomonas*, *Kathablepharidida*, *Cryptomonadales*) were most associated with *Leptolyngbyaceae* bacteria indicating similar feeding behaviour across diverse cryptophytes. Alveolates (*Askenasia*) and Stramenopiles (*Pedinellales*, *Olisthodiscs*, *Cylindrotheca*, *Fragilariopsis*, *Paraphysomonas*) were associated with *Endomicrobiaceae* bacteria indicating widespread intracellular prokaryotic symbiosis across the TSAR supergroup. These microeukaryote-prokaryote associations were stable across treatments and sampling intervals.

4.5 Sequencing targets and their implications

In a cautionary review about the use of molecular sequencing tools for the study of microeukaryotes, Keeling and Del Campo [252] talk about how little we know about

ecosystem-level compositions, interactions, and functions of microbial communities in nature. While amplicon sequencing methods (with their initial biases and later corrections) have added to our knowledge of community compositions, knowledge of individual components' functions and interactions are still limited. Great collaborative efforts have been made to cultivate environmental microeukaryotes, develop genetic tools for them, and create new model organisms from a diverse range of protist taxa to generate insights into general eukaryote cell biology and evolution of cellular pathways [253]. But the vast majority of microeukaryotes are uncultivated [80]. Sequence data might provide the only realistic means of acquiring any biological observations about uncultivated protists at all [252].

Across prokaryotes and microeukaryotes, small subunit ribosomal RNA (SSU rRNA) has been the most common target choice for amplicon sequencing. And across both prokaryotes and microeukaryotes, biases associated with amplicon sequencing exist [254; 70]. Due to the prevalence of short-read sequencing, small hypervariable regions of the 18S SSU rRNA, particularly the V4 and V9 regions, have been used for inferring microeukaryote community composition [252]. The choice of amplicon target influences the observation of community composition, for example, sequencing amplicons of the V4 region excludes euglenozoans [255; 252]. Different microeukaryote taxa can have from just 30 copies of the SSU rRNA gene to over 12,000 copies [256; 257; 252]. Differences like these, in addition to sequence amplification bias, can severely bias the inferred relative abundances of community components. Amplicon sequencing also cannot differentiate between (potentially more important) microbes active functionally and interactively and those lying dormant, instead using a simpler paradigm of higher abundance equating to higher importance [252]. These drawbacks aside, amplicon sequencing can still provide broad inexpensive community-level information about diversity and co-occurrence, but falls well short of providing meaningful insights about functions and interactions.

Some of the above shortcomings and biases are mitigated by using single-cell transcriptomics or scRNAseq. For starters, it adds the whole paradigm of functional activity that was unavailable from amplicon sequencing data. This enables the identification of sub-population level functional profiles, by means of gene expression clustering, differential expression analyses, and metabolic mapping. By including UMIs (chapter **III** vs **II**), accuracy of these inferences are improved further. Even for the task of taxonomic identification, scRNAseq forgoes amplification and sequence target bias by generating random rRNA gene fragments. Although inefficient due to no targeted amplification of particular gene regions, it has less likelihood of bias. The taxonomic identification of the microeukaryotes enables broad culture-free study of environmental microeukaryotes, similar to amplicon sequencing, however additionally, the taxonomic identification of the individual-level prokaryotic associates enables new inferences of microeukaryote-prokaryote interactions beyond co-occurrence.

Similar to scRNAseq, single-cell genomics as well can be used to infer identities and interactions, in addition to functional potential instead of functional activity. Although genome-based phylogenetic classification provides greater resolution than the rRNA gene (or the even smaller individual hypervariable region) due to the comparison of a larger fraction of the genome, the extensive history of horizontal gene transfer complicates it [42]. Additionally, the order of magnitude difference between microeukaryote and prokaryote genomes sizes and additionally the loss of genes and reduction of genomes in prokaryotic symbionts [258] makes the simultaneous sequencing of microeukaryote and prokaryote genomes lopsided and inferences of interactions difficult. The inference of functional potential is also hindered by the complex transcriptional mechanisms, as described in section 1.2. Nevertheless, in an ongoing project, we are investigating the utility and the various analysis methods of single-cell genomics and single assembled genomes for inferring microeukaryote-prokaryote interactions and functional complementation potential.

All of the aforementioned sequencing techniques generally use short sequence reads of approximately 150 bp. Assembly of these short sequence reads into transcripts or genes and other essential steps in analysis of sequencing data such as identification of splice sites, coding regions, and alternative isoforms, and annotation of coding sequences are challenging in single-cell transcriptomics and genomics using short sequence reads. A potential solution to this is the use of long read sequencing (such as offerings from Oxford Nanopore Technologies, Pacific Biosciences, MGI Tech, and Roche) to sequence whole transcripts and skip assembly altogether. Over the past decade, there have been attempts of using this methodology, but they have been hampered by transcript length bias, low read throughput, and insufficient read accuracy [259]. With newer long-read sequencers and sequencing library preparation protocols and mitigation of bias and accuracy deficiencies, the era of long read sequencing for transcriptomics [177] as well as rRNA amplicon sequencing [260] may soon be at hand.

4.6 Single-cell transcriptomics for culture-free reference-free microeukaryote study: Potential and planned steps

From chapters **II** and **III**, we saw that single-cell transcriptomics has the potential to decipher microeukaryotic diversity in complex natural ecosystems, in terms of population-level taxonomic identity, sub-population-level transcriptomic profiles and metabolic states, and individual-level prokaryotic associations. However, such decipherment of the entire microeukaryote diversity is limited by throughput. With the need to sequence transcripts from around 100 cells and obtain around 100M reads from each microeukaryote species in a population, deciphering functions and interactions for all microeukaryotes in complex natural ecosystems comprising 100s to

1000s of species of microeukaryotes will require isolating over 100,000 cells and sequencing billions of reads.

Methods like Smart-seq that rely on FACS for high-throughput cell isolation in microplate wells do not scale well due to limitations of human, material, and financial resources. Alternative high-throughput full-length single-cell transcriptomics methods which do scale to million cells or more exist or are in development, but have not been tested on environmental microeukaryotes. Two such families of methods are described below.

Split-pool barcoding based methods chemically fix and permeabilize cells such that each individual cell can be used as a separate reaction vessel [261; 262]. Then, the sample of fixed cells is split across 96 wells, different oligonucleotide barcodes are attached to cDNA in different wells, and cells across different wells are then pooled together. This split-pool cycle is repeated 3-4 times. After the final pooling, library prep, and sequencing, the unique barcode sequences can be used to distinguish transcript read sequences originating from over a million distinct cells. Split-pool barcoding based scRNAseq kits are commercially available through Parse Biosciences and Scale Biosciences.

Droplet based methods generate thousands to millions of droplets capturing a single cell in a droplet, lyse the cells, and uniquely barcode the cDNA from each droplet using random oligonucleotides [263; 264]. While older methods relied on sophisticated microfluidics devices to generate the droplets, PIP-seq uses a standard benchtop vortexer for droplet generation and encapsulation of cells [265]. sc-rDSeq combines droplet generation using microfluidics, split-and-pool barcoding, and not-so-random hexamer primers that deplete rRNA sequences, for full-length total RNA sequencing [266].

Apart from throughput, another limitation in the previously used methods (Smart-seq2 and Smart-seq3xpress) is the targeting of protein-coding mRNA. Although high abundance of rRNA led to obtaining carryover rRNA sequences in our data, microRNA and long non-coding RNA etc. were lost. Non-coding RNA is transcribed from nearly 70% of the genome [267]. Methods like Smart-seq-total [268], VASA-seq [269], and TotalX [270] expand the single-cell profiling to beyond the protein-coding mRNA.

A caveat to using single-cell transcriptomics to understand microeukaryote-prokaryote interactions is the limited ability to resolve the type of interactions without prior knowledge of the microeukaryotes' behaviour. In both chapters **II** and **III**, we identified prokaryotic associates of individual microeukaryotes, but could not definitively resolve where on the spectrum of microeukaryote-prokaryotic interactions–from syntrophy to bacterial farming [18]–each one of these associations occurred in. Our inferences about these interactions are based on the *a priori* known mixotrophic or heterotrophic behaviour of the microeukaryotes studied as well as some signals confirming these behaviours. Further improvement of computational methodologies

related to functional annotation and metabolic mapping has the potential to reduce the burden on prior knowledge of microeukaryote behaviour.

In chapter **III**, we demonstrated the potential of scRNAseq in the elucidation of microeukaryote functional activity and microeukaryote-prokaryote interactions in the mixotrophic subset of the natural microeukaryote community. In the future, we hope to expand this to the full natural diversity of microeukaryotes. As a first step toward that, we plan to test Parse Biosciences' WT Mini kit, based on split-and-pool barcoding, to sequence transcripts from 10,000 microeukaryotic cells. This methodology has previously only been tested on mammalian cells and model microeukaryotes like *Saccharomyces cerevisiae*. Testing and validation is needed to mitigate challenges associated with complex cell walls, low transcript content, and potential sampling bias. But once optimized, it has the potential to easily scale to millions of cells.

5 Conclusion

Chapter **I** investigated broad interactions between prokaryotic and eukaryotic communities in aquatic ecosystems using amplicon sequencing. Chapter **II** investigated the utility of scRNAseq for reference-free microeukaryote and microeukaryote-prokaryote interaction study. Chapter **III** expanded this investigation to a reference-free and culture-free community of mixotrophic microeukaryotes.

In conclusion, aquatic ecosystems host a myriad variety of interacting microeukaryotes and prokaryotes, a majority of which do not have pure cultures or reference genomes and transcriptomes. Culture-free metagenome assembled genomes have added to the reference information for prokaryotes, however microeukaryotes are complex and do not easily accede to genome assembly and annotation from metagenomes. Here we demonstrate the utility of full length single cell transcriptome sequencing for reference-free and culture-free characterisation of microeukaryotes' functional activity and cross-domain interactions with prokaryotes.

List of References

[1] Claire Patterson. Age of meteorites and the earth. *Geochimica et Cosmochimica Acta*, 10(4): 230–237, October 1956. ISSN 0016-7037. doi: 10.1016/0016-7037(56)90036-9. URL http://dx.doi.org/10.1016/0016-7037(56)90036-9.

[2] Arthur Holmes. How old is the earth? *Transactions of the Edinburgh Geological Society*, 16 (3):313–333, January 1956. ISSN 2052-9414. doi: 10.1144/transed.16.3.313. URL http://dx.doi.org/10.1144/transed.16.3.313.

[3] Simon A. Wilde, John W. Valley, William H. Peck, and Colin M. Graham. Evidence from detrital zircons for the existence of continental crust and oceans on the earth 4.4 gyr ago. *Nature*, 409(6817):175–178, January 2001. ISSN 1476-4687. doi: 10.1038/35051550. URL http://dx.doi.org/10.1038/35051550.

[4] Antonio Lazcano and Stanley L Miller. The origin and early evolution of life: Prebiotic chemistry, the pre-rna world, and time. *Cell*, 85(6):793–798, June 1996. ISSN 0092-8674. doi: 10.1016/s0092-8674(00)81263-5. URL http://dx.doi.org/10.1016/S0092-8674(00)81263-5.

[5] Norio Kitadai and Shigenori Maruyama. Origins of building blocks of life: A review. *Geoscience Frontiers*, 9(4):1117–1153, July 2018. ISSN 1674-9871. doi: 10.1016/j.gsf.2017.07.007. URL http://dx.doi.org/10.1016/j.gsf.2017.07.007.

[6] Stanley L. Miller and Harold C. Urey. Organic compound synthesis on the primitive earth: Several questions about the origin of life have been answered, but much remains to be studied. *Science*, 130(3370):245–251, July 1959. ISSN 1095-9203. doi: 10.1126/science.130.3370.245. URL http://dx.doi.org/10.1126/science.130.3370.245.

[7] Christopher Chyba and Carl Sagan. Endogenous production, exogenous delivery and impact-shock synthesis of organic molecules: an inventory for the origins of life. *Nature*, 355(6356): 125–132, January 1992. ISSN 1476-4687. doi: 10.1038/355125a0. URL http://dx.doi.org/10.1038/355125a0.

[8] Abigail C. Allwood, Malcolm R. Walter, Balz S. Kamber, Craig P. Marshall, and Ian W. Burch. Stromatolite reef from the early archaean era of australia. *Nature*, 441(7094):714–718, June 2006. ISSN 1476-4687. doi: 10.1038/nature04764. URL http://dx.doi.org/10.1038/nature04764.

[9] C R Woese, O Kandler, and M L Wheelis. Towards a natural system of organisms: proposal for the domains archaea, bacteria, and eucarya. *Proceedings of the National Academy of Sciences*, 87(12):4576–4579, June 1990. ISSN 1091-6490. doi: 10.1073/pnas.87.12.4576. URL http://dx.doi.org/10.1073/pnas.87.12.4576.

[10] Paul G. Falkowski, Tom Fenchel, and Edward F. Delong. The microbial engines that drive earth's biogeochemical cycles. *Science*, 320(5879):1034–1039, May 2008. ISSN 1095-9203. doi: 10.1126/science.1153213. URL http://dx.doi.org/10.1126/science.1153213.

[11] Ramiro Logares, Stéphane Audic, David Bass, Lucie Bittner, Christophe Boutte, Richard Christen, Jean-Michel Claverie, Johan Decelle, John R. Dolan, Micah Dunthorn, Bente Edvardsen, Angélique Gobet, Wiebe H.C.F. Kooistra, Frédéric Mahé, Fabrice Not, Hiroyuki Ogata, Jan Pawlowski, Massimo C. Pernice, Sarah Romac, Kamran Shalchian-Tabrizi, Nathalie Simon, Thorsten Stoeck, Sébastien Santini, Raffaele Siano, Patrick Wincker, Adriana Zingone, Thomas A. Richards, Colomban de Vargas, and Ramon Massana. Patterns of Rare and Abun-

dant Marine Microbial Eukaryotes. *Current Biology*, 24(8):813–821, April 2014. ISSN 09609822. doi: 10.1016/j.cub.2014.02.050. URL https://linkinghub.elsevier.com/retrieve/pii/S0960982214002188.

[12] E. Haeckel. Generelle morphologie der organismen. 1866. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-0003485236&partnerID=40&md5=30bbead2b77d982a402e877090a67485.

[13] H.F. Copeland. The kingdoms of organisms. *The Quarterly Review of Biology*, 13(4):383 – 420, 1938. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-0043148540&partnerID=40&md5=7c54b12cf53c8e4c77095ddd2cd105f6.

[14] R.H. Whittaker. On the broad classification of organisms. *The Quarterly review of biology*, 34:210 – 226, 1959. doi: 10.1086/402733. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-0344872581&doi=10.1086%2f402733&partnerID=40&md5=a1b52f3e61beb18eed35ba8c8547b432.

[15] Brett J. Baker, Valerie De Anda, Kiley W. Seitz, Nina Dombrowski, Alyson E. Santoro, and Karen G. Lloyd. Diversity, ecology and evolution of archaea. *Nature Microbiology*, 5(7):887–900, May 2020. ISSN 2058-5276. doi: 10.1038/s41564-020-0715-z. URL http://dx.doi.org/10.1038/s41564-020-0715-z.

[16] Fabien Burki, Andrew J. Roger, Matthew W. Brown, and Alastair G. B. Simpson. The New Tree of Eukaryotes. *Trends in Ecology & Evolution*, 35(1):43–55, January 2020. ISSN 0169-5347. doi: 10.1016/j.tree.2019.08.008. URL https://www.cell.com/trends/ecology-evolution/abstract/S0169-5347(19)30257-5. Publisher: Elsevier.

[17] Laura A. Katz. Origin and diversification of eukaryotes. *Annual Review of Microbiology*, 66(1):411–427, October 2012. ISSN 1545-3251. doi: 10.1146/annurev-micro-090110-102808. URL http://dx.doi.org/10.1146/annurev-micro-090110-102808.

[18] Filip Husnik, Daria Tashyreva, Vittorio Boscaro, Emma E. George, Julius Lukeš, and Patrick J. Keeling. Bacterial and archaeal symbioses with protists. *Current Biology*, 31(13):R862–R877, July 2021. ISSN 0960-9822. doi: 10.1016/j.cub.2021.05.049. URL http://dx.doi.org/10.1016/j.cub.2021.05.049.

[19] William F. Martin, Sriram Garg, and Verena Zimorski. Endosymbiotic theories for eukaryote origin. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1678):20140330, September 2015. ISSN 1471-2970. doi: 10.1098/rstb.2014.0330. URL http://dx.doi.org/10.1098/rstb.2014.0330.

[20] Alessia Rossi, Alessio Bellone, Sergei I. Fokin, Vittorio Boscaro, and Claudia Vannini. Detecting Associations Between Ciliated Protists and Prokaryotes with Culture-Independent Single-Cell Microbiomics: a Proof-of-Concept Study. *Microbial Ecology*, 78(1):232–242, July 2019. ISSN 1432-184X. doi: 10.1007/s00248-018-1279-9. URL https://doi.org/10.1007/s00248-018-1279-9.

[21] Vittorio Boscaro, Vittoria Manassero, Patrick J. Keeling, and Claudia Vannini. Single-cell microbiomics unveils distribution and patterns of microbial symbioses in the natural environment. *Microbial Ecology*, 85(1):307–316, January 2022. ISSN 1432-184X. doi: 10.1007/s00248-021-01938-x. URL http://dx.doi.org/10.1007/s00248-021-01938-x.

[22] Xiaoxin Zhang, Luping Bi, Eleni Gentekaki, Jianmin Zhao, Pingping Shen, and Qianqian Zhang. Culture-independent single-cell pacbio sequencing reveals epibiotic variovorax and nucleus associated mycoplasma in the microbiome of the marine benthic protist geleia sp. yt (ciliophora, karyorelictea). *Microorganisms*, 11(6):1500, June 2023. ISSN 2076-2607. doi: 10.3390/microorganisms11061500. URL http://dx.doi.org/10.3390/microorganisms11061500.

[23] Marit F Markussen Bjorbækmo, Andreas Evenstad, Line Lieblein Røsæg, Anders K Krabberød, and Ramiro Logares. The planktonic protist interactome: where do we stand after a century of research? *The ISME Journal*, 14(2):544–559, November 2019. ISSN 1751-7370. doi: 10.1038/s41396-019-0542-5. URL http://dx.doi.org/10.1038/s41396-019-0542-5.

[24] Ernst Mayr et al. The biological species concept. In *Species Concepts and Phylogenetic Theory* Wheeler and Meier [29], pages 17–29. ISBN 9780231101424. URL https://books.google.com/books/about/Species_Concepts_and_Phylogenetic_Theory.html?hl=&id=1k4LCU2q6FMC.

[25] M.F. Claridge, H. A. Dawah, and M.R. (ed.) Wilson. Practical approaches to species concepts for living organisms. In *Species: The Units of Biodiversity* Claridge et al. [26], pages 1–15. ISBN 9780412631207.

[26] M.F. Claridge, H. A. Dawah, and M.R. (ed.) Wilson. Chapman & Hall, London, UK, mar 31 1997. ISBN 9780412631207.

[27] Leigh Van Valen. Ecological species, multispecies, and oaks. *TAXON*, 25(2–3):233–239, May 1976. ISSN 1996-8175. doi: 10.2307/1219444. URL http://dx.doi.org/10.2307/1219444.

[28] Quentin D Wheeler. The phylogenetic species concept (sensu wheeler and platnick). In *Species Concepts and Phylogenetic Theory* Wheeler and Meier [29], pages 55–69. ISBN 9780231101424. URL https://books.google.com/books/about/Species_Concepts_and_Phylogenetic_Theory.html?hl=&id=1k4LCU2q6FMC.

[29] Quentin D. Wheeler and Rudolf Meier. Columbia University Press, 2000. ISBN 9780231101424. URL https://books.google.com/books/about/Species_Concepts_and_Phylogenetic_Theory.html?hl=&id=1k4LCU2q6FMC.

[30] D.L. Hull. The ideal species concept–and why we can't get it. In *Species: The Units of Biodiversity* Claridge et al. [26], pages 357–380. ISBN 9780412631207.

[31] R.L. Mayden. A hierarchy of species concepts: the denouement in the saga of the species problem. In *Species: The Units of Biodiversity* Claridge et al. [26], pages 381–324. ISBN 9780412631207.

[32] Ramon Rosselló-Mora and Peter Kämpfer. *Defining Microbial Diversity-the Species Concept for Prokaryotic and Eukaryotic Microorganisms*, page 29–39. ASM Press, April 2014. ISBN 9781555812676. doi: 10.1128/9781555817770.ch3. URL http://dx.doi.org/10.1128/9781555817770.ch3.

[33] Brian P. Hedlund, Maria Chuvochina, Philip Hugenholtz, Konstantinos T. Konstantinidis, Alison E. Murray, Marike Palmer, Donovan H. Parks, Alexander J. Probst, Anna-Louise Reysenbach, Luis M. Rodriguez-R, Ramon Rossello-Mora, Iain C. Sutcliffe, Stephanus N. Venter, and William B. Whitman. Seqcode: a nomenclatural code for prokaryotes described from sequence data. *Nature Microbiology*, September 2022. ISSN 2058-5276. doi: 10.1038/s41564-022-01214-9. URL http://dx.doi.org/10.1038/s41564-022-01214-9.

[34] Daiqing Liao. Gene conversion drives within genic sequences: Concerted evolution of ribosomal rna genes in bacteria and archaea. *Journal of Molecular Evolution*, 51(4):305–317, October 2000. ISSN 0022-2844. doi: 10.1007/s002390010093. URL http://dx.doi.org/10.1007/s002390010093.

[35] George E. Fox, Carl R. Woese, and Kenneth R. Pechman. Comparative cataloging of 16s ribosomal ribonucleic acid: Molecular approach to procaryotic systematics. *International Journal of Systematic and Evolutionary Microbiology*, 27(1):44–57, January 1977. ISSN 1466-5034. doi: 10.1099/00207713-27-1-44. URL http://dx.doi.org/10.1099/00207713-27-1-44.

[36] Carl R. Woese. Interpreting the universal phylogenetic tree. *Proceedings of the National Academy of Sciences*, 97(15):8392–8396, July 2000. ISSN 1091-6490. doi: 10.1073/pnas.97.15.8392. URL http://dx.doi.org/10.1073/pnas.97.15.8392.

[37] Ravi Jain, Maria C. Rivera, and James A. Lake. Horizontal gene transfer among genomes: The complexity hypothesis. *Proceedings of the National Academy of Sciences*, 96(7):3801–3806, March 1999. ISSN 1091-6490. doi: 10.1073/pnas.96.7.3801. URL http://dx.doi.org/10.1073/pnas.96.7.3801.

[38] Enrico Coen, Tom Strachan, and Gabriel Dover. Dynamics of concerted evolution of ribosomal dna and histone gene families in the melanogaster species subgroup of drosophila. *Journal of*

Molecular Biology, 158(1):17–35, June 1982. ISSN 0022-2836. doi: 10.1016/0022-2836(82)90448-x. URL http://dx.doi.org/10.1016/0022-2836(82)90448-X.

[39] G J Olsen and C R Woese. Ribosomal rna: a key to phylogeny. *The FASEB Journal*, 7(1):113–123, January 1993. ISSN 1530-6860. doi: 10.1096/fasebj.7.1.8422957. URL http://dx.doi.org/10.1096/fasebj.7.1.8422957.

[40] Pablo Yarza, Pelin Yilmaz, Elmar Pruesse, Frank Oliver Glöckner, Wolfgang Ludwig, Karl-Heinz Schleifer, William B. Whitman, Jean Euzéby, Rudolf Amann, and Ramon Rosselló-Móra. Uniting the classification of cultured and uncultured bacteria and archaea using 16s rrna gene sequences. *Nature Reviews Microbiology*, 12(9):635–645, August 2014. ISSN 1740-1534. doi: 10.1038/nrmicro3330. URL http://dx.doi.org/10.1038/nrmicro3330.

[41] Carl R. Woese and George E. Fox. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences*, 74(11):5088–5090, November 1977. ISSN 1091-6490. doi: 10.1073/pnas.74.11.5088. URL http://dx.doi.org/10.1073/pnas.74.11.5088.

[42] Philip Hugenholtz, Maria Chuvochina, Aharon Oren, Donovan H Parks, and Rochelle M Soo. Prokaryotic taxonomy and nomenclature in the age of big sequence data. *The ISME Journal*, 15(7):1879–1892, April 2021. ISSN 1751-7370. doi: 10.1038/s41396-021-00941-x. URL http://dx.doi.org/10.1038/s41396-021-00941-x.

[43] David M. Ward, Roland Weller, and Mary M. Bateson. 16s rrna sequences reveal numerous uncultured microorganisms in a natural community. *Nature*, 345(6270):63–65, May 1990. ISSN 1476-4687. doi: 10.1038/345063a0. URL http://dx.doi.org/10.1038/345063a0.

[44] Stephen J. Giovannoni, Theresa B. Britschgi, Craig L. Moyer, and Katharine G. Field. Genetic diversity in sargasso sea bacterioplankton. *Nature*, 345(6270):60–63, May 1990. ISSN 1476-4687. doi: 10.1038/345060a0. URL http://dx.doi.org/10.1038/345060a0.

[45] Thomas S. B. Schmidt, João F. Matias Rodrigues, and Christian von Mering. Ecological consistency of ssu rrna-based operational taxonomic units at a global scale. *PLoS Computational Biology*, 10(4):e1003594, April 2014. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003594. URL http://dx.doi.org/10.1371/journal.pcbi.1003594.

[46] Danny S. Tuckwell, Matthew J. Nicholson, Christopher S. McSweeney, Michael K. Theodorou, and Jayne L. Brookman. The rapid assignment of ruminal fungi to presumptive genera using its1 and its2 rna secondary structures to produce group-specific fingerprints. *Microbiology*, 151(5):1557–1567, May 2005. ISSN 1465-2080. doi: 10.1099/mic.0.27689-0. URL http://dx.doi.org/10.1099/mic.0.27689-0.

[47] Christian Quast, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, 41(D1):D590–D596, January 2013. ISSN 0305-1048. doi: 10.1093/nar/gks1219. URL https://doi.org/10.1093/nar/gks1219.

[48] Donovan H Parks, Maria Chuvochina, Christian Rinke, Aaron J Mussig, Pierre-Alain Chaumeil, and Philip Hugenholtz. Gtdb: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Research*, 50(D1):D785–D794, September 2021. ISSN 1362-4962. doi: 10.1093/nar/gkab776. URL http://dx.doi.org/10.1093/nar/gkab776.

[49] James R. Cole, Qiong Wang, Jordan A. Fish, Benli Chai, Donna M. McGarrell, Yanni Sun, C. Titus Brown, Andrea Porras-Alfaro, Cheryl R. Kuske, and James M. Tiedje. Ribosomal database project: data and tools for high throughput rrna analysis. *Nucleic Acids Research*, 42(D1):D633–D642, November 2013. ISSN 1362-4962. doi: 10.1093/nar/gkt1244. URL http://dx.doi.org/10.1093/nar/gkt1244.

[50] Daniel McDonald, Yueyu Jiang, Metin Balaban, Kalen Cantrell, Qiyun Zhu, Antonio Gonzalez, James T. Morton, Giorgia Nicolaou, Donovan H. Parks, Søren M. Karst, Mads Albertsen, Philip Hugenholtz, Todd DeSantis, Se Jin Song, Andrew Bartko, Aki S. Havulinna, Pekka Jousilahti, Susan Cheng, Michael Inouye, Teemu Niiranen, Mohit Jain, Veikko Salomaa, Leo Lahti, Siavash

Mirarab, and Rob Knight. Greengenes2 unifies microbial data in a single reference tree. *Nature Biotechnology*, 42(5):715–718, July 2023. ISSN 1546-1696. doi: 10.1038/s41587-023-01845-1. URL http://dx.doi.org/10.1038/s41587-023-01845-1.

[51] Laure Guillou, Dipankar Bachar, Stéphane Audic, David Bass, Cédric Berney, Lucie Bittner, Christophe Boutte, Gaétan Burgaud, Colomban De Vargas, Johan Decelle, Javier Del Campo, John R. Dolan, Micah Dunthorn, Bente Edvardsen, Maria Holzmann, Wiebe H.C.F. Kooistra, Enrique Lara, Noan Le Bescot, Ramiro Logares, Frédéric Mahé, Ramon Massana, Marina Montresor, Raphael Morard, Fabrice Not, Jan Pawlowski, Ian Probert, Anne-Laure Sauvadet, Raffaele Siano, Thorsten Stoeck, Daniel Vaulot, Pascal Zimmermann, and Richard Christen. The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Research*, 41(D1):D597–D604, November 2012. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gks1160. URL http://academic.oup.com/nar/article/41/D1/D597/1064851/The-Protist-Ribosomal-Reference-database-PR2-a.

[52] Kessy Abarenkov, R Henrik Nilsson, Karl-Henrik Larsson, Andy F S Taylor, Tom W May, Tobias Guldberg Frøslev, Julia Pawlowska, Björn Lindahl, Kadri Põldmaa, Camille Truong, Duong Vu, Tsuyoshi Hosoya, Tuula Niskanen, Timo Piirmann, Filipp Ivanov, Allan Zirk, Marko Peterson, Tanya E Cheeke, Yui Ishigami, Arnold Tobias Jansson, Thomas Stjernegaard Jeppesen, Erik Kristiansson, Vladimir Mikryukov, Joseph T Miller, Ryoko Oono, Francisco J Ossandon, Joana Paupério, Irja Saar, Dmitry Schigel, Ave Suija, Leho Tedersoo, and Urmas Kõljalg. The unite database for molecular identification and taxonomic communication of fungi and other eukaryotes: sequences, taxa and classifications reconsidered. *Nucleic Acids Research*, 52(D1): D791–D797, November 2023. ISSN 1362-4962. doi: 10.1093/nar/gkad1039. URL http://dx.doi.org/10.1093/nar/gkad1039.

[53] coidb. URL https://github.com/biodiversitydata-se/coidb/wiki/Usage.

[54] Johan Decelle, Sarah Romac, Rowena F. Stern, El Mahdi Bendif, Adriana Zingone, Stéphane Audic, Michael D. Guiry, Laure Guillou, Désiré Tessier, Florence Le Gall, Priscillia Gourvil, Adriana L. Dos Santos, Ian Probert, Daniel Vaulot, Colomban de Vargas, and Richard Christen. Phyto¡scp¿ref¡/scp¿: a reference database of the plastidial 16s ¡scp¿rrna¡/scp¿ gene of photosynthetic eukaryotes with curated taxonomy. *Molecular Ecology Resources*, 15(6):1435–1445, April 2015. ISSN 1755-0998. doi: 10.1111/1755-0998.12401. URL http://dx.doi.org/10.1111/1755-0998.12401.

[55] Leho Tedersoo, Mahdieh S Hosseyni Moghaddam, Vladimir Mikryukov, Ali Hakimzadeh, Mohammad Bahram, R Henrik Nilsson, Iryna Yatsiuk, Stefan Geisen, Arne Schwelm, Kasia Piwosz, Marko Prous, Sirje Sildever, Dominika Chmolowska, Sonja Rueckert, Pavel Skaloud, Peeter Laas, Marco Tines, Jae-Ho Jung, Ji Hye Choi, Saad Alkahtani, and Sten Anslan. EUKARYOME: the rRNA gene reference database for identification of all eukaryotes. *Database*, 2024:baae043, February 2024. ISSN 1758-0463. doi: 10.1093/database/baae043. URL https://doi.org/10.1093/database/baae043.

[56] Mark Blaxter, Jenna Mann, Tom Chapman, Fran Thomas, Claire Whitton, Robin Floyd, and Eyualem Abebe. Defining operational taxonomic units using dna barcode data. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1462):1935–1943, September 2005. ISSN 1471-2970. doi: 10.1098/rstb.2005.1725. URL http://dx.doi.org/10.1098/rstb.2005.1725.

[57] Victor Kunin, Anna Engelbrektson, Howard Ochman, and Philip Hugenholtz. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environmental Microbiology*, 12(1):118–123, December 2009. ISSN 1462-2920. doi: 10.1111/j.1462-2920.2009.02051.x. URL http://dx.doi.org/10.1111/j.1462-2920.2009.02051.x.

[58] Benjamin J. Callahan, Paul J. McMurdie, Michael J. Rosen, Andrew W. Han, Amy Jo A. Johnson, and Susan P. Holmes. Dada2: High-resolution sample inference from illumina amplicon data. *Nature Methods*, 13(7):581 – 583, 2016. doi: 10.1038/nmeth.3869.

[59] Amnon Amir, Daniel McDonald, Jose A. Navas-Molina, Evguenia Kopylova, James T. Morton, Zhenjiang Zech Xu, Eric P. Kightley, Luke R. Thompson, Embriette R. Hyde, Antonio Gonzalez, and Rob Knight. Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems*, 2(2), April 2017. ISSN 2379-5077. doi: 10.1128/msystems.00191-16. URL http://dx.doi.org/10.1128/msystems.00191-16.

[60] R.C. Edgar. Unoise2: improved error-correction for illumina 16s and its amplicon sequencing. *BioRxiv*, 2016.

[61] Benjamin J Callahan, Joan Wong, Cheryl Heiner, Steve Oh, Casey M Theriot, Ajay S Gulati, Sarah K McGill, and Michael K Dougherty. High-throughput amplicon sequencing of the full-length 16s rrna gene with single-nucleotide resolution. *Nucleic Acids Research*, 47 (18):e103–e103, July 2019. ISSN 1362-4962. doi: 10.1093/nar/gkz569. URL http://dx.doi.org/10.1093/nar/gkz569.

[62] Benjamin J Callahan, Paul J McMurdie, and Susan P Holmes. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal*, 11(12): 2639–2643, July 2017. ISSN 1751-7370. doi: 10.1038/ismej.2017.119. URL http://dx.doi.org/10.1038/ismej.2017.119.

[63] Alejandro P. Rooney and Todd J. Ward. Evolution of a large ribosomal rna multigene family in filamentous fungi: Birth and death of a concerted evolution paradigm. *Proceedings of the National Academy of Sciences*, 102(14):5084–5089, March 2005. ISSN 1091-6490. doi: 10.1073/pnas.0409689102. URL http://dx.doi.org/10.1073/pnas.0409689102.

[64] Tomoyuki Iwanaga, Kazushi Anzawa, and Takashi Mochizuki. Variations in ribosomal dna copy numbers in a genome of trichophyton interdigitale. *Mycoses*, 63(10):1133–1140, September 2020. ISSN 1439-0507. doi: 10.1111/myc.13163. URL http://dx.doi.org/10.1111/myc.13163.

[65] A. P. Rooney. Mechanisms underlying the evolution and maintenance of functionally heterogeneous 18s rrna genes in apicomplexans. *Molecular Biology and Evolution*, 21(9):1704–1711, May 2004. ISSN 1537-1719. doi: 10.1093/molbev/msh178. URL http://dx.doi.org/10.1093/molbev/msh178.

[66] Sirinthorn Sunthornthummas, Rujipat Wasitthankasem, Pimonwan Phokhaphan, Nirinya Sudtachat, Alisa Wilantho, Chumpol Ngamphiw, Wanwisa Chareanchim, and Sissades Tongsima. Unveiling the impact of 16s rrna gene intergenomic variation on primer design and gut microbiome profiling. *Frontiers in Microbiology*, 16, May 2025. ISSN 1664-302X. doi: 10.3389/fmicb.2025.1573920. URL http://dx.doi.org/10.3389/fmicb.2025.1573920.

[67] Marcel Martinez-Porchas, Enrique Villalpando-Canchola, Luis Enrique Ortiz Suarez, and Francisco Vargas-Albores. How conserved are the conserved 16s-rrna regions? *PeerJ*, 5:e3036, February 2017. ISSN 2167-8359. doi: 10.7717/peerj.3036. URL http://dx.doi.org/10.7717/peerj.3036.

[68] Martin A. Fischer, Simon Güllert, Sven C. Neulinger, Wolfgang R. Streit, and Ruth A. Schmitz. Evaluation of 16s rrna gene primer pairs for monitoring microbial community structures showed high reproducibility within and low comparability between datasets generated with multiple archaeal and bacterial primer pairs. *Frontiers in Microbiology*, 7, August 2016. ISSN 1664-302X. doi: 10.3389/fmicb.2016.01297. URL http://dx.doi.org/10.3389/fmicb.2016.01297.

[69] Benjamin Hillmann, Gabriel A. Al-Ghalith, Robin R. Shields-Cutler, Qiyun Zhu, Daryl M. Gohl, Kenneth B. Beckman, Rob Knight, and Dan Knights. Evaluating the information content of shallow shotgun metagenomics. *mSystems*, 3(6), October 2018. ISSN 2379-5077. doi: 10.1128/msystems.00069-18. URL http://dx.doi.org/10.1128/msystems.00069-18.

[70] Oldřich Bartoš, Martin Chmel, and Iva Swierczková. The overlooked evolutionary dynamics of 16s rrna revises its role as the "gold standard" for bacterial species identification. *Scientific Reports*, 14(1), April 2024. ISSN 2045-2322. doi: 10.1038/s41598-024-59667-3. URL http://dx.doi.org/10.1038/s41598-024-59667-3.

[71] Bo Yang, Yong Wang, and Pei-Yuan Qian. Sensitivity and correlation of hypervariable regions in 16s rrna genes in phylogenetic analysis. *BMC Bioinformatics*, 17(1), March 2016. ISSN 1471-2105. doi: 10.1186/s12859-016-0992-y. URL http://dx.doi.org/10.1186/s12859-016-0992-y.

[72] Shahed U. A. Shazib, Ragib Ahsan, Marie Leleu, George B. McManus, Laura A. Katz, and Luciana F. Santoferrara. Phylogenomic workflow for uncultivable microbial eukaryotes using single-cell RNA sequencing - a case study with planktonic ciliates (ciliophora, oligotrichea). *Molecular Phylogenetics and Evolution*, 204:108239, March 2025. ISSN 1055-7903. doi: 10.1016/j.ympev.2024.108239. URL https://www.sciencedirect.com/science/article/pii/S1055790324002318.

[73] R. H. Whittaker. Evolution and measurement of species diversity. *TAXON*, 21(2–3):213–251, May 1972. ISSN 1996-8175. doi: 10.2307/1218190. URL http://dx.doi.org/10.2307/1218190.

[74] Camilo Mora, Derek P. Tittensor, Sina Adl, Alastair G. B. Simpson, and Boris Worm. How many species are there on earth and in the ocean? *PLoS Biology*, 9(8):e1001127, August 2011. ISSN 1545-7885. doi: 10.1371/journal.pbio.1001127. URL http://dx.doi.org/10.1371/journal.pbio.1001127.

[75] Kenneth J. Locey and Jay T. Lennon. Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences*, 113(21):5970–5975, May 2016. ISSN 1091-6490. doi: 10.1073/pnas.1521291113. URL http://dx.doi.org/10.1073/pnas.1521291113.

[76] John J. Wiens. How many species are there on earth? progress and problems. *PLOS Biology*, 21(11):e3002388, November 2023. ISSN 1545-7885. doi: 10.1371/journal.pbio.3002388. URL http://dx.doi.org/10.1371/journal.pbio.3002388.

[77] Yinon M. Bar-On, Rob Phillips, and Ron Milo. The biomass distribution on Earth. *Proceedings of the National Academy of Sciences*, 115(25):6506–6511, June 2018. doi: 10.1073/pnas.1711842115. URL https://www.pnas.org/doi/10.1073/pnas.1711842115. Publisher: Proceedings of the National Academy of Sciences.

[78] Camille Parmesan. Ecological and evolutionary responses to recent climate change. *Annual Review of Ecology, Evolution, and Systematics*, 37(1):637–669, December 2006. ISSN 1545-2069. doi: 10.1146/annurev.ecolsys.37.091305.110100. URL http://dx.doi.org/10.1146/annurev.ecolsys.37.091305.110100.

[79] Sean L. Maxwell, Richard A. Fuller, Thomas M. Brooks, and James E. M. Watson. Biodiversity: The ravages of guns, nets and bulldozers. *Nature*, 536(7615):143–145, August 2016. ISSN 1476-4687. doi: 10.1038/536143a. URL http://dx.doi.org/10.1038/536143a.

[80] Michael S. Rappé and Stephen J. Giovannoni. The uncultured microbial majority. *Annual Review of Microbiology*, 57(1):369–394, October 2003. ISSN 1545-3251. doi: 10.1146/annurev.micro.57.030502.090759. URL http://dx.doi.org/10.1146/annurev.micro.57.030502.090759.

[81] R I Amann, W Ludwig, and K H Schleifer. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiological Reviews*, 59(1):143–169, March 1995. ISSN 0146-0749. doi: 10.1128/mr.59.1.143-169.1995. URL http://dx.doi.org/10.1128/mr.59.1.143-169.1995.

[82] Lin-Xing Chen, Karthik Anantharaman, Alon Shaiber, A. Murat Eren, and Jillian F. Banfield. Accurate and complete genomes from metagenomes. *Genome Research*, 30(3):315–333, March 2020. ISSN 1549-5469. doi: 10.1101/gr.258640.119. URL http://dx.doi.org/10.1101/gr.258640.119.

[83] Owen L. Petchey and Kevin J. Gaston. Functional diversity (fd), species richness and community composition. *Ecology Letters*, 5(3):402–411, May 2002. ISSN 1461-0248. doi: 10.1046/j.1461-0248.2002.00339.x. URL http://dx.doi.org/10.1046/j.1461-0248.2002.00339.x.

[84] P. Calow. Towards a definition of functional ecology. *Functional Ecology*, 1(1):57, 1987. ISSN 0269-8463. doi: 10.2307/2389358. URL http://dx.doi.org/10.2307/2389358.

[85] Jan Bengtsson. Which species? what kind of diversity? which ecosystem function? some problems in studies of relations between biodiversity and ecosystem function. *Applied Soil Ecology*, 10(3):191–199, November 1998. ISSN 0929-1393. doi: 10.1016/s0929-1393(98)00120-6. URL http://dx.doi.org/10.1016/S0929-1393(98)00120-6.

[86] Anne E. Bernhard and John J. Kelly. Editorial: Linking ecosystem function to microbial diversity. *Frontiers in Microbiology*, 7, June 2016. ISSN 1664-302X. doi: 10.3389/fmicb.2016.01041. URL http://dx.doi.org/10.3389/fmicb.2016.01041.

[87] Diane K. Stoecker, Per Juel Hansen, David A. Caron, and Aditee Mitra. Mixotrophy in the marine plankton. *Annual Review of Marine Science*, 9(1):311–335, January 2017. ISSN 1941-0611. doi: 10.1146/annurev-marine-010816-060617. URL http://dx.doi.org/10.1146/annurev-marine-010816-060617.

[88] Engràcia Costa, Julio Pérez, and Jan-Ulrich Kreft. Why is metabolic labour divided in nitrification? *Trends in Microbiology*, 14(5):213–219, May 2006. ISSN 0966-842X. doi: 10.1016/j.tim.2006.03.006. URL http://dx.doi.org/10.1016/j.tim.2006.03.006.

[89] Brandon E.L. Morris, Ruth Henneberger, Harald Huber, and Christine Moissl-Eichinger. Microbial syntrophy: interaction for the common good. *FEMS Microbiology Reviews*, 37(3):384–406, May 2013. ISSN 1574-6976. doi: 10.1111/1574-6976.12019. URL http://dx.doi.org/10.1111/1574-6976.12019.

[90] Annette B.G. Janssen, Sven Teurlincx, Shuqing An, Jan H. Janse, Hans W. Paerl, and Wolf M. Mooij. Alternative stable states in large shallow lakes? *Journal of Great Lakes Research*, 40(4):813 – 826, 2014. doi: 10.1016/j.jglr.2014.09.019.

[91] Anita Narwani, Marta Reyes, Aaron Louis Pereira, Hannele Penson, Stuart R. Dennis, Samuel Derrer, Piet Spaak, and Blake Matthews. Interactive effects of foundation species on ecosystem functioning and stability in response to disturbance. *Proceedings of the Royal Society B: Biological Sciences*, 286(1913), 2019. doi: 10.1098/rspb.2019.1857.

[92] Barbara J. Campbell, Liying Yu, John F. Heidelberg, and David L. Kirchman. Activity of abundant and rare bacteria in a coastal ocean. *Proceedings of the National Academy of Sciences*, 108(31):12776–12781, July 2011. ISSN 1091-6490. doi: 10.1073/pnas.1101405108. URL http://dx.doi.org/10.1073/pnas.1101405108.

[93] DA Caron and PD Countway. Hypotheses on the role of the protistan rare biosphere in a changing world. *Aquatic Microbial Ecology*, 57:227–238, November 2009. ISSN 1616-1564. doi: 10.3354/ame01352. URL http://dx.doi.org/10.3354/ame01352.

[94] Mitchell L. Sogin, Hilary G. Morrison, Julie A. Huber, David Mark Welch, Susan M. Huse, Phillip R. Neal, Jesus M. Arrieta, and Gerhard J. Herndl. Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences*, 103(32):12115–12120, August 2006. ISSN 1091-6490. doi: 10.1073/pnas.0605127103. URL http://dx.doi.org/10.1073/pnas.0605127103.

[95] Aaron M. Ellison. Foundation species, non-trophic interactions, and the value of being common. *iScience*, 13:254–268, March 2019. ISSN 2589-0042. doi: 10.1016/j.isci.2019.02.020. URL http://dx.doi.org/10.1016/j.isci.2019.02.020.

[96] Aditya Jeevannavar, Anita Narwani, Blake Matthews, Piet Spaak, Jeanine Brantschen, Elvira Mächler, Florian Altermatt, and Manu Tamminen. Foundation species stabilize an alternative eutrophic state in nutrient-disturbed ponds via selection on microbial community. *Frontiers in Microbiology*, 15, April 2024. ISSN 1664-302X. doi: 10.3389/fmicb.2024.1310374. URL https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2024.1310374/full. Publisher: Frontiers.

[97] S.M. Thomaz and E.R. Cunha. The role of macrophytes in habitat structuring in aquatic ecosystems: methods of measurement, causes and consequences on animal assemblages' composition and biodiversity. *Acta Limnologica Brasiliensia*, 22(2):218 – 236, 2010.

[98] S. Declerck, M. Vanderstukken, A. Pals, K. Muylaert, and L. De Meester. Plankton biodiversity along a gradient of productivity and its mediation by macrophytes. *Ecology*, 88(9):2199 – 2210, 2007. doi: 10.1890/07-0048.1.

[99] Richard B. Hallick, Ling Hong, Robert G. Drager, Mitchell R. Favreau, Amparo Monfort, Bernard Orsat, Albert Spielmann, and Erhard Stutz. Complete sequence of *Euglena gracilis* chloroplast dna. *Nucleic Acids Research*, 21(15):3537–3544, 1993. ISSN 1362-4962. doi: 10.1093/nar/21.15.3537. URL http://dx.doi.org/10.1093/nar/21.15.3537.

[100] Francis A. Tablizo and Arturo O. Lluisma. The mitochondrial genome of the red alga kappaphycus striatus ("green sacol" variety): Complete nucleotide sequence, genome structure and organization, and comparative analysis. *Marine Genomics*, 18:155–161, December 2014. ISSN 1874-7787. doi: 10.1016/j.margen.2014.05.006. URL http://dx.doi.org/10.1016/j.margen.2014.05.006.

[101] Isao Masuda, Motomichi Matsuzaki, and Kiyoshi Kita. Extensive frameshift at all agg and ccc codons in the mitochondrial cytochrome c oxidase subunit 1 gene of perkinsus marinus (alveolata; dinoflagellata). *Nucleic Acids Research*, 38(18):6186–6194, May 2010. ISSN 1362-4962. doi: 10.1093/nar/gkq449. URL http://dx.doi.org/10.1093/nar/gkq449.

[102] David Roy Smith and Patrick J. Keeling. Protists and the wild, wild west of gene expression: New frontiers, lawlessness, and misfits. *Annual Review of Microbiology*, 70:161 – 178, 2016. doi: 10.1146/annurev-micro-102215-095448.

[103] Daniel Straub, Nia Blackwell, Adrian Langarica-Fuentes, Alexander Peltzer, Sven Nahnsen, and Sara Kleindienst. Interpretations of environmental microbial community studies are biased by the selected 16s rrna (gene) amplicon sequencing pipeline. *Frontiers in Microbiology*, 11, 2020. doi: 10.3389/fmicb.2020.550420.

[104] Javier Florenza, Manu Tamminen, and Stefan Bertilsson. Uncovering microbial inter-domain interactions in complex communities. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374:20190087, 10 2019. doi: 10.1098/rstb.2019.0087.

[105] Karla Iveth Aguilera-Campos, Julie Boisard, Viktor Törnblom, Jon Jerlström-Hultqvist, Ada Behncké-Serra, Elena Aramendia Cotillas, and Courtney Weir Stairs. Anaerobic breviate protist survival in microcosms depends on microbiome metabolic function. *The ISME Journal*, August 2025. ISSN 1751-7370. doi: 10.1093/ismejo/wraf171. URL http://dx.doi.org/10.1093/ismejo/wraf171.

[106] Alessia Rossi, Alessio Bellone, Sergei I. Fokin, Vittorio Boscaro, and Claudia Vannini. Detecting associations between ciliated protists and prokaryotes with culture-independent single-cell microbiomics: a proof-of-concept study. *Microbial Ecology*, 78(1):232–242, November 2018. ISSN 1432-184X. doi: 10.1007/s00248-018-1279-9. URL http://dx.doi.org/10.1007/s00248-018-1279-9.

[107] Ruben Schulte-Hillen, Jakob K. Giesler, Thomas Mock, Nigel Belshaw, Uwe John, Tilmann Harder, Nancy Kühne, Stefan Neuhaus, and Sylke Wohlrab. A crispr-cas9 assisted analysis of single-cell microbiomes for identifying rare bacterial taxa in phycospheres of diatoms. *bioRxiv*, 2024. doi: 10.1101/2024.11.10.622248. URL https://www.biorxiv.org/content/early/2024/11/10/2024.11.10.622248.

[108] Sarah J Spencer, Manu V Tamminen, Sarah P Preheim, Mira T Guo, Adrian W Briggs, Ilana L Brito, David A Weitz, Leena K Pitkänen, Francois Vigneault, Marko P Juhani Virta, and Eric J Alm. Massively parallel sequencing of single cells by epicpcr links functional genes with phylogenetic markers. *The ISME Journal*, 10(2):427–436, September 2015. ISSN 1751-7370. doi: 10.1038/ismej.2015.124. URL http://dx.doi.org/10.1038/ismej.2015.124.

[109] Eric G. Sakowski, Keith Arora-Williams, Funing Tian, Ahmed A. Zayed, Olivier Zablocki, Matthew B. Sullivan, and Sarah P. Preheim. Interaction dynamics and virus–host range for estuarine actinophages captured by epicpcr. *Nature Microbiology*, 6(5):630–642, February 2021. ISSN 2058-5276. doi: 10.1038/s41564-021-00873-4. URL http://dx.doi.org/10.1038/s41564-021-00873-4.

[110] Bingliang Xie, Jian Wang, Yong Nie, Jing Tian, Zerui Wang, Dongwei Chen, Beiyu Hu, Xiao-Lei Wu, and Wenbin Du. Type iv pili trigger episymbiotic association of saccharibacteria with its bacterial host. *Proceedings of the National Academy of Sciences*, 119(49), December 2022.

ISSN 1091-6490. doi: 10.1073/pnas.2215990119. URL http://dx.doi.org/10.1073/pnas.2215990119.

[111] Sebastian C. Treitli, Martin Kolisko, Filip Husník, Patrick J. Keeling, and Vladimír Hampl. Revealing the metabolic capacity of streblomastix strix and its bacterial symbionts using single-cell metagenomics. *Proceedings of the National Academy of Sciences*, 116(39):19675–19684, September 2019. ISSN 1091-6490. doi: 10.1073/pnas.1910793116. URL http://dx.doi.org/10.1073/pnas.1910793116.

[112] Bing Zhang, Liwen Xiao, Liping Lyu, Fangqing Zhao, and Miao Miao. Exploring the landscape of symbiotic diversity and distribution in unicellular ciliated protists. *Microbiome*, 12(1), May 2024. ISSN 2049-2618. doi: 10.1186/s40168-024-01809-w. URL http://dx.doi.org/10.1186/s40168-024-01809-w.

[113] Frederik Schulz, Ying Yan, Agnes K.M. Weiner, Ragib Ahsan, Laura A. Katz, and Tanja Woyke. Protists as mediators of complex microbial and viral associations. *bioRxiv*, 2024. doi: 10.1101/2024.12.29.630703. URL https://www.biorxiv.org/content/early/2024/12/30/2024.12.29.630703.

[114] Martin Kolisko, Vittorio Boscaro, Fabien Burki, Denis H. Lynn, and Patrick J. Keeling. Single-cell transcriptomics for microbial eukaryotes. *Current Biology*, 24(22):R1081–R1082, November 2014. ISSN 0960-9822. doi: 10.1016/j.cub.2014.10.026. URL https://www.sciencedirect.com/science/article/pii/S0960982214013207.

[115] Yaohan Jiang, Xiao Chen, Chundi Wang, Liping Lyu, Saleh A. Al-Farraj, Naomi A. Stover, and Feng Gao. Genes and proteins expressed at different life cycle stages in the model protist Euplotes vannus revealed by both transcriptomic and proteomic approaches. *Science China Life Sciences*, 68(1):232–248, January 2025. ISSN 1869-1889. doi: 10.1007/s11427-023-2605-9. URL https://doi.org/10.1007/s11427-023-2605-9.

[116] Adam J Reid, Arthur M Talman, Hayley M Bennett, Ana R Gomes, Mandy J Sanders, Christopher J R Illingworth, Oliver Billker, Matthew Berriman, and Mara KN Lawniczak. Single-cell RNA-seq reveals hidden transcriptional variation in malaria parasites. *eLife*, 7:e33105, March 2018. ISSN 2050-084X. doi: 10.7554/eLife.33105. URL https://doi.org/10.7554/eLife.33105. Publisher: eLife Sciences Publications, Ltd.

[117] Virginia M. Howick, Andrew J. C. Russell, Tallulah Andrews, Haynes Heaton, Adam J. Reid, Kedar Natarajan, Hellen Butungi, Tom Metcalf, Lisa H. Verzier, Julian C. Rayner, Matthew Berriman, Jeremy K. Herren, Oliver Billker, Martin Hemberg, Arthur M. Talman, and Mara K. N. Lawniczak. The Malaria Cell Atlas: Single parasite transcriptomes across the complete Plasmodium life cycle. *Science*, 365(6455):eaaw2619, August 2019. doi: 10.1126/science.aaw2619. URL https://www.science.org/doi/full/10.1126/science.aaw2619. Publisher: American Association for the Advancement of Science.

[118] Mariona Nadal-Ribelles, Saiful Islam, Wu Wei, Pablo Latorre, Michelle Nguyen, Eulàlia de Nadal, Francesc Posas, and Lars M. Steinmetz. Sensitive high-throughput single-cell RNA-seq reveals within-clonal transcript correlations in yeast populations. *Nature Microbiology*, 4 (4):683–692, April 2019. ISSN 2058-5276. doi: 10.1038/s41564-018-0346-9. URL https://www.nature.com/articles/s41564-018-0346-9. Publisher: Nature Publishing Group.

[119] Malika Saint, François Bertaux, Wenhao Tang, Xi-Ming Sun, Laurence Game, Anna Köferle, Jürg Bähler, Vahid Shahrezaei, and Samuel Marguerat. Single-cell imaging and RNA sequencing reveal patterns of gene expression heterogeneity during fission yeast growth and adaptation. *Nature Microbiology*, 4(3):480–491, March 2019. ISSN 2058-5276. doi: 10.1038/s41564-018-0330-4. URL https://www.nature.com/articles/s41564-018-0330-4. Publisher: Nature Publishing Group.

[120] Anders K. Krabberød, Russell J.S. Orr, Jon Bråte, Tom Kristensen, Kjell R. Bjørklund, and Kamran Shalchian-Tabrizi. Single cell transcriptomics, mega-phylogeny, and the genetic basis of morphological innovations in rhizaria. *Molecular Biology and Evolution*, 34(7):1557–1573,

March 2017. ISSN 1537-1719. doi: 10.1093/molbev/msx075. URL http://dx.doi.org/10.1093/molbev/msx075.

[121] Yuki Nishimura, Masato Otagiri, Masahiro Yuki, Michiru Shimizu, Jun-ichi Inoue, Shigeharu Moriya, and Moriya Ohkuma. Division of functional roles for termite gut protists revealed by single-cell transcriptomes. *The ISME Journal*, 14(10):2449–2460, October 2020. ISSN 1751-7362. doi: 10.1038/s41396-020-0698-z. URL https://doi.org/10.1038/s41396-020-0698-z.

[122] Caroline Juéry, Adria Auladell, Zoltan Füssy, Fabien Chevalier, Daniel P Yee, Eric Pelletier, Erwan Corre, Andrew E Allen, Daniel J Richter, and Johan Decelle. Transportome remodeling of a symbiotic microalga inside a planktonic host. *The ISME Journal*, 18(1):wrae239, January 2024. ISSN 1751-7362. doi: 10.1093/ismejo/wrae239. URL https://doi.org/10.1093/ismejo/wrae239.

[123] Weitian Zhou, Weishan Zhao, Shiman Yang, Fanya Nie, Daji Luo, Ming Li, Wenxiang Li, Hong Zou, and Guitang Wang. Single-cell transcriptome profiles and e-64 inhibitor data reveal the essential role of cysteine proteases in the ontogeny of *Ichthyophthirius multifiliis*. *Fish & Shellfish Immunology*, 154:109979, November 2024. ISSN 1050-4648. doi: 10.1016/j.fsi.2024.109979. URL https://www.sciencedirect.com/science/article/pii/S1050464824006247.

[124] Ying Yan, Xyrus X. Maurer-Alcalá, Rob Knight, Sergei L. Kosakovsky Pond, and Laura A. Katz. Single-Cell Transcriptomics Reveal a Correlation between Genome Architecture and Gene Family Evolution in Ciliates. *mBio*, 10(6):10.1128/mbio.02524–19, December 2019. doi: 10.1128/mbio.02524-19. URL https://journals.asm.org/doi/10.1128/mbio.02524-19. Publisher: American Society for Microbiology.

[125] Henning Onsbring, Mahwash Jamy, and Thijs J. G. Ettema. RNA Sequencing of Stentor Cell Fragments Reveals Transcriptional Changes during Cellular Regeneration. *Current Biology*, 28(8):1281–1288.e3, April 2018. ISSN 0960-9822. doi: 10.1016/j.cub.2018.02.055. URL https://www.cell.com/current-biology/abstract/S0960-9822(18)30240-9. Publisher: Elsevier.

[126] Elizabeth C. Cooney, Corey C. Holt, Elisabeth Hehenberger, Jayd A. Adams, Brian S. Leander, and Patrick J. Keeling. Investigation of heterotrophs reveals new insights in dinoflagellate evolution. *Molecular Phylogenetics and Evolution*, 196:108086, July 2024. ISSN 1055-7903. doi: 10.1016/j.ympev.2024.108086. URL https://www.sciencedirect.com/science/article/pii/S1055790324000782.

[127] Henning Onsbring, Alexander K. Tice, Brandon T. Barton, Matthew W. Brown, and Thijs J. G. Ettema. An efficient single-cell transcriptomics workflow for microbial eukaryotes benchmarked on Giardia intestinalis cells. *BMC Genomics*, 21(1):1–9, December 2020. ISSN 1471-2164. doi: 10.1186/s12864-020-06858-7. URL https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-020-06858-7. Publisher: BioMed Central.

[128] Vesna Grujčić, Sami Saarenpää, John Sundh, Bengt Sennblad, Benjamin Norgren, Meike Latz, Stefania Giacomello, Rachel A. Foster, and Anders F. Andersson. Towards high-throughput parallel imaging and single-cell transcriptomics of microbial eukaryotic plankton. *PLOS ONE*, 19(1):e0296672, January 2024. ISSN 1932-6203. doi: 10.1371/journal.pone.0296672. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0296672. Publisher: Public Library of Science.

[129] Daniel Tillett and Brett A. Neilan. Xanthogenate nucleic acid isolation from cultured and environmental cyanobacteria. *Journal of Phycology*, 36(1):251 – 258, 2000. doi: 10.1046/j.1529-8817.2000.99079.x.

[130] Kristy Deiner, Jean-Claude Walser, Elvira Mächler, and Florian Altermatt. Choice of capture and extraction methods affect detection of freshwater biodiversity from environmental dna. *Biological Conservation*, 183:53 – 63, 2015. doi: 10.1016/j.biocon.2014.11.018.

[131] Maureen D. Keller, Rhonda C. Selvin, Wolfgang Claus, and Robert R. L. Guillard. Media for

the culture of oceanic ultraphytoplankton. *Journal of Phycology*, 23(4):633 – 638, 1987. doi: 10.1111/j.1529-8817.1987.tb04217.x.

[132] John A. Berges, Daniel J. Franklin, and Paul J. Harrison. Evolution of an artificial seawater medium: Improvements in enriched seawater, artificial water over the last two decades. *Journal of Phycology*, 37(6):1138 – 1145, 2001. doi: 10.1046/j.1529-8817.2001.01052.x.

[133] Anna Klindworth, Elmar Pruesse, Timmy Schweer, Jörg Peplies, Christian Quast, Matthias Horn, and Frank Oliver Glöckner. Evaluation of general 16s ribosomal rna gene pcr primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Research*, 41(1):e1, 2013. doi: 10.1093/nar/gks808.

[134] Luisa W. Hugerth, Emilie E. L. Muller, Yue O. O. Hu, Laura A. M. Lebrun, Hugo Roume, Daniel Lundin, Paul Wilmes, and Anders F. Andersson. Systematic design of 18s rrna gene primers for determining eukaryotic diversity in microbial consortia. *PLoS ONE*, 9(4), 2014. doi: 10.1371/journal.pone.0095567.

[135] J. Geller, C. Meyer, M. Parker, and H. Hawk. Redesign of pcr primers for mitochondrial cytochrome c oxidase subunit i for marine invertebrates and application in all-taxa biotic surveys. *Molecular Ecology Resources*, 13(5):851 – 861, 2013. doi: 10.1111/1755-0998.12138.

[136] Matthieu Leray, Joy Y. Yang, Christopher P. Meyer, Suzanne C. Mills, Natalia Agudelo, Vincent Ranwez, Joel T. Boehm, and Ryuji J. Machida. A new versatile primer set targeting a short fragment of the mitochondrial coi region for metabarcoding metazoan diversity: Application for characterizing coral reef fish gut contents. *Frontiers in Zoology*, 10(1), 2013. doi: 10.1186/1742-9994-10-34.

[137] Simone Picelli, Omid R. Faridani, Åsa K. Björklund, Gösta Winberg, Sven Sagasser, and Rickard Sandberg. Full-length rna-seq from single cells using smart-seq2. *Nature Protocols*, 9(1):171 – 181, 2014. doi: 10.1038/nprot.2014.006.

[138] Michael Hagemann-Jensen, Christoph Ziegenhain, and Rickard Sandberg. Scalable single-cell RNA sequencing from full transcripts with Smart-seq3xpress. *Nature Biotechnology*, 40 (10):1452–1457, October 2022. ISSN 1546-1696. doi: 10.1038/s41587-022-01311-4. URL https://www.nature.com/articles/s41587-022-01311-4. Publisher: Nature Publishing Group.

[139] Michael Hagemann-Jensen, Christoph Ziegenhain, Ping Chen, Daniel Ramsköld, Gert-Jan Hendriks, Anton J. M. Larsson, Omid R. Faridani, and Rickard Sandberg. Single-cell RNA counting at allele and isoform resolution using Smart-seq3. *Nature Biotechnology*, 38(6):708–714, June 2020. ISSN 1546-1696. doi: 10.1038/s41587-020-0497-0. URL https://www.nature.com/articles/s41587-020-0497-0. Publisher: Nature Publishing Group.

[140] Philip A. Ewels, Alexander Peltzer, Sven Fillinger, Harshil Patel, Johannes Alneberg, Andreas Wilm, Maxime Ulysse Garcia, Paolo Di Tommaso, and Sven Nahnsen. The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology*, 38(3):276 – 278, 2020. doi: 10.1038/s41587-020-0439-x.

[141] Paolo DI Tommaso, Maria Chatzou, Evan W. Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35(4):316 – 319, 2017. doi: 10.1038/nbt.3820.

[142] M. Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal*, 17(1):10 – 12, 2011.

[143] S. Andrews. Fastqc. *FastQC*, 2010.

[144] Philip Ewels, Måns Magnusson, Sverker Lundin, and Max Käller. Multiqc: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19):3047, 2016. doi: 10.1093/bioinformatics/btw354. URL +http://dx.doi.org/10.1093/bioinformatics/btw354.

[145] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10–12, May 2011. ISSN 2226-6089. doi: 10.14806/ej.17.1.200. URL https://journal.embnet.org/index.php/embnetjournal/article/view/200. Number: 1.

[146] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, April 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu170. URL http://dx.doi.org/10.1093/bioinformatics/btu170.

[147] Shifu Chen. Ultrafast one-pass fastq data preprocessing, quality control, and deduplication using fastp. *iMeta*, 2(2), May 2023. ISSN 2770-596X. doi: 10.1002/imt2.107. URL http://dx.doi.org/10.1002/imt2.107.

[148] Felix Krueger, Frankie James, Phil Ewels, Ebrahim Afyounian, Michael Weinstein, Benjamin Schuster-Boeckler, Gert Hulselmans, and sclamons. FelixKrueger/TrimGalore: v0.6.10 - add default decompression path, February 2023. URL https://zenodo.org/records/7598955.

[149] Brian Bushnell. BBMap, July 2025. URL https://sourceforge.net/projects/bbmap/.

[150] Mikkel Schubert, Stinus Lindgreen, and Ludovic Orlando. Adapterremoval v2: rapid adapter trimming, identification, and read merging. *BMC Research Notes*, 9(1), February 2016. ISSN 1756-0500. doi: 10.1186/s13104-016-1900-2. URL http://dx.doi.org/10.1186/s13104-016-1900-2.

[151] Wouter De Coster and Rosa Rademakers. Nanopack2: population-scale evaluation of long-read sequencing data. *Bioinformatics*, 39(5), May 2023. ISSN 1367-4811. doi: 10.1093/bioinformatics/btad311. URL http://dx.doi.org/10.1093/bioinformatics/btad311.

[152] Robert C. Edgar. Search and clustering orders of magnitude faster than blast. *Bioinformatics*, 26(19):2460–2461, August 2010. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq461. URL http://dx.doi.org/10.1093/bioinformatics/btq461.

[153] Robert Schmieder and Robert Edwards. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27(6):863–864, January 2011. ISSN 1367-4803. doi: 10.1093/bioinformatics/btr026. URL http://dx.doi.org/10.1093/bioinformatics/btr026.

[154] Torbjørn Rognes, Tomáš Flouri, Ben Nichols, Christopher Quince, and Frédéric Mahé. VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, 4:e2584, October 2016. ISSN 2167-8359. doi: 10.7717/peerj.2584. URL https://peerj.com/articles/2584. Publisher: PeerJ Inc.

[155] Derrick E. Wood, Jennifer Lu, and Ben Langmead. Improved metagenomic analysis with kraken 2. *Genome Biology*, 20(1), 2019. doi: 10.1186/s13059-019-1891-0.

[156] Benjamin Buchfink, Klaus Reuter, and Hajk-Georg Drost. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*, 18(4):366–368, April 2021. ISSN 1548-7105. doi: 10.1038/s41592-021-01101-x. URL https://www.nature.com/articles/s41592-021-01101-x. Publisher: Nature Publishing Group.

[157] Steven Wingett. Stevenwingett/fastq-screen: Fastq screen v0.15.1, 2022. URL https://zenodo.org/record/5838377.

[158] Zhi-Luo Deng, Philipp C Münch, René Mreches, and Alice C McHardy. Rapid and accurate identification of ribosomal rna sequences via deep learning. *Nucleic Acids Research*, 50(10):E60, 2022. doi: 10.1093/nar/gkac112.

[159] Susannah J. Salter, Michael J. Cox, Elena M. Turek, Szymon T. Calus, William O. Cookson, Miriam F. Moffatt, Paul Turner, Julian Parkhill, Nicholas J. Loman, and Alan W. Walker. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology*, 12(1), 2014. doi: 10.1186/s12915-014-0087-z.

[160] Leonid A. Kulakov, Morven B. McAlister, Kimberly L. Ogden, Michael J. Larkin, and John F. O'Hanlon. Analysis of bacteria contaminating ultrapure water in industrial systems. *Applied and Environmental Microbiology*, 68(4):1548–1555, April 2002. ISSN 1098-5336. doi: 10.1128/aem.68.4.1548-1555.2002. URL http://dx.doi.org/10.1128/AEM.68.4.1548-1555.2002.

[161] MB McAlister, LA Kulakov, JF O'Hanlon, MJ Larkin, and KL Ogden. Survival and nutritional requirements of three bacteria isolated from ultrapure water. *Journal of Industrial Microbiology and Biotechnology*, 29(2):75–82, August 2002. ISSN 1476-5535. doi: 10.1038/sj.jim.7000273. URL http://dx.doi.org/10.1038/sj.jim.7000273.

[162] Jennifer Sue Gray, Janette Marie Birmingham, and Jenifer Imig Fenton. Got black swimming dots in your cell culture? identification of achromobacter as a novel cell culture contaminant. *Biologicals*, 38(2):273–277, March 2010. ISSN 1045-1056. doi: 10.1016/j.biologicals.2009.09.006. URL http://dx.doi.org/10.1016/j.biologicals.2009.09.006.

[163] M. Maiwald, H.-J. Ditton, H.-G. Sonntag, and M. von Knebel Doeberitz. Characterization of contaminating dna in taq polymerase which occurs during amplification with a primer set for legionella 5s ribosomal rna. *Molecular and Cellular Probes*, 8(1):11–14, February 1994. ISSN 0890-8508. doi: 10.1006/mcpr.1994.1002. URL http://dx.doi.org/10.1006/mcpr.1994.1002.

[164] Niclas Grahn, Margaretha Olofsson, Katarina Ellnebo-Svedlund, Hans-JÃ¼rg Monstein, and Jon Jonasson. Identification of mixed bacterial dna contamination in broad-range pcr amplification of 16s rdna v1 and v3 variable regions by pyrosequencing of cloned amplicons. *FEMS Microbiology Letters*, 219(1):87–91, February 2003. ISSN 1574-6968. doi: 10.1016/s0378-1097(02)01190-4. URL http://dx.doi.org/10.1016/S0378-1097(02)01190-4.

[165] Tamimount Mohammadi, Henk W. Reesink, Christina M.J.E. Vandenbroucke-Grauls, and Paul H.M. Savelkoul. Removal of contaminating dna from commercial nucleic acid extraction kit reagents. *Journal of Microbiological Methods*, 61(2):285–288, May 2005. ISSN 0167-7012. doi: 10.1016/j.mimet.2004.11.018. URL http://dx.doi.org/10.1016/j.mimet.2004.11.018.

[166] Martin Laurence, Christos Hatzis, and Douglas E. Brash. Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes. *PLoS ONE*, 9(5):e97876, May 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0097876. URL http://dx.doi.org/10.1371/journal.pone.0097876.

[167] Robyn J. Wright, André M. Comeau, and Morgan G. I. Langille. From defaults to databases: parameter and database choice dramatically impact the performance of metagenomic taxonomic classification tools. *Microbial Genomics*, 9(3), March 2023. ISSN 2057-5858. doi: 10.1099/mgen.0.000949. URL http://dx.doi.org/10.1099/mgen.0.000949.

[168] Venket Raghavan, Louis Kraft, Fantin Mesny, and Linda Rigerte. A simple guide to de novo transcriptome assembly and annotation. *Briefings in Bioinformatics*, 23(2), January 2022. ISSN 1477-4054. doi: 10.1093/bib/bbab563. URL http://dx.doi.org/10.1093/bib/bbab563.

[169] Manfred G. Grabherr, Brian J. Haas, Moran Yassour, Joshua Z. Levin, Dawn A. Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, Zehua Chen, Evan Mauceli, Nir Hacohen, Andreas Gnirke, Nicholas Rhind, Federica Di Palma, Bruce W. Birren, Chad Nusbaum, Kerstin Lindblad-Toh, Nir Friedman, and Aviv Regev. Full-length transcriptome assembly from rna-seq data without a reference genome. *Nature Biotechnology*, 29(7):644 – 652, 2011. doi: 10.1038/nbt.1883.

[170] Elena Bushmanova, Dmitry Antipov, Alla Lapidus, and Andrey D Prjibelski. rnaspades: a de novo transcriptome assembler and its application to rna-seq data. *GigaScience*, 8(9), September 2019. ISSN 2047-217X. doi: 10.1093/gigascience/giz100. URL http://dx.doi.org/10.1093/gigascience/giz100.

[171] Yinlong Xie, Gengxiong Wu, Jingbo Tang, Ruibang Luo, Jordan Patterson, Shanlin Liu, Weihua Huang, Guangzhu He, Shengchang Gu, Shengkang Li, Xin Zhou, Tak-Wah Lam, Yingrui Li, Xun Xu, Gane Ka-Shu Wong, and Jun Wang. Soapdenovo-trans: de novo transcriptome assembly with short rna-seq reads. *Bioinformatics*, 30(12):1660–1666, February 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu077. URL http://dx.doi.org/10.1093/bioinformatics/btu077.

[172] Marcel H. Schulz, Daniel R. Zerbino, Martin Vingron, and Ewan Birney. Oases:robustde novorna-seq assembly across the dynamic range of expression levels. *Bioinformatics*, 28(8): 1086–1092, February 2012. ISSN 1367-4811. doi: 10.1093/bioinformatics/bts094. URL http://dx.doi.org/10.1093/bioinformatics/bts094.

[173] Gordon Robertson, Jacqueline Schein, Readman Chiu, Richard Corbett, Matthew Field, Shaun D Jackman, Karen Mungall, Sam Lee, Hisanaga Mark Okada, Jenny Q Qian, Malachi Griffith, Anthony Raymond, Nina Thiessen, Timothee Cezard, Yaron S Butterfield, Richard Newsome, Simon K Chan, Rong She, Richard Varhol, Baljit Kamoh, Anna-Liisa Prabhu, Angela Tam, YongJun Zhao, Richard A Moore, Martin Hirst, Marco A Marra, Steven J M Jones, Pamela A Hoodless, and Inanc Birol. De novo assembly and analysis of rna-seq data. *Nature Methods*, 7 (11):909–912, October 2010. ISSN 1548-7105. doi: 10.1038/nmeth.1517. URL http://dx.doi.org/10.1038/nmeth.1517.

[174] Yu Peng, Henry C. M. Leung, Siu-Ming Yiu, Ming-Ju Lv, Xin-Guang Zhu, and Francis Y. L. Chin. Idba-tran: a more robust de novo de bruijn graph assembler for transcriptomes with uneven expression levels. *Bioinformatics*, 29(13):i326–i334, June 2013. ISSN 1367-4811. doi: 10.1093/bioinformatics/btt219. URL http://dx.doi.org/10.1093/bioinformatics/btt219.

[175] Ka Ming Nip, Readman Chiu, Chen Yang, Justin Chu, Hamid Mohamadi, René L. Warren, and Inanc Birol. Rna-bloom enables reference-free and reference-guided sequence assembly for single-cell transcriptomes. *Genome Research*, 30(8):1191–1200, August 2020. ISSN 1549-5469. doi: 10.1101/gr.260174.119. URL http://dx.doi.org/10.1101/gr.260174.119.

[176] Jeffrey A. Martin and Zhong Wang. Next-generation transcriptome assembly. *Nature Reviews Genetics*, 12(10):671–682, September 2011. ISSN 1471-0064. doi: 10.1038/nrg3068. URL http://dx.doi.org/10.1038/nrg3068.

[177] Carolina Monzó, Tianyuan Liu, and Ana Conesa. Transcriptomics in the era of long-read sequencing. *Nature Reviews Genetics*, March 2025. ISSN 1471-0064. doi: 10.1038/s41576-025-00828-z. URL http://dx.doi.org/10.1038/s41576-025-00828-z.

[178] Donald M. Bryant, Kimberly Johnson, Tia DiTommaso, Timothy Tickle, Matthew Brian Couger, Duygu Payzin-Dogru, Tae J. Lee, Nicholas D. Leigh, Tzu-Hsing Kuo, Francis G. Davis, Joel Bateman, Sevara Bryant, Anna R. Guzikowski, Stephanie L. Tsai, Steven Coyne, William W. Ye, Robert M. Freeman, Leonid Peshkin, Clifford J. Tabin, Aviv Regev, Brian J. Haas, and Jessica L. Whited. A Tissue-Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration Factors. *Cell Reports*, 18(3):762–776, January 2017. ISSN 2211-1247. doi: 10.1016/j.celrep.2016.12.063. URL https://www.cell.com/cell-reports/abstract/S2211-1247(16)31770-3. Publisher: Elsevier.

[179] Ben Langmead and Steven L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359, April 2012. ISSN 1548-7105. doi: 10.1038/nmeth.1923. URL https://www.nature.com/articles/nmeth.1923. Publisher: Nature Publishing Group.

[180] Mosè Manni, Matthew R Berkeley, Mathieu Seppey, Felipe A Simão, and Evgeny M Zdobnov. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution*, 38(10):4647–4654, October 2021. ISSN 1537-1719. doi: 10.1093/molbev/msab199. URL https://doi.org/10.1093/molbev/msab199.

[181] Neng Huang and Heng Li. compleasm: a faster and more accurate reimplementation of BUSCO. *Bioinformatics*, 39(10):btad595, October 2023. ISSN 1367-4811. doi: 10.1093/bioinformatics/btad595. URL https://doi.org/10.1093/bioinformatics/btad595.

[182] Evgeny M Zdobnov, Dmitry Kuznetsov, Fredrik Tegenfeldt, Mosè Manni, Matthew Berkeley, and Evgenia V Kriventseva. OrthoDB in 2020: evolutionary and functional annotations of orthologs. *Nucleic Acids Research*, 49(D1):D389–D393, January 2021. ISSN 0305-1048. doi: 10.1093/nar/gkaa1009. URL https://doi.org/10.1093/nar/gkaa1009.

[183] B. Haas and A. Papanicolaou. Transdecoder (find coding regions within transcripts). *Trans-*

Decoder (Find Coding Regions Within Transcripts), 2016. URL https://github.com/TransDecoder/TransDecoder.

[184] Limin Fu, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152, December 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts565. URL https://doi.org/10.1093/bioinformatics/bts565.

[185] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):1–10, March 2009. ISSN 1474-760X. doi: 10.1186/gb-2009-10-3-r25. URL https://genomebiology.biomedcentral.com/articles/10.1186/gb-2009-10-3-r25. Number: 3 Publisher: BioMed Central.

[186] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21, October 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts635. URL http://dx.doi.org/10.1093/bioinformatics/bts635.

[187] Bo Li and Colin N. Dewey. Rsem: Accurate transcript quantification from rna-seq data with or without a reference genome. *BMC Bioinformatics*, 12, 2011. doi: 10.1186/1471-2105-12-323.

[188] Nicolas L Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic rna-seq quantification. *Nature Biotechnology*, 34(5):525–527, April 2016. ISSN 1546-1696. doi: 10.1038/nbt.3519. URL http://dx.doi.org/10.1038/nbt.3519.

[189] Rob Patro, Geet Duggal, Michael I Love, Rafael A Irizarry, and Carl Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14(4):417–419, March 2017. ISSN 1548-7105. doi: 10.1038/nmeth.4197. URL http://dx.doi.org/10.1038/nmeth.4197.

[190] Tom Sean Smith, Andreas Heger, and Ian Sudbery. UMI-tools: Modelling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Research*, page gr.209601.116, January 2017. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.209601.116. URL http://genome.cshlp.org/content/early/2017/01/18/gr.209601.116. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[191] Mark D Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of rna-seq data. *Genome Biology*, 11(3), March 2010. ISSN 1474-760X. doi: 10.1186/gb-2010-11-3-r25. URL http://dx.doi.org/10.1186/gb-2010-11-3-r25.

[192] Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. Blast+: Architecture and applications. *BMC Bioinformatics*, 10, 2009. doi: 10.1186/1471-2105-10-421.

[193] The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D617, January 2025. ISSN 1362-4962. doi: 10.1093/nar/gkae1010. URL https://doi.org/10.1093/nar/gkae1010.

[194] K. D. Pruitt. Ncbi reference sequence (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, 33(Database issue):D501–D504, December 2004. ISSN 1362-4962. doi: 10.1093/nar/gki025. URL http://dx.doi.org/10.1093/nar/gki025.

[195] Baris E. Suzek, Yuqi Wang, Hongzhan Huang, Peter B. McGarvey, and Cathy H. Wu. Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, November 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu739. URL http://dx.doi.org/10.1093/bioinformatics/btu739.

[196] Nicola Bordin, Ian Sillitoe, Vamsi Nallapareddy, Clemens Rauer, Su Datt Lam, Vaishali P. Waman, Neeladri Sen, Michael Heinzinger, Maria Littmann, Stephanie Kim, Sameer Velankar,

Martin Steinegger, Burkhard Rost, and Christine Orengo. AlphaFold2 reveals commonalities and novelties in protein structure space for 21 model organisms. *Communications Biology*, 6(1):1–12, February 2023. ISSN 2399-3642. doi: 10.1038/s42003-023-04488-9. URL https://www.nature.com/articles/s42003-023-04488-9. Publisher: Nature Publishing Group.

[197] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N. Kinch, R. Dustin Schaeffer, Claudia Millán, Hahnbeom Park, Carson Adams, Caleb R. Glassman, Andy DeGiovanni, Jose H. Pereira, Andria V. Rodrigues, Alberdina A. van Dijk, Ana C. Ebrecht, Diederik J. Opperman, Theo Sagmeister, Christoph Buhlheller, Tea Pavkov-Keller, Manoj K. Rathinaswamy, Udit Dalwadi, Calvin K. Yip, John E. Burke, K. Christopher Garcia, Nick V. Grishin, Paul D. Adams, Randy J. Read, and David Baker. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, August 2021. ISSN 1095-9203. doi: 10.1126/science.abj8754. URL http://dx.doi.org/10.1126/science.abj8754.

[198] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan Dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, March 2023. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/10.1126/science.ade2574.

[199] Michel van Kempen, Stephanie S. Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron L. M. Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with Foldseek. *Nature Biotechnology*, 42(2):243–246, February 2024. ISSN 1546-1696. doi: 10.1038/s41587-023-01773-0. URL https://www.nature.com/articles/s41587-023-01773-0. Publisher: Nature Publishing Group.

[200] Stephen K Burley, Charmi Bhikadiya, Chunxiao Bi, Sebastian Bittrich, Li Chen, Gregg V Crichlow, Cole H Christie, Kenneth Dalenberg, Luigi Di Costanzo, Jose M Duarte, Shuchismita Dutta, Zukang Feng, Sai Ganesan, David S Goodsell, Sutapa Ghosh, Rachel Kramer Green, Vladimir Guranović, Dmytro Guzenko, Brian P Hudson, Catherine L Lawson, Yuhe Liang, Robert Lowe, Harry Namkoong, Ezra Peisach, Irina Persikova, Chris Randle, Alexander Rose, Yana Rose, Andrej Sali, Joan Segura, Monica Sekharan, Chenghua Shao, Yi-Ping Tao, Maria Voigt, John D Westbrook, Jasmine Y Young, Christine Zardecki, and Marina Zhuravleva. RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Research*, 49(D1):D437–D451, January 2021. ISSN 0305-1048. doi: 10.1093/nar/gkaa1038. URL https://doi.org/10.1093/nar/gkaa1038.

[201] Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, Augustin Zídek, Tim Green, Kathryn Tunyasuvunakool, Stig Petersen, John Jumper, Ellen Clancy, Richard Green, Ankur Vora, Mira Lutfi, Michael Figurnov, Andrew Cowie, Nicole Hobbs, Pushmeet Kohli, Gerard Kleywegt, Ewan Birney, Demis Hassabis, and Sameer Velankar. Alphafold protein structure database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50(D1):D439 – D444, 2022. doi: 10.1093/nar/gkab1061.

[202] Gene Ontology Consortium. The gene ontology project in 2008. *Nucleic Acids Research*, 36: D440–D444, November 2007. ISSN 1362-4962. doi: 10.1093/nar/gkm883. URL http://dx.doi.org/10.1093/nar/gkm883.

[203] Gene Ontology Consortium. *The Gene Ontology Handbook*. Springer New York, 2017. ISBN 9781493937431. doi: 10.1007/978-1-4939-3743-1. URL http://dx.doi.org/10.1007/978-1-4939-3743-1.

[204] Minoru Kanehisa, Miho Furumichi, Yoko Sato, Yuriko Matsuura, and Mari Ishiguro-Watanabe.

KEGG: biological systems database as a model of the real world. *Nucleic Acids Research*, 53 (D1):D672–D677, January 2025. ISSN 1362-4962. doi: 10.1093/nar/gkae909. URL https://doi.org/10.1093/nar/gkae909.

[205] Ana Conesa, Stefan Götz, Juan Miguel García-Gómez, Javier Terol, Manuel Talón, and Montserrat Robles. Blast2go: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 21(18):3674–3676, August 2005. ISSN 1367-4803. doi: 10.1093/bioinformatics/bti610. URL http://dx.doi.org/10.1093/bioinformatics/bti610.

[206] Takuya Aramaki, Romain Blanc-Mathieu, Hisashi Endo, Koichi Ohkubo, Minoru Kanehisa, Susumu Goto, and Hiroyuki Ogata. Kofamkoala: Kegg ortholog assignment based on profile hmm and adaptive score threshold. *Bioinformatics*, 36(7):2251–2252, November 2019. ISSN 1367-4811. doi: 10.1093/bioinformatics/btz859. URL http://dx.doi.org/10.1093/bioinformatics/btz859.

[207] Philip Jones, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, John Maslen, Alex Mitchell, Gift Nuka, Sebastien Pesseat, Antony F. Quinn, Amaia Sangrador-Vegas, Maxim Scheremetjew, Siew-Yit Yong, Rodrigo Lopez, and Sarah Hunter. Interproscan 5: genome-scale protein function classification. *Bioinformatics*, 30(9):1236–1240, January 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu031. URL http://dx.doi.org/10.1093/bioinformatics/btu031.

[208] Matthias Blum, Hsin-Yu Chang, Sara Chuguransky, Tiago Grego, Swaathi Kandasaamy, Alex Mitchell, Gift Nuka, Typhaine Paysan-Lafosse, Matloob Qureshi, Shriya Raj, Lorna Richardson, Gustavo A Salazar, Lowri Williams, Peer Bork, Alan Bridge, Julian Gough, Daniel H Haft, Ivica Letunic, Aron Marchler-Bauer, Huaiyu Mi, Darren A Natale, Marco Necci, Christine A Orengo, Arun P Pandurangan, Catherine Rivoire, Christian J A Sigrist, Ian Sillitoe, Narmada Thanki, Paul D Thomas, Silvio C E Tosatto, Cathy H Wu, Alex Bateman, and Robert D Finn. The interpro protein families and domains database: 20 years on. *Nucleic Acids Research*, 49(D1): D344–D354, November 2020. ISSN 1362-4962. doi: 10.1093/nar/gkaa977. URL http://dx.doi.org/10.1093/nar/gkaa977.

[209] Vladimir Gligorijević, P. Douglas Renfrew, Tomasz Kosciolek, Julia Koehler Leman, Daniel Berenberg, Tommi Vatanen, Chris Chandler, Bryn C. Taylor, Ian M. Fisk, Hera Vlamakis, Ramnik J. Xavier, Rob Knight, Kyunghyun Cho, and Richard Bonneau. Structure-based protein function prediction using graph convolutional networks. *Nature Communications*, 12(1), 2021. doi: 10.1038/s41467-021-23303-9.

[210] Felix Teufel, José Juan Almagro Armenteros, Alexander Rosenberg Johansen, Magnús Halldór Gíslason, Silas Irby Pihl, Konstantinos D. Tsirigos, Ole Winther, Søren Brunak, Gunnar von Heijne, and Henrik Nielsen. SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nature Biotechnology*, 40(7):1023–1025, July 2022. ISSN 1546-1696. doi: 10.1038/s41587-021-01156-3. URL https://www.nature.com/articles/s41587-021-01156-3. Publisher: Nature Publishing Group.

[211] Jeppe Hallgren, Konstantinos D. Tsirigos, Mads Damgaard Pedersen, José Juan Almagro Armenteros, Paolo Marcatili, Henrik Nielsen, Anders Krogh, and Ole Winther. DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks, April 2022. URL http://biorxiv.org/lookup/doi/10.1101/2022.04.08.487609.

[212] Eric P. Nawrocki and Sean R. Eddy. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, 29(22):2933–2935, November 2013. ISSN 1367-4803. doi: 10.1093/bioinformatics/btt509. URL https://doi.org/10.1093/bioinformatics/btt509.

[213] Carlos P Cantalapiedra, Ana Hernández-Plaza, Ivica Letunic, Peer Bork, and Jaime Huerta-Cepas. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution*, 38(12):5825–5829, December 2021. ISSN 1537-1719. doi: 10.1093/molbev/msab293. URL https://doi.org/10.1093/molbev/msab293.

[214] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL https://www.R-project.org/.

[215] Jari Oksanen, Gavin L. Simpson, F. Guillaume Blanchet, Roeland Kindt, Pierre Legendre, Peter R. Minchin, R.B. O'Hara, Peter Solymos, M. Henry H. Stevens, Eduard Szoecs, Helene Wagner, Matt Barbour, Michael Bedward, Ben Bolker, Daniel Borcard, Tuomas Borman, Gustavo Carvalho, Michael Chirico, Miquel De Caceres, Sebastien Durand, Heloisa Beatriz Antoniazi Evangelista, Rich FitzJohn, Michael Friendly, Brendan Furneaux, Geoffrey Hannigan, Mark O. Hill, Leo Lahti, Cameron Martino, Dan McGlinn, Marie-Helene Ouellette, Eduardo Ribeiro Cunha, Tyler Smith, Adrian Stier, Cajo J.F. Ter Braak, and James Weedon. *vegan: Community Ecology Package*, 2025. URL https://vegandevs.github.io/vegan/. R package version 2.8-0.

[216] Thomaz F. S. Bastiaanssen, Thomas P. Quinn, and Amy Loughman. Bugs as features (part 1): concepts and foundations for the compositional data analysis of the microbiome–gut–brain axis. *Nature Mental Health*, 1(12):930–938, December 2023. ISSN 2731-6076. doi: 10.1038/s44220-023-00148-3. URL http://dx.doi.org/10.1038/s44220-023-00148-3.

[217] M. O. Hill. Diversity and evenness: A unifying notation and its consequences. *Ecology*, 54(2): 427–432, March 1973. ISSN 1939-9170. doi: 10.2307/1934352. URL http://dx.doi.org/10.2307/1934352.

[218] Erich Schubert and Michael Gertz. *Intrinsic t-Stochastic Neighbor Embedding for Visualization and Outlier Detection*, page 188–203. Springer International Publishing, 2017. ISBN 9783319684741. doi: 10.1007/978-3-319-68474-1_13. URL http://dx.doi.org/10.1007/978-3-319-68474-1_13.

[219] John Healy and Leland McInnes. Uniform manifold approximation and projection. *Nature Reviews Methods Primers*, 4(1), November 2024. ISSN 2662-8449. doi: 10.1038/s43586-024-00363-x. URL http://dx.doi.org/10.1038/s43586-024-00363-x.

[220] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):1–21, December 2014. ISSN 1474-760X. doi: 10.1186/s13059-014-0550-8. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8. Number: 12 Publisher: BioMed Central.

[221] Huijuan Zhou, Kejun He, Jun Chen, and Xianyang Zhang. LinDA: linear models for differential abundance analysis of microbiome compositional data. *Genome Biology*, 23(1):1–23, December 2022. ISSN 1474-760X. doi: 10.1186/s13059-022-02655-5. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-022-02655-5. Number: 1 Publisher: BioMed Central.

[222] Huang Lin and Shyamal Das Peddada. Analysis of compositions of microbiomes with bias correction. *Nature Communications*, 11(1):3514, July 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-17041-7. URL https://www.nature.com/articles/s41467-020-17041-7. Publisher: Nature Publishing Group.

[223] Andrew D. Fernandes, Jennifer NS Reid, Jean M. Macklaim, Thomas A. McMurrough, David R. Edgell, and Gregory B. Gloor. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome*, 2(1):1–13, December 2014. ISSN 2049-2618. doi: 10.1186/2049-2618-2-15. URL https://microbiomejournal.biomedcentral.com/articles/10.1186/2049-2618-2-15. Number: 1 Publisher: BioMed Central.

[224] Himel Mallick, Ali Rahnavard, Lauren J. McIver, Siyuan Ma, Yancong Zhang, Long H. Nguyen, Timothy L. Tickle, George Weingart, Boyu Ren, Emma H. Schwager, Suvo Chatterjee, Kelsey N. Thompson, Jeremy E. Wilkinson, Ayshwarya Subramanian, Yiren Lu, Levi Waldron, Joseph N. Paulson, Eric A. Franzosa, Hector Corrada Bravo, and Curtis Huttenhower. Multivariable association discovery in population-scale meta-omics studies. *PLOS Computational Biology*, 17(11):e1009442, November 2021. ISSN 1553-7358. doi: 10.1371/journal.pcbi.

1009442. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1009442. Publisher: Public Library of Science.

[225] Yunshun Chen, Lizhong Chen, Aaron T L Lun, Pedro L Baldoni, and Gordon K Smyth. edger v4: powerful differential analysis of sequencing data with expanded functionality and improved support for small counts and larger datasets. *Nucleic Acids Research*, 53(2), January 2025. ISSN 1362-4962. doi: 10.1093/nar/gkaf018. URL http://dx.doi.org/10.1093/nar/gkaf018.

[226] Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47, January 2015. ISSN 0305-1048. doi: 10.1093/nar/gkv007. URL http://dx.doi.org/10.1093/nar/gkv007.

[227] Bryan D. Martin, Daniela Witten, and Amy D. Willis. Modeling microbial abundances and dysbiosis with beta-binomial regression. *The Annals of Applied Statistics*, 14(1), March 2020. ISSN 1932-6157. doi: 10.1214/19-aoas1283. URL http://dx.doi.org/10.1214/19-AOAS1283.

[228] Tiantian Liu, Hongyu Zhao, and Tao Wang. An empirical bayes approach to normalization and differential abundance testing for microbiome data. *BMC Bioinformatics*, 21(1), June 2020. ISSN 1471-2105. doi: 10.1186/s12859-020-03552-z. URL http://dx.doi.org/10.1186/s12859-020-03552-z.

[229] Stefano Mangiola, Alexandra J. Roth-Schulze, Marie Trussart, Enrique Zozaya-Valdés, Mengyao Ma, Zijie Gao, Alan F. Rubin, Terence P. Speed, Heejung Shim, and Anthony T. Papenfuss. sccomp: Robust differential composition and variability analysis for single-cell data. *Proceedings of the National Academy of Sciences*, 120(33), August 2023. ISSN 1091-6490. doi: 10.1073/pnas.2203828120. URL http://dx.doi.org/10.1073/pnas.2203828120.

[230] Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn, and Lior Pachter. Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. *Nature Protocols*, 7(3):562–578, March 2012. ISSN 1750-2799. doi: 10.1038/nprot.2012.016. URL http://dx.doi.org/10.1038/nprot.2012.016.

[231] Sonia Tarazona, Fernando García, Alberto Ferrer, Joaquín Dopazo, and Ana Conesa. Noiseq: a rna-seq differential expression method robust for sequencing depth biases. *EMBnet.journal*, 17 (B):18, February 2012. ISSN 2226-6089. doi: 10.14806/ej.17.b.265. URL http://dx.doi.org/10.14806/ej.17.B.265.

[232] Harold Pimentel, Nicolas L Bray, Suzette Puente, Páll Melsted, and Lior Pachter. Differential analysis of rna-seq incorporating quantification uncertainty. *Nature Methods*, 14(7):687–690, June 2017. ISSN 1548-7105. doi: 10.1038/nmeth.4324. URL http://dx.doi.org/10.1038/nmeth.4324.

[233] Juliana Costa-Silva, Douglas Domingues, and Fabricio Martins Lopes. RNA-Seq differential expression analysis: An extended review and a software tool. *PLOS ONE*, 12(12):e0190152, December 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0190152. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0190152. Publisher: Public Library of Science.

[234] Jacob T. Nearing, Gavin M. Douglas, Molly G. Hayes, Jocelyn MacDonald, Dhwani K. Desai, Nicole Allward, Casey M. A. Jones, Robyn J. Wright, Akhilesh S. Dhanani, André M. Comeau, and Morgan G. I. Langille. Microbiome differential abundance methods produce different results across 38 datasets. *Nature Communications*, 13(1):342, January 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-28034-z. URL https://www.nature.com/articles/s41467-022-28034-z. Publisher: Nature Publishing Group.

[235] Juho Pelto, Kari Auranen, Janne V Kujala, and Leo Lahti. Elementary methods provide more replicable results in microbial differential abundance analysis. *Briefings in Bioinformatics*, 26 (2), March 2025. ISSN 1477-4054. doi: 10.1093/bib/bbaf130. URL http://dx.doi.org/10.1093/bib/bbaf130.

[236] Weijun Luo and Cory Brouwer. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics*, 29(14):1830–1831, July 2013. ISSN 1367-4803. doi: 10.1093/bioinformatics/btt285. URL https://doi.org/10.1093/bioinformatics/btt285.

[237] Emden R. Gansner and Stephen C. North. An open graph visualization system and its applications to software engineering. *Software: Practice and Experience*, 30(11):1203–1233, 2000. ISSN 1097-024X. doi: 10.1002/1097-024x(200009)30:11⟨1203::aid-spe338⟩3.0.co;2-n. URL http://dx.doi.org/10.1002/1097-024X(200009)30:11<1203::AID-SPE338>3.0.CO;2-N.

[238] Gennady Korotkevich, Vladimir Sukhov, Nikolay Budin, Boris Shpak, Maxim N. Artyomov, and Alexey Sergushichev. Fast gene set enrichment analysis, February 2021. URL https://www.biorxiv.org/content/10.1101/060012v3. Pages: 060012 Section: New Results.

[239] Matthias C. Rillig and Michael Bonkowski. Microplastic and soil protists: A call for research. *Environmental Pollution*, 241:1128–1131, October 2018. ISSN 0269-7491. doi: 10.1016/j.envpol.2018.04.147. URL http://dx.doi.org/10.1016/j.envpol.2018.04.147.

[240] Sarika Kumari, Komal A Chandarana, and Natarajan Amaresan. Protists as a potential microbial tool for environmental microplastics remediation. *Environmental Science: Processes and Impacts*, 2025. ISSN 2050-7895. doi: 10.1039/d5em00623f. URL http://dx.doi.org/10.1039/D5EM00623F.

[241] Robert L. Beschta and William J. Ripple. Wolves, trophic cascades, and rivers in the olympic national park, usa. *Ecohydrology*, 1(2):118–130, January 2008. ISSN 1936-0592. doi: 10.1002/eco.12. URL http://dx.doi.org/10.1002/eco.12.

[242] Kristine M Averill, David A Mortensen, Erica A H Smithwick, Susan Kalisz, William J McShea, Norman A Bourg, John D Parker, Alejandro A Royo, Marc D Abrams, David K Apsley, Bernd Blossey, Douglas H Boucher, Kai L Caraher, Antonio DiTommaso, Sarah E Johnson, Robert Masson, and Victoria A Nuzzo. A regional assessment of white-tailed deer effects on plant invasion. *AoB PLANTS*, 10(1), December 2017. ISSN 2041-2851. doi: 10.1093/aobpla/plx047. URL http://dx.doi.org/10.1093/aobpla/plx047.

[243] Alexandra Z. Worden, Michael J. Follows, Stephen J. Giovannoni, Susanne Wilken, Amy E. Zimmerman, and Patrick J. Keeling. Rethinking the marine carbon cycle: Factoring in the multifarious lifestyles of microbes. *Science*, 347(6223), 2015. doi: 10.1126/science.1257594.

[244] Anita Narwani, Marta Reyes, Aaron Louis Pereira, Hannele Penson, Stuart R. Dennis, Samuel Derrer, Piet Spaak, and Blake Matthews. Interactive effects of foundation species on ecosystem functioning and stability in response to disturbance. *Proceedings of the Royal Society B: Biological Sciences*, 286(1913):20191857, October 2019. doi: 10.1098/rspb.2019.1857. URL https://royalsocietypublishing.org/doi/10.1098/rspb.2019.1857. Publisher: Royal Society.

[245] Vasilis Dakos, Blake Matthews, Andrew P. Hendry, Jonathan Levine, Nicolas Loeuille, Jon Norberg, Patrik Nosil, Marten Scheffer, and Luc De Meester. Ecosystem tipping points in an evolving world. *Nature Ecology and Evolution*, 3(3):355 – 362, 2019. doi: 10.1038/s41559-019-0797-2.

[246] Aidan C. Parte, Joaquim Sardà Carbasse, Jan P. Meier-Kolthoff, Lorenz C. Reimer, and Markus Göker. List of prokaryotic names with standing in nomenclature (lpsn) moves to the dsmz. *International Journal of Systematic and Evolutionary Microbiology*, 70(11):5607–5612, November 2020. ISSN 1466-5034. doi: 10.1099/ijsem.0.004332. URL http://dx.doi.org/10.1099/ijsem.0.004332.

[247] Jian-Yu Jiao, Lan Liu, Zheng-Shuang Hua, Bao-Zhu Fang, En-Min Zhou, Nimaichand Salam, Brian P Hedlund, and Wen-Jun Li. Microbial dark matter coming to light: challenges and opportunities. *National Science Review*, 8(3), December 2020. ISSN 2053-714X. doi: 10.1093/nsr/nwaa280. URL http://dx.doi.org/10.1093/nsr/nwaa280.

[248] Mark Blaxter. Revealing the dark matter of the genome. *Science*, 330(6012):1758–1759, December 2010. ISSN 1095-9203. doi: 10.1126/science.1200700. URL http://dx.doi.org/10.1126/science.1200700.

[249] Mark T. W. Ebbert, Tanner D. Jensen, Karen Jansen-West, Jonathon P. Sens, Joseph S. Reddy, Perry G. Ridge, John S. K. Kauwe, Veronique Belzil, Luc Pregent, Minerva M. Carrasquillo, Dirk Keene, Eric Larson, Paul Crane, Yan W. Asmann, Nilufer Ertekin-Taner, Steven G. Younkin, Owen A. Ross, Rosa Rademakers, Leonard Petrucelli, and John D. Fryer. Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. *Genome Biology*, 20(1), May 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1707-2. URL http://dx.doi.org/10.1186/s13059-019-1707-2.

[250] Zhenfeng Liu, Sarah K Hu, Victoria Campbell, Avery O Tatters, Karla B Heidelberg, and David A Caron. Single-cell transcriptomics of small microbial eukaryotes: limitations and potential. *The ISME Journal*, 11(5):1282–1285, May 2017. ISSN 1751-7362. doi: 10.1038/ismej.2016.190. URL https://doi.org/10.1038/ismej.2016.190.

[251] David M. Suter, Nacho Molina, David Gatfield, Kim Schneider, Ueli Schibler, and Felix Naef. Mammalian genes are transcribed with widely different bursting kinetics. *Science*, 332(6028): 472 – 474, 2011. doi: 10.1126/science.1198817.

[252] Patrick J. Keeling and Javier Del Campo. Marine protists are not just big bacteria. *Current Biology*, 27(11):R541–R549, June 2017. ISSN 0960-9822. doi: 10.1016/j.cub.2017.03.075. URL http://dx.doi.org/10.1016/j.cub.2017.03.075.

[253] Drahomíra Faktorová, R. Ellen R. Nisbet, José A. Fernández Robledo, Elena Casacuberta, Lisa Sudek, Andrew E. Allen, Manuel Ares, Cristina Aresté, Cecilia Balestreri, Adrian C. Barbrook, Patrick Beardslee, Sara Bender, David S. Booth, François-Yves Bouget, Chris Bowler, Susana A. Breglia, Colin Brownlee, Gertraud Burger, Heriberto Cerutti, Rachele Cesaroni, Miguel A. Chiurillo, Thomas Clemente, Duncan B. Coles, Jackie L. Collier, Elizabeth C. Cooney, Kathryn Coyne, Roberto Docampo, Christopher L. Dupont, Virginia Edgcomb, Elin Einarsson, Pía A. Elustondo, Fernan Federici, Veronica Freire-Beneitez, Nastasia J. Freyria, Kodai Fukuda, Paulo A. García, Peter R. Girguis, Fatma Gomaa, Sebastian G. Gornik, Jian Guo, Vladimír Hampl, Yutaka Hanawa, Esteban R. Haro-Contreras, Elisabeth Hehenberger, Andrea Highfield, Yoshihisa Hirakawa, Amanda Hopes, Christopher J. Howe, Ian Hu, Jorge Ibañez, Nicholas A. T. Irwin, Yuu Ishii, Natalia Ewa Janowicz, Adam C. Jones, Ambar Kachale, Konomi Fujimura-Kamada, Binnypreet Kaur, Jonathan Z. Kaye, Eleanna Kazana, Patrick J. Keeling, Nicole King, Lawrence A. Klobutcher, Noelia Lander, Imen Lassadi, Zhuhong Li, Senjie Lin, Jean-Claude Lozano, Fulei Luan, Shinichiro Maruyama, Tamara Matute, Cristina Miceli, Jun Minagawa, Mark Moosburner, Sebastián R. Najle, Deepak Nanjappa, Isabel C. Nimmo, Luke Noble, Anna M. G. Novák Vanclová, Mariusz Nowacki, Isaac Nuñez, Arnab Pain, Angela Piersanti, Sandra Pucciarelli, Jan Pyrih, Joshua S. Rest, Mariana Rius, Deborah Robertson, Albane Ruaud, Iñaki Ruiz-Trillo, Monika A. Sigg, Pamela A. Silver, Claudio H. Slamovits, G. Jason Smith, Brittany N. Sprecher, Rowena Stern, Estienne C. Swart, Anastasios D. Tsaousis, Lev Tsypin, Aaron Turkewitz, Jernej Turnšek, Matus Valach, Valérie Vergé, Peter von Dassow, Tobias von der Haar, Ross F. Waller, Lu Wang, Xiaoxue Wen, Glen Wheeler, April Woods, Huan Zhang, Thomas Mock, Alexandra Z. Worden, and Julius Lukeš. Genetic tool development in marine protists: emerging model organisms for experimental cell biology. *Nature Methods*, 17(5):481–494, April 2020. ISSN 1548-7105. doi: 10.1038/s41592-020-0796-x. URL http://dx.doi.org/10.1038/s41592-020-0796-x.

[254] Donovan H Parks, Maria Chuvochina, David W Waite, Christian Rinke, Adam Skarshewski, Pierre-Alain Chaumeil, and Philip Hugenholtz. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology*, 36(10):996–1004, August 2018. ISSN 1546-1696. doi: 10.1038/nbt.4229. URL http://dx.doi.org/10.1038/nbt.4229.

[255] Ramon Massana, Angélique Gobet, Stéphane Audic, David Bass, Lucie Bittner, Christophe Boutte, Aurélie Chambouvet, Richard Christen, Jean-Michel Claverie, Johan Decelle, John R.

Dolan, Micah Dunthorn, Bente Edvardsen, Irene Forn, Dominik Forster, Laure Guillou, Olivier Jaillon, Wiebe H. C. F. Kooistra, Ramiro Logares, Frédéric Mahé, Fabrice Not, Hiroyuki Ogata, Jan Pawlowski, Massimo C. Pernice, Ian Probert, Sarah Romac, Thomas Richards, Sébastien Santini, Kamran Shalchian-Tabrizi, Raffaele Siano, Nathalie Simon, Thorsten Stoeck, Daniel Vaulot, Adriana Zingone, and Colomban de Vargas. Marine protist diversity in ¡scp¿e¡/scp¿uropean coastal waters and sediments as revealed by high-throughput sequencing. *Environmental Microbiology*, 17(10):4035–4049, August 2015. ISSN 1462-2920. doi: 10.1111/1462-2920.12955. URL http://dx.doi.org/10.1111/1462-2920.12955.

[256] Raquel Rodríguez-Martínez, Matthias Labrenz, Javier Del Campo, Irene Forn, Klaus Jürgens, and Ramon Massana. Distribution of the uncultured protist mast-4 in the indian ocean, drake passage and mediterranean sea assessed by real-time quantitative pcr. *Environmental Microbiology*, 11(2):397–408, January 2009. ISSN 1462-2920. doi: 10.1111/j.1462-2920.2008.01779.x. URL http://dx.doi.org/10.1111/j.1462-2920.2008.01779.x.

[257] Fei Zhu, Ramon Massana, Fabrice Not, Dominique Marie, and Daniel Vaulot. Mapping of picoeucaryotes in marine ecosystems with quantitative pcr of the 18s rrna gene. *FEMS Microbiology Ecology*, 52(1):79–92, March 2005. ISSN 1574-6941. doi: 10.1016/j.femsec.2004.10.006. URL http://dx.doi.org/10.1016/j.femsec.2004.10.006.

[258] David J. MartÃnez-Cano, Mariana Reyes-Prieto, Esperanza MartÃnez-Romero, Laila P. Partida-MartÃnez, Amparo Latorre, AndrÃ©s Moya, and Luis Delaye. Evolution of small prokaryotic genomes. *Frontiers in Microbiology*, 5, January 2015. ISSN 1664-302X. doi: 10.3389/fmicb.2014.00742. URL http://dx.doi.org/10.3389/fmicb.2014.00742.

[259] Ashley Byrne, Charles Cole, Roger Volden, and Christopher Vollmers. Realizing the potential of full-length transcriptome sequencing. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1786):20190097, October 2019. ISSN 1471-2970. doi: 10.1098/rstb.2019.0097. URL http://dx.doi.org/10.1098/rstb.2019.0097.

[260] Christina Karmisholt Overgaard, Mahwash Jamy, Simona Radutoiu, Fabien Burki, and Morten Kam Dahl Dueholm. Benchmarking long-read sequencing strategies for obtaining ¡scp¿asv¡/scp¿-resolved ¡scp¿rrna¡/scp¿ operons from environmental microeukaryotes. *Molecular Ecology Resources*, 24(7), July 2024. ISSN 1755-0998. doi: 10.1111/1755-0998.13991. URL http://dx.doi.org/10.1111/1755-0998.13991.

[261] Alexander B. Rosenberg, Charles M. Roco, Richard A. Muscat, Anna Kuchina, Paul Sample, Zizhen Yao, Lucas T. Graybuck, David J. Peeler, Sumit Mukherjee, Wei Chen, Suzie H. Pun, Drew L. Sellers, Bosiljka Tasic, and Georg Seelig. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*, 360(6385):176–182, April 2018. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aam8999. URL https://www.science.org/doi/10.1126/science.aam8999.

[262] Beth K. Martin, Chengxiang Qiu, Eva Nichols, Melissa Phung, Rula Green-Gladden, Sanjay Srivatsan, Ronnie Blecher-Gonen, Brian J. Beliveau, Cole Trapnell, Junyue Cao, and Jay Shendure. Optimized single-nucleus transcriptional profiling by combinatorial indexing. *Nature Protocols*, 18(1):188–207, January 2023. ISSN 1750-2799. doi: 10.1038/s41596-022-00752-0. URL https://www.nature.com/articles/s41596-022-00752-0. Publisher: Nature Publishing Group.

[263] Evan Z. Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R. Bialas, Nolan Kamitaki, Emily M. Martersteck, John J. Trombetta, David A. Weitz, Joshua R. Sanes, Alex K. Shalek, Aviv Regev, and Steven A. McCarroll. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, May 2015. ISSN 0092-8674. doi: 10.1016/j.cell.2015.05.002. URL http://dx.doi.org/10.1016/j.cell.2015.05.002.

[264] Rapolas Zilionis, Juozas Nainys, Adrian Veres, Virginia Savova, David Zemmour, Allon M Klein, and Linas Mazutis. Single-cell barcoding and sequencing using droplet microfluidics. *Nature Protocols*, 12(1):44–73, December 2016. ISSN 1750-2799. doi: 10.1038/nprot.2016.154. URL http://dx.doi.org/10.1038/nprot.2016.154.

[265] Iain C. Clark, Kristina M. Fontanez, Robert H. Meltzer, Yi Xue, Corey Hayford, Aaron May-Zhang, Chris D'Amato, Ahmad Osman, Jesse Q. Zhang, Pabodha Hettige, Jacob S. A. Ishibashi, Cyrille L. Delley, Daniel W. Weisgerber, Joseph M. Replogle, Marco Jost, Kiet T. Phong, Vanessa E. Kennedy, Cheryl A. C. Peretz, Esther A. Kim, Siyou Song, William Karlon, Jonathan S. Weissman, Catherine C. Smith, Zev J. Gartner, and Adam R. Abate. Microfluidics-free single-cell genomics with templated emulsification. *Nature Biotechnology*, 41(11):1557–1566, March 2023. ISSN 1546-1696. doi: 10.1038/s41587-023-01685-z. URL http://dx.doi.org/10.1038/s41587-023-01685-z.

[266] Xue Sun, Shir-Liya Dadon, Dena Ennis, Wenpeng Fan, Muhammad Awawdy, Eli Reuveni, Adi Alajem, and Oren Ram. Expanding single-cell toolbox with a cost-effective full-length total rna droplet-based sequencing technology. *bioRxiv*, 2025. doi: 10.1101/2025.02.23.639726. URL https://www.biorxiv.org/content/early/2025/02/27/2025.02.23.639726.

[267] Elizabeth Pennisi. Encode project writes eulogy for junk dna. *Science*, 337(6099):1159–1161, September 2012. ISSN 1095-9203. doi: 10.1126/science.337.6099.1159. URL http://dx.doi.org/10.1126/science.337.6099.1159.

[268] Alina Isakova, Norma Neff, and Stephen R. Quake. Single-cell quantification of a broad rna spectrum reveals unique noncoding patterns associated with cell types and states. *Proceedings of the National Academy of Sciences*, 118(51), December 2021. ISSN 1091-6490. doi: 10.1073/pnas.2113568118. URL http://dx.doi.org/10.1073/pnas.2113568118.

[269] Fredrik Salmen, Joachim De Jonghe, Tomasz S. Kaminski, Anna Alemany, Guillermo E. Parada, Joe Verity-Legg, Ayaka Yanagida, Timo N. Kohler, Nicholas Battich, Floris van den Brekel, Anna L. Ellermann, Alfonso Martinez Arias, Jennifer Nichols, Martin Hemberg, Florian Hollfelder, and Alexander van Oudenaarden. High-throughput total rna sequencing in single cells using vasa-seq. *Nature Biotechnology*, 40(12):1780–1793, June 2022. ISSN 1546-1696. doi: 10.1038/s41587-022-01361-8. URL http://dx.doi.org/10.1038/s41587-022-01361-8.

[270] Alina Isakova, Daniel Dan Liu, Ivana Cvijovic, Rahul Sinha, Anna E Eastman, Sirle Saul, Angela Detweiler, Norma Neff, Shirit Einav, Irving L Weissman, and Stephen R. Quake. Beyond polya: scalable single-cell total rna-seq unifies coding and non-coding transcriptomics. *bioRxiv*, 2025. doi: 10.1101/2025.08.08.669394. URL https://www.biorxiv.org/content/early/2025/08/10/2025.08.08.669394.

Microorganisms are the most dominant forms of life in Earth's history. Interactions between prokaryotic and eukaryotic micro organisms play key roles in ecosystem functioning and can take many forms. This thesis aims to add single-cell transcriptomics as a high-throughput high-resolution tool to the microbial ecology toolbox to characterize natural microeukaryote function and their interactions with prokaryotes. Most environmental microeukary otes, including many in this thesis, are thought to be uncultivable in laboratory conditions and are absent from reference genome or transcriptome databases. By extending single-cell transcrip tomics to such uncultured organisms without references, the stage is set for studying microeukaryote-prokaryote functions and interactions in the stressful, heterogeneous, and ephemeral conditions that they naturally occur in.

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

Painosalama, Turku, Finland 2025

ISBN 978-952-02-0497-6 (PRINT)
ISBN 978-952-02-0498-3 (PDF)
ISSN 0082-6979 (PRINT)
ISSN 2343-3183 (ONLINE)