RESEARCH ARTICLE

Towards high-throughput parallel imaging and single-cell transcriptomics of microbial eukaryotic plankton

Vesna Grujčić[1,2]*, Sami Saarenpää[1], John Sundh[3], Bengt Sennblad[4], Benjamin Norgren[2], Meike Latz[1], Stefania Giacomello[1], Rachel A. Foster[2], Anders F. Andersson[1]*

1 Science for Life Laboratory, Department of Gene Technology, KTH Royal Institute of Technology, Stockholm, Sweden, 2 Department of Ecology, Environment and Plant Sciences, Stockholm University, Stockholm, Sweden, 3 Science for Life Laboratory, Department of Biochemistry and Biophysics, National Bioinformatics Infrastructure Sweden, Stockholm University, Solna, Sweden, 4 Science for Life Laboratory, Dept of Cell and Molecular Biology, National Bioinformatics Infrastructure Sweden, Uppsala University, Uppsala, Sweden

* anders.andersson@scilifelab.se (AFA); vesna.grujcic@scilifelab.se (VG)

Abstract

Single-cell transcriptomics has the potential to provide novel insights into poorly studied microbial eukaryotes. Although several such technologies are available and benchmarked on mammalian cells, few have been tested on protists. Here, we applied a microarray single-cell sequencing (MASC-seq) technology, that generates microscope images of cells in parallel with capturing their transcriptomes, on three species representing important plankton groups with different cell structures; the ciliate *Tetrahymena thermophila*, the diatom *Phaeodactylum tricornutum*, and the dinoflagellate *Heterocapsa* sp. Both the cell fixation and permeabilization steps were adjusted. For the ciliate and dinoflagellate, the number of transcripts of microarray spots with single cells were significantly higher than for background spots, and the overall expression patterns were correlated with that of bulk RNA, while for the much smaller diatom cells, it was not possible to separate single-cell transcripts from background. The MASC-seq method holds promise for investigating "microbial dark matter", although further optimizations are necessary to increase the signal-to-noise ratio.

Introduction

Planktonic microbial eukaryotes (protists) function as primary producers, grazers and parasites, and are important in carbon and nutrient cycling [1–4]. Currently, however, we know relatively little about the biology of most microbial eukaryotes and even less about how they interact and function at the ecosystem level [5]. The genome is the blueprint for understanding the physiology, ecology and evolution of an organism, but the majority of protists are difficult to culture and therefore obtaining genomic references for them has been challenging, and to date there are only a few reference genomes available [6, 7]. Shotgun metagenomics and binning allows recovering genomes without cultivation, and is effective for prokaryotes [8, 9].

However, due to the typically low relative abundance and higher complexity of eukaryotic genomes, this approach has so far had more limited success for protists (but see e.g. [10]).

Genomes of uncultivated plankton can also be obtained through single-cell sequencing [11]. Since eukaryotic genomes are complex, an attractive alternative is to sequence the transcriptome (RNA sequencing; RNA-seq), which allows characterization of expressed genes without having to sequence the sometimes vast non-coding regions that are common in eukaryotic genomes. Single-cell RNA sequencing (scRNA-seq) emerged about a decade ago, as a powerful tool in biomedical research to study cell-to-cell variability [12] and has mainly been applied to mammalian cells. Applications of the method to cell types such as plants [13] and other eukaryotes [14, 15] are slowly emerging but still scarce, since the process requires vigorous adaptation of the protocol. Application of scRNA-seq to microbial eukaryotes, in comparison with bulk RNA-seq, has great potential for revealing ecophysiological properties from basic traits such as phototrophy, phagotrophy or osmotrophy, to more elaborate functions such as nutrient uptake, storage systems and interactions with other microbes [16, 17]. Concurrently, our understanding of cell states in microbial eukaryotes is limited and there are ongoing research efforts to better understand the diversity and complexity of their biology. Some microbial eukaryotes can transition between vegetative cells and cysts or spores and some can undergo sexual reproduction with different cell types [18, 19]. These different cell types could be considered as states but they are much less studied compared to those in multicellular eukaryotes [20]. Currently, very few microbial eukaryotes are available in the cultures and the ones that are, might not reflect the ecologically most important genera. The application of single cell RNA sequencing (scRNA-seq) will likely enhance our understanding of this area.

To date, only a few different scRNA-seq protocols and methods have been tested on microbial eukaryotes, mostly due to difficulty of adapting protocols to extensive variation in cell size and cell surface structure. Recent studies using versions of SMART-seq showed performances comparable to bulk RNA-seq on bigger ciliate cells (50–100 µm) without hard cell walls [21, 22]; however, achieved only low gene recovery and high stochasticity of gene expression on small cells with cell walls (a haptophyte—8 µm and a dinoflagellate—15 µm) [23]. Another technology, the 10X Genomics Chromium platform widely used in biomedical research, showed results comparable to bulk RNA-seq on the photosynthetic unicellular algae *Chlamydomonas* with no cell wall [24], indicating that cell size and cell wall are determining factors in application of scRNA-seq protocols to microbial eukaryotes.

In this study we applied MASC-seq (massive and parallel microarray sequencing) [25], a scRNA-seq method that enables sequencing of hundreds of cells at the same time, and that links microscope images with the single cell transcriptome data. For this method, fixed cells are spread over a glass slide that contains array with 100 µm-sized spots of DNA capture probes with spot-specific indices. The cells are first imaged using a scanning microscope and then permeabilized, releasing their RNA out of the cells which binds to the probes on the array. cDNA is synthesized, harvested and sequenced, and, using the spot-specific barcode-sequences, cDNA sequences stemming from a specific spot (i.e., cell) can be linked to the microscope image of the corresponding cell. However, until now, the MASC-seq method has only been applied to mammalian cells.

The aim of this study was to test and adapt the MASC-seq method for application on unicellular eukaryotic plankton. We applied and optimized the method on three cultured plankton representatives, abundant in communities of aquatic environments and with different cell sizes and surface structures; *Phaeodactylum tricornutum* (a diatom, silica and polysaccharide cell walls [26]), *Heterocapsa* sp. (a dinoflagellate, cellulose thecal plates [27]) and *Tetrahymena thermophila* (a ciliate, lipid membrane [28]). We optimized several steps in the protocol to

make it more suitable for planktonic cells and compared the results from MASC-seq generated single cell transcriptomes to bulk RNA sequencing.

Results

Adapting the MASC-seq method for eukaryotic single-celled plankton

In order to adapt the MASC-seq protocol to eukaryotic plankton, we focused on three species, each representing an important group of plankton: *P. tricornutum* (a diatom), *Heterocapsa sp.* (a dinoflagellate) and *T. thermophila* (a ciliate). The three species were cultured individually and cells were harvested at the exponential growth phase. The following steps of the MASC-seq protocol were optimized using these cells: cell fixation, cell permeabilization and cell attachment to the specialized microarray slide. These tests were performed on optimization slides and library preparation slides (Fig 1B and 1D) and based on the results we recommend the following adjustments of the existing protocol [25].

Cell fixation

We tested how two different cell fixatives (formaldehyde and methanol) affect cDNA synthesis and final library quality. Here, we used cells of *P. tricornotum* and *Heterocapsa sp.* as models. The two methods showed comparable fluorescent cDNA signals on optimization slides, but, as expected, the Bioanalyzer traces of final sequencing libraries showed that methanol gave longer final library products (Fig 1E and S1 Fig). Consequently, we used methanol as the fixative of choice when preparing the libraries for sequencing. Related to the fixation step, we additionally compared prefixing the cells during the collection (using methanol as a fixative) with keeping them unfixed and then, during the first stages of the protocol, fixing them on the slide (as done by Vickovic et al. [25]). Our results indicated that cell morphology was improved when cells were prefixed during the collection, since the staining (see below) was more even and cells retained their shape better (Fig 1E and S2 Fig).

Cell permeabilization

Vickovic et al. [25] used pepsin for 30s when permeabilizing cells of the human adenocarcinoma cell line. In addition to pepsin, we tested lysozyme and 0.01 M HCL (S1 Fig) which are both commonly used in catalyzed reporter deposition fluorescent *in situ* hybridization (CARD-FISH) assays for permeabilizing eukaryotic plankton [30]. However, of the three tested permeabilization methods, only pepsin rendered positive fluorescent cDNA signals on the optimization slides (Fig 1E). Cells of *P. tricornutum* and *Heterocapsa sp.* were permeabilized with pepsin for 5 min and *T. thermophila* for 1 min due to the difference in cell structures. *T. thermophila* has phospholipid membrane, similar to mammalian cells meaning shorter permeabilization time compared to *P. tricornutum* and *Heterocapsa sp.*, which both have hard cell walls.

We further compared solely using pepsin or combining it with freeze-thaw cycles in liquid nitrogen and room temperature water (details in Methods section). Freeze-thaw cycles increased concentrations of final libraries (Fig 1E and S3 Fig) for *P. tricornutum* and *Heterocapsa sp.*, but resulted in disintegration of the less robust *T. thermophila* cells (S4 Fig).

Cell attachment

After the cells have been imaged, they undergo different incubations and washing steps (see Methods section), which may lead to cell loss before their mRNA has been captured by the surface probes. To estimate the extent of cell loss, we used *Heterocapsa sp.* as a model and
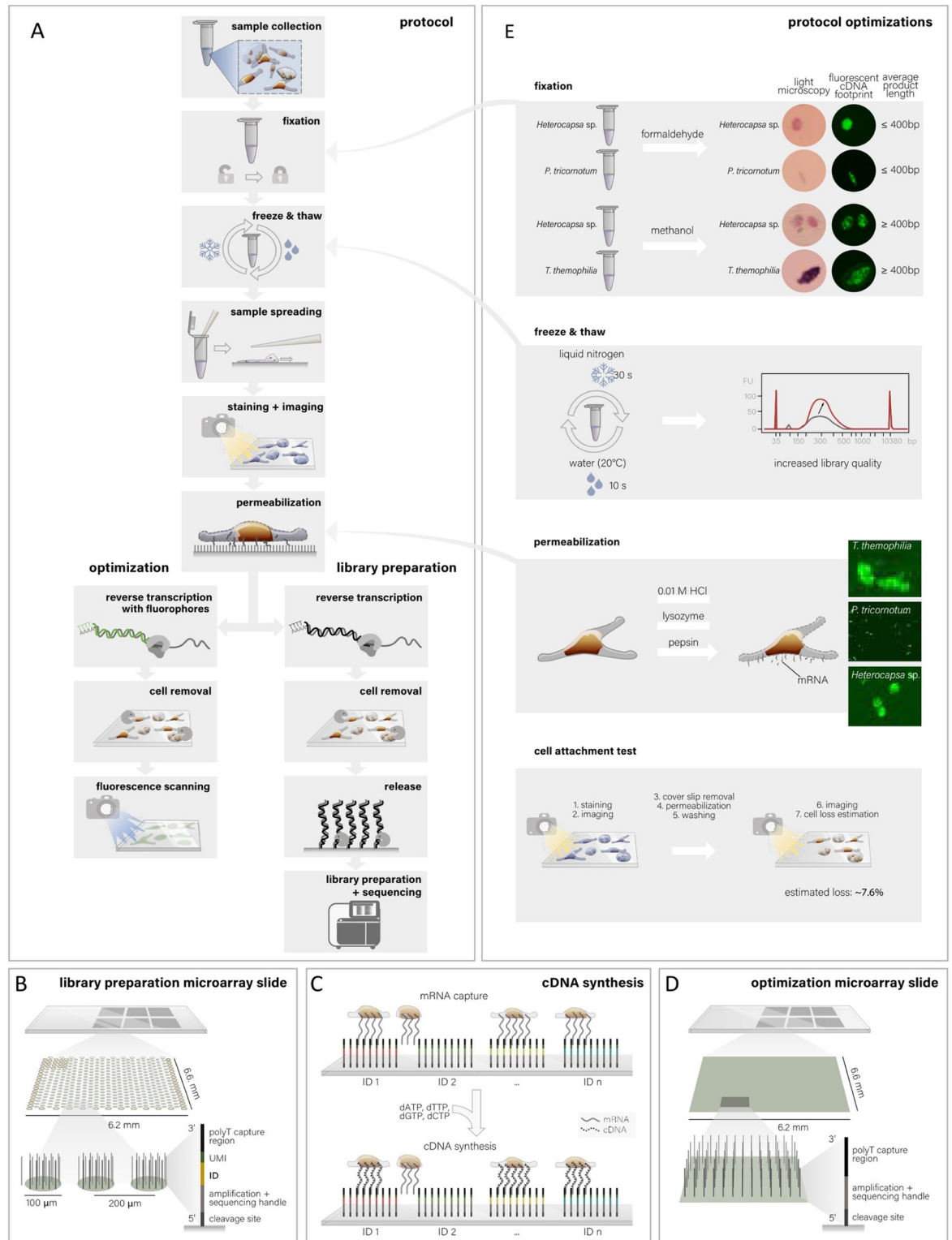
Fig 1. Overview of the MASC-seq method. A) Samples are collected as described in the Methods section and fixed. Eukaryotic plankton that have hard cell wall covering undergo freeze-thaw cycling in liquid nitrogen to pre-permeabilize cell walls. Subsequent steps are attachment to the slide, staining, imaging and permeabilization. After permeabilization, polyadenylated transcripts leaking out of the cells are captured by the surface probes, which function as primers in the following reverse transcription reaction. Cells are removed, cDNA is enzymatically cleaved off, harvested, amplified and processed for Illumina sequencing (path to the right in panel A). **C**) The slides have six

separate arrays that can accommodate different experiments. Each array contains ~2000 spots with oligo-dT probes that have spot-specific barcodes (IDs) and UMIs. **D**) During cDNA synthesis, a cell positioned on top of a certain spot will get the spot-specific ID incorporated into its cDNA that can later be linked to the image of the spot. For optimization purposes, a special type of optimization arrays with oligo-dT probes covering the whole surface (**B**) can be used. Here, the reverse transcription is conducted using fluorescently labeled nucleotides, and after cell removal, the slide is scanned and the fluorescence cDNA footprint will correspond to the amount of cDNA synthesized (path to the left in panel A). Library preparation and sequencing are not conducted in this case. **E**) Optimizations conducted to make the MASC-seq protocol more suitable for single-celled eukaryotic plankton. From top to bottom: tested fixatives, formaldehyde (2%) and methanol (absolute) both gave positive results on optimization array slide, methanol produced longer final library molecules. Two different enzymes (pepsin and lysozyme) and acid (0.01 M HCL) were tested for cell permeabilization and only pepsin gave positive results. Freeze-thaw cycling in liquid nitrogen and room temperature (3 cycles) improved the amount of product in the final libraries. A cell attachment test was performed to estimate cell loss during the different steps of the protocol. Fig illustrations in A, B and D are redrawn with permission from [29], Springer Nature Limited.

calculated that around 7% of the cells fall off during the different stages of the protocol (Fig 1E). This is an important aspect to consider when comparing hematoxylin and eosin (H&E) images and images of fluorescent cDNA synthesis.

MASC-seq applied on individual plankton species

To evaluate quality and quantity of data produced with the MASC-seq method we sequenced four libraries, two for *T. thermophila* (Tet1 and Tet2), one for *Heterocapsa* sp. (Het) and one for *P. tricornotum* (Pha), using the optimized protocol. By manual inspection of the microscopic images of the arrays, we subdivided spots on each array into the following four groups: single cells (1 cell per spot), pairs of cells (2 cells), clusters of cells (>2 cells), and background (0 cells) (Fig 2). There was a general pattern that spots with more cells displayed higher expression (more expressed genes [i.e., expressed reference transcripts, see Methods section] and more unique transcripts) than those with fewer cells (Fig 2). Background spots had significantly lower expression than single cell spots in all libraries except *P. tricornotum*, and single-cell spots had significantly lower expression than double-cell spots in all libraries except *Heterocapsa* sp. (Fig 2). By comparing the number of uniquely mapped reads per spot for single-cell spots and background spots, we estimated the precision in the readings to range from 0.17 for *P. tricornotum* to 0.73 for *T. thermophila* Tet2 (see S2 Table).

We compared the gene expression obtained from the cells on the array with that of total RNA extracted from the bulk cultures and pipetted onto the array (Fig 3). Bulk samples for RNA were collected at the same time as cells for MASC-seq, and the amount of RNA was adjusted to correspond to the expected total amount of RNA in the ca. 2000 cells in the MASC-seq experiments (see below). While determining the true in vivo expression levels of the cells in the cultures remains elusive, we can assume that the RNA extracted from the cultures closely approximates real expression. The total expression (number of transcripts) per gene was significantly correlated between cells and RNA for all three species, but the linear correlations were stronger for the two *T. thermophila* replicates and *Heterocapsa* sp. (Pearson $r = 0.64–0.73$) than for *P. tricornotum* ($r = 0.43$), likely due to the much larger number of transcripts detected in the former. The correlations were also lower than between the two replicated MASC-seq experiments in *T. thermophila* ($r = 0.88$) (Fig 3), which may indicate that the cell permeabilization, cDNA synthesis and library preparation in this protocol probably give rise to some transcript-specific biases not present in the more complete lysis of cells in the bulk RNA preparations.

The much lower number of expressed genes and transcripts per cell in *P. tricornotum* (on average 4.9 genes and 8.0 transcripts for single-cell spots) compared to *Heterocapsa* sp. (305 genes and 338 transcripts) and *T. thermophila* (148 and 194 genes, and 171 and 254 transcripts, for the two libraries) may partly be explained by the smaller size of *P. tricornotum* and the
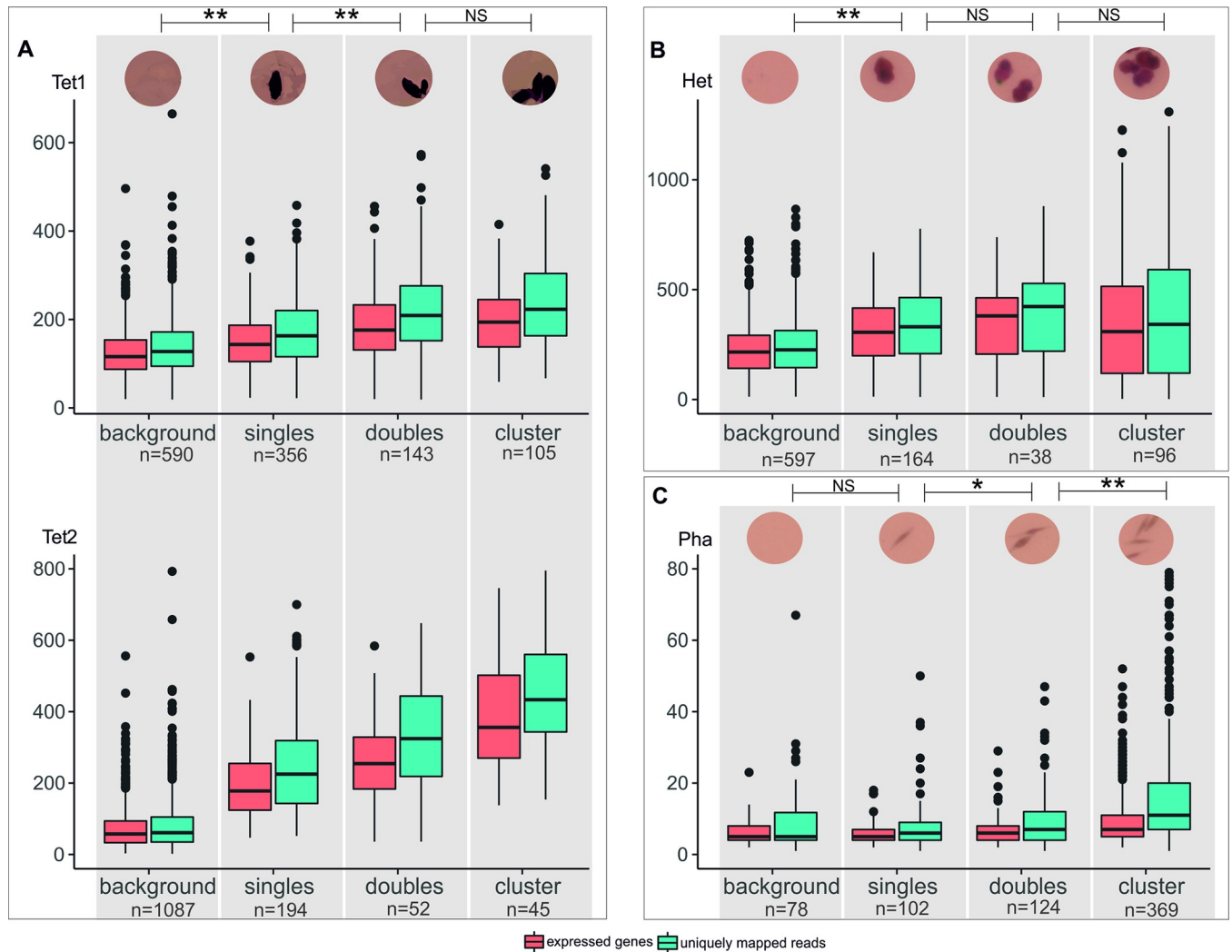
Fig 2. Number of expressed genes and unique transcripts per spot for different spot categories. Background (empty spots), singles (one cell), doubles (two cells) and clusters (>2 cells) are identified based on the microscopic images and MASC-seq sequencing libraries of **A**) *T. thermophila* (two libraries; Tet1 and Tet2), **B**) *Heterocapsa* sp. (Het) and **C**) *P. tricornutum* (Pha). An example image of a spot is shown above each spot category and species. The number of spots (n) for each category and library are given. Significance levels of Wilcoxon rank-sum tests comparing number of unique transcripts per spot between pairs of spot categories are indicated, with * = p-value < 0.05; ** = p-value < 0.0001; and NS = non-significant.

https://doi.org/10.1371/journal.pone.0296672.g002

expected lower number of mRNA molecules inside each cell. *P. tricornutum* cells have a volume of 60–170 μm³ [31] as compared to 1,000–4,000 μm³ and 15,000 μm³ for *Heterocapsa* sp. [31] and *T. thermophila* [32], respectively. In line with this, based on the amount of RNA that we could harvest from cultures of each cell type, we estimated that *P. tricornutum* had 1,250–6,250 mRNA molecules per cell compared to *Heterocapsa* sp. with 27,500–137,500 and *T. thermophila* with 46,490–232,450 molecules [23] (Table 1). However, cell size cannot be the only explanation behind the variation in detected transcripts per cell, as illustrated by the variation between the two *T. thermophila* libraries. Other factors such as optimal cell fixation and cell permeabilization conditions, optimal amplification of library molecules, and sequencing depth are likely important, although much larger variability in scRNA transcripts was observed in protist cells compared to mammalian cells [23].

In order to test the effect of sequencing depth on the number of detected transcripts per cell, we subsampled the datasets to different levels (Fig 4). This showed that a plateau in the
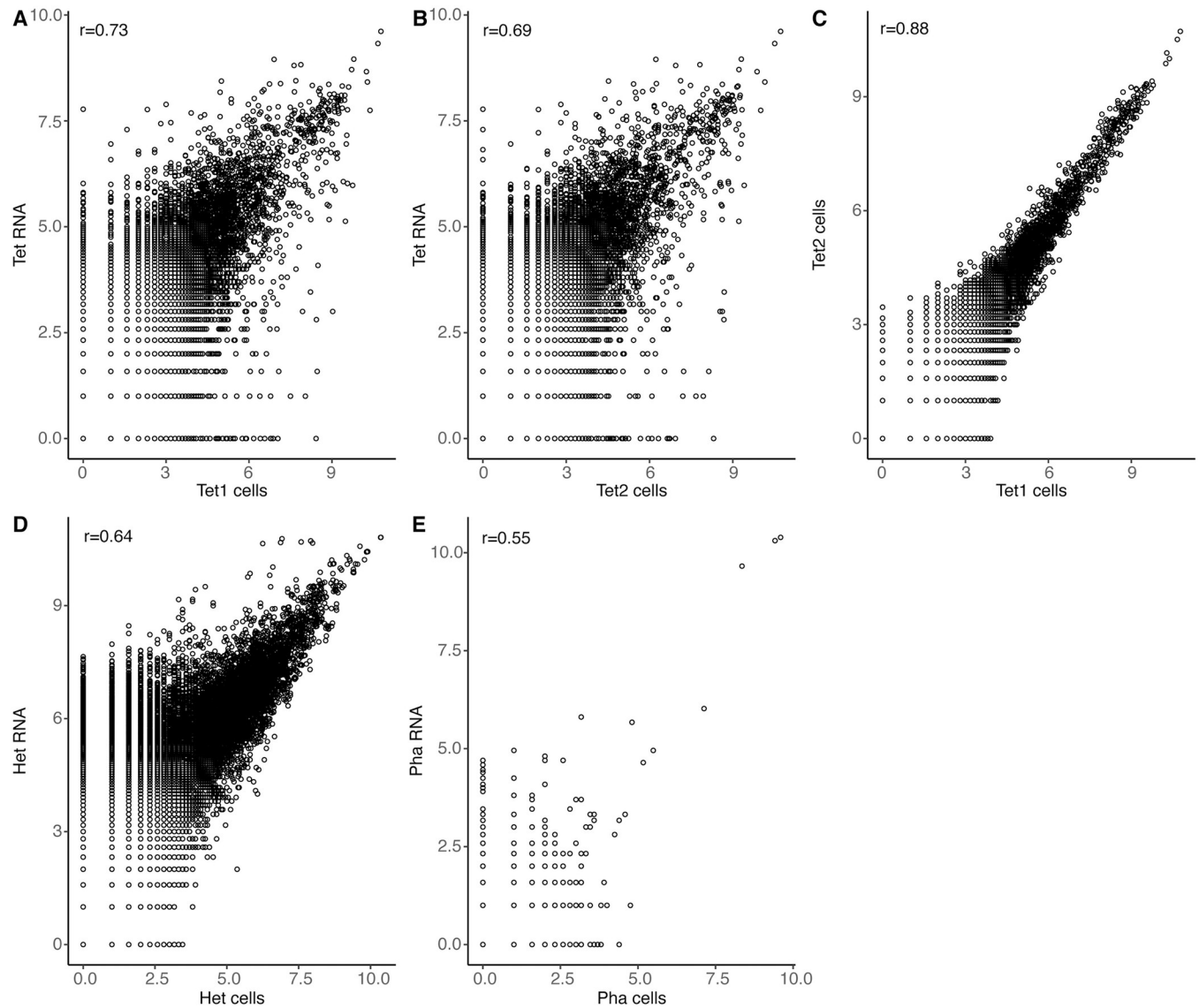
Fig 3. Scatter plots illustrating correlations between sequencing libraries. Each circle represents one gene, and its position indicates its total counts (number of unique transcripts) in the two libraries compared. **A-C)** *T. thermophila*. Bulk RNA added to the array vs. MASC-seq data (cells) for Tet1 (A) and Tet2 (B). MASC-seq data for Tet1 vs Tet2 (C). **D)** Bulk RNA added to the array vs. MASC-seq data for *Heterocapsa* sp. **E)** Bulk RNA added to the array vs. MASC-seq data for *P. tricornutum*. Pearson correlation coefficients (r) are given for each subplot.

https://doi.org/10.1371/journal.pone.0296672.g003

Table 1. Cell sizes and estimated number of total RNA and mRNA molecules per cell.

	Cell biovolume (μm^3)	Total RNA molecules per cell[a]	mRNA molecules per cell[b]
P. tricornutum	60–170 μm^3	125 000	1250–6290
Heterocapsa sp.	1000–4000 μm^3	2 750 000	27 500–137 500
T. thermophila	15000 μm^3	4 649 000	46 490–232 450

The number of RNA molecules calculated from the amount of total RNA extracted and the number of cells used for the RNA extractions.

[a]Assuming RNA extraction efficiency was 50% and an average mRNA transcript length of 1000 nt

[b]Assuming 1–5% of total RNA was mRNA. (according to Liu et al. 2017)

https://doi.org/10.1371/journal.pone.0296672.t001

Fig 4. Saturation curves for each library. *T. thermophila* (Tet1; A, B), *Heterocapsa* sp. (Het; C, D) and *P. tricornutum* (Pha, E, F) illustr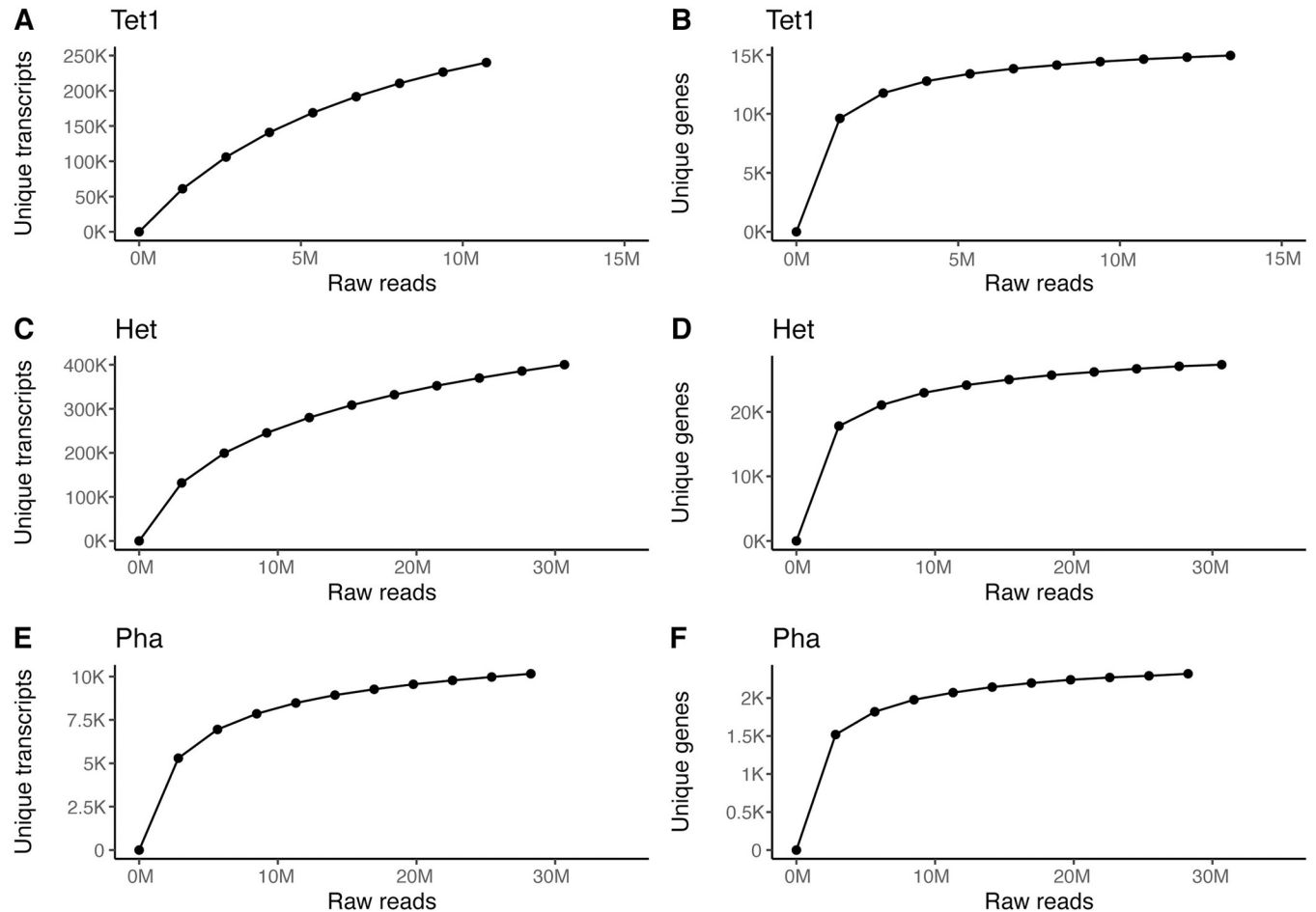ating how many unique transcripts (A, C, E) and genes (B, D, F) that were detected as a function of sequencing depth. The sequencing depth (number of read pairs) are plotted on the x-axes and the unique transcripts or unique gene counts on the y-axes.

detected number of transcripts and genes was reached before the full library had been analyzed for *P. tricornutum* (Fig 4E and 4F) while for *Heterocapsa* sp., and even more so for *T. thermophila*, the number of transcripts was more steadily increasing with sequencing depth (Fig 4A and 4C). This indicates that more sequencing would not have rendered substantially more data for *P. tricornotum* and that rather the number of unique molecules in the final library was the limitation.

We investigated the ten most highly expressed genes, in spots that contained single cells, in more detail. In both the Tet1 and Tet2 libraries, genes for transmembrane proteins, papain family cysteine proteases and ribosomal proteins were the most highly expressed (Table 2). In the Het library, the most highly expressed genes encoded photosynthetic proteins peridinin-chl a protein, chloroplast proteins and major basic nuclear protein. We did not perform this analysis for the Pha library since gene counts in this case were very low.

Discussion

To the best of our knowledge, this study is the first to apply high-throughput parallel imaging and single-cell transcriptomics to microbial eukaryotic plankton. A major advantage of the

Table 2. Top 10 most highly expressed genes (*i.e.* reference transcripts) in single-cell libraries of *Heterocapsa* sp. (Het) and *T. thermophila* (Tet1 and Tet2).

Het singles	reference transcript ID	protein description	Organisms	gene accession ID
	TrINITY_DN4782_c1_g1_i13	peridinin-chl a protein	Heterocapsa pygmaea; NCBI:txid35672	CAC19481.1
	TRINITY_DN4782_c1_g1_i7	unnamed protein product	Heterocapsa pygmaea; NCBI:txid35672	CAE8622356.1
	TRINITY_DN2086_c0_g1_i21	major basic nuclear protein	Polarella glacialis; NCBI:txid89957	AAL61531.1
	TRINITY_DN2086_c0_g1_i8	major basic nuclear protein	Polarella glacialis; NCBI:txid89957	AAL61531.1
	TRINITY_DN2086_c0_g1_i32	major basic nuclear protein	Polarella glacialis; NCBI:txid89957	AAL61531.1
	TRINITY_DN2957_c0_g2_i4	chloroplast ATPH isoform 5	Heterocapsa triquetra; NCBI:txid66468	AAW80675.1
	TRINITY_DN4782_c1_g1_i4	peridinin-chl a protein	Heterocapsa pygmaea; NCBI:txid35672	CAC19481.1
	TRINITY_DN2086_c0_g1_i18	major basic nuclear protein	Pfiesteria piscicida; NCBI:txid71001	AAL61531.1
	TRINITY_DN2957_c0_g2_i3	chloroplast ATPH isoform 5	Heterocapsa triquetra; NCBI:txid66468	AAW80675.1
	TRINITY_DN2086_c0_g1_i14	major basic nuclear protein	Pfiesteria piscicida; NCBI:txid71001	AAL61531.1
Tet1 singles	**reference transcript ID**	**protein description**		**gene accession ID**
	g13052.t1	transmembrane protein, putative	Tetrahymena thermophila SB210; NCBI: txid312017	XP_001032220.2
	g3965.t1	papain family cysteine protease	Tetrahymena thermophila SB210; NCBI: txid312018	XP_001027342.1
	g1442.t1	hypothetical protein TTHERM_001080487	Tetrahymena thermophila SB210; NCBI: txid312019	XP_012655449.1
	g7083.t1	transmembrane protein, putative	Tetrahymena thermophila SB210; NCBI: txid312020	XP_012654551.1
	g3276.t1	hypothetical protein TTHERM_000760291	Tetrahymena thermophila SB210; NCBI: txid312021	XM_012800270.1
	g25171.t1	hypothetical protein TTHERM_01125120	Tetrahymena thermophila SB210; NCBI: txid312022	XP_001030140.1
	g16201.t1	papain family cysteine protease	Tetrahymena thermophila SB210; NCBI: txid312023	XP_012653036.1
	g10027.t1	60S ribosomal protein L31, putative	Tetrahymena thermophila SB210; NCBI: txid312024	XP_001031042.2
	g162.t1	ubiquitin-40S ribosomal protein S27a	Tetrahymena thermophila SB210; NCBI: txid312025	XP_001025217.1
	g16902.t1	40S ribosomal protein S23	Tetrahymena thermophila SB210; NCBI: txid312026	XP_001011429.2
Tet2 singles	**reference transcript ID**	**protein description**		**gene accession ID**
	g13052.t1	transmembrane protein, putative	Tetrahymena thermophila SB210; NCBI: txid312022	XP_001032220.2
	g3965.t1	papain family cysteine protease	Tetrahymena thermophila SB210; NCBI: txid312023	XP_001027342.1
	g1442.t1	hypothetical protein TTHERM_001080487	Tetrahymena thermophila SB210; NCBI: txid312024	XP_012655449.1
	g7083.t1	transmembrane protein, putative	Tetrahymena thermophila SB210; NCBI: txid312025	XP_012654551.1
	g3276.t1	hypothetical protein TTHERM_000760291	Tetrahymena thermophila SB210; NCBI: txid312026	XM_012800270.1
	g16902.t1	40S ribosomal protein S23	Tetrahymena thermophila SB210; NCBI: txid312027	XP_001011429.2
	g14283.t1	60S ribosomal protein L36a	Tetrahymena thermophila SB210; NCBI: txid312028	XP_001010053.1
	g25171.t1	hypothetical protein TTHERM_01125120	Tetrahymena thermophila SB210; NCBI: txid312029	XP_001030140.1

(*Continued*)

Table 2. (Continued)

Het singles	reference transcript ID	protein description	Organisms	gene accession ID
	g3569.t1	Crystal Structure Of The Eukaryotic 40s Ribosomal Subunit In Complex With Initiation Factor 1.	Tetrahymena thermophila SB210; NCBI: txid312030	2XZM_Z
	g10027.t1	60S ribosomal protein L31, putative	Tetrahymena thermophila SB210; NCBI: txid312031	XP_001031042.2

Annotations were deduced by BLASTx searches against NCBI nr; reference transcript ID, protein description, organism and accession number of the best matches are given.

MASC-seq method is the generation of images of the cells for which RNA is sequenced. This has great potential when studying uncultured microbial eukaryotes whose morphology and ecological roles are largely unknown. A microscope image of a cell can reveal information that cannot be interfered from genetic data alone, such as morphology, pigmentation, symbiotic relationships [33]. The value of this combined approach is emphasized by two recent low-throughput studies [34, 35] which imaged cells prior to amplifying transcriptomes, unveiling distinct morphological characteristics in the Apicomplexa phylum that would have been completely lost without the imaging component. By running the method on natural samples it has the potential to provide morphological data to microbial "dark matter" [8, 36] for which only DNA sequence data exist to date.

Most single-cell transcriptomics protocols have been developed for and benchmarked on human and other mammalian cells but few have been tested on protists [16]. Planktonic protists are very diverse in terms of cell structure and thorough studies are needed to gain insights in a methods' performance on different protist types. Therefore, our study represents a valuable contribution to the field.

Challenges and optimization strategies in the MASC-seq method for single-cell analysis of microbial eukaryotic plankton

The MASC-seq method worked fairly well on the ciliate (*T. thermophila*), and dinoflagellate (*Heterocapsa* sp.) species, with median values of 148–305 genes and 171–338 transcripts detected per cell for single-cell spots, and significantly more data for single-cell than background spots. Moreover, total expression levels displayed moderately high correlations between MASC-seq from cells and from bulk RNA (r = 0.55–0.73, Fig 3). The diatom (*P. tricornutum*), however, gave much lower counts (5/8 genes/transcripts per cell) and expression for single-cell spots was not significantly different from the background spots. This is probably related to the diatom's much smaller cell volume and lower number of mRNA molecules per cell [23]; the ratio between estimated number of mRNA molecules per cell and number of detected transcripts for single-cell spots was actually in the same order of magnitude for the different species and libraries (156, 81, 271 and 183 for Pha, Het, Tet1 and Tet2, respectively). The reason why not more of the mRNA molecules of the cells (of all three species) were recovered is likely a combination of RNA degradation, incomplete diffusion out of the cells, incomplete binding to the underlying probes and cDNA synthesis, uneven amplification of the library, and insufficient sequencing depth. Which of these factors are more important is not clear, but our analysis indicated that sequencing depth was not the major limitation.

Although empty spots showed lower gene/transcript counts than spots with cells, the presence of background signal is still problematic since it indicates that cells leak RNA that diffuse

to other spots, resulting in transcripts in some instances being assigned to the wrong cells. This issue will be particularly problematic when the method is run on microbial communities with species of varying sizes, since background signals from the larger cells will mask the foreground signal of the smaller. This can to some extent be alleviated by pre-selecting cells within a certain size range, for example by serial filtration. It is not clear to what extent the background signal comes from RNA already present in the suspension when the cells are smeared on the array, or are derived from diffusion from cells on the array after permeabilization with pepsin. The freeze-thaw cycle that was added before the cells were smeared could potentially contribute to RNA in the suspension, however we observe background signals also for the *Tetrahymena* libraries where this was not applied. Interestingly, the Tet2 library shows a much higher signal-to-background ratio than the Tet1 library, indicating that it is possible to further optimize conditions to lower the problem of background signal.

Another challenge of the protocol is obtaining an even distribution of cells over the array surface. Smearing of a cell suspension is a fast procedure but it can result in clumping of cells in different regions of the array. And even when cells are uniformly distributed, only a fraction of the spots will harbor a single cell, and also these may have other cells directly outside in the surrounding array area. In the original MASC-seq publication [25], an alternative approach of precision fluorescence-activated cell sorting (FACS) of cells into spots was also applied, which results in a larger fraction of single-cell spots and less background noise. However, this requires a specialized instrument and optimizations for different cell types, and most protists cannot be easily sorted by FACS. Another alternative to smearing could be to add the cells to the array in a larger suspension volume and performing a gentle swing bucket centrifugation of the slide as described in Colin et al. [37]. Employing this approach could improve the uniform distribution of cells across the array, potentially improving cell attachment. This would, in turn, reduce cell loss during the various washing steps. Furthermore, capture spot resolution and technologies continue to improve [38, 39] reaching a resolution smaller than the diameter of a typical cell (0.5–1 μm), thereby increasing the likelihood that only one cell will overlap a spot.

State of the art and comparation to other studies in the field

In the current landscape of single cell transcriptomics technologies and protocols, a limited number has been applied exclusively to aquatic single cell eukaryotes. Among related studies, most have employed SMART-seq2 [40, 41] or its modified versions [14, 42], which are considered the gold standard for aquatic protists. This method involves manual isolation of cells or fluorescent cell sorting, making it more time-consuming compared to the MASC-seq method, especially when imaging is performed prior to isolation. However, despite its time-consuming nature, the majority of studies conducted in this field have utilized the SMART-seq2 method, for exploring certain genes in the context of microbial eukaryotes' phylogeny and evolution [41–43] and for gene discovery with method benchmarking [14, 21, 23].

One of the earliest studies, conducted by Kolisko et al. [21], applied this method to ciliates, including T. thermophila, which is also utilized in our current study. Nevertheless, it is challenging to directly compare the results of our study with theirs, as the chemistry of library preparation, protocols, and data analysis differ significantly. Furthermore, while our study generated transcriptomes from 356 and 194 single cells in the Tet1 and Tet2 libraries, respectively, their study focused on transcriptome generation from a one single cell only.

Another study, by Liu et al. [23], that similarly to ours used representatives of marine single celled plankton, more specifically the dinoflagellate *Karlodium veneficum* and the haptophyte *Prymnesium parvum* detected very low transcript recovery and attributed it to small cell size,

which is comparable to our study but only for Phae library while others did not show such a low recovery.

Similar to our study, recent applications of the SMART-seq2 protocol on protists also tested the protocol with different adjustments. However, the protocol was tested on *Giardia intestinalis* [14], a human parasite which has a different membrane and naturally grows in very different environment compared to aquatic protists. Nevertheless, the gene recovery (average 4524 to 4992) was much higher compared to our study which could be attributed to either technological or biological differences.

Recent research has utilized microfluidics, to enable rapid, high-throughput cell sorting. One study utilized the 10x Genomics Chromium system for the analysis of the marine photosynthetic algae, Chlamydomonas [24]. This technology facilitated the sequencing of an impressive 28,690 cells from three samples, detecting an average of 823 genes per cell. By comparison, in our Tet1 and Tet2 libraries, we sequenced 550 single cells combined, identifying an average of 162 genes.

In another groundbreaking study, the emulsion droplet-based method mDrop-seq was used to examine the yeast species *Candida albicans* and *Saccharomyces cerevisiae* [44]. This method demonstrated a considerably higher throughput compared to the presented MASC-seq method used in this study, with a capacity to process between 5,000 and 10,000 cells per sample and an average gene detection that ranged from 480 to 587 per cell. However, its application was limited to well studied model organisms and its effectiveness on aquatic microbial eukaryotes remains unexplored. These findings underscore the potential of microfluidic approaches, paving the way for their application in studying aquatic microbial eukaryotes.

In conclusion, the MASC-seq method, with its unique ability to image cells before transcriptomics, stands as a powerful tool in the study of microbial eukaryotic plankton. Despite challenges and the need for further optimization, its adaptability makes it an important asset to current methods, holding the potential to illuminate unknown aspects of microbial 'dark matter' and advance our understanding of this vital component of marine food webs.

Methods

Microarray slides

In this study two types of slides were used, optimization slides and library preparation slides; both slides were prepared on CodeLink-activated microscope glass slides (SurModics) as described by Ståhl et al. [45]. Optimization slides (Fig 1B) contained six array surfaces with uniform probes. The reverse transcription (RT) probes contained poly-T20VN oligonucleotides and were immobilized by printing on the array surface. Optimization slides are used for optimizing the key steps of the protocol: cell attachment, permeabilization, cell removal by analyzing the fluorescent cDNA 'footprint' that remained after successful RNA capture and removal of the cell (Fig 1A).

Library preparation slides (Fig 1C) were printed in six identical 6.3×6.7 mm^2 capture arrays. Each array includes 66 frame spots for orientation purpose and 1934 barcoded spots for capturing gene expression information. The spots have a diameter of ~ 100 μm and are arranged in a centered rectangular lattice pattern: each spot has four surrounding spots with a center-to-center distance of 150 μm, forming a 310 μm x 325 μm rectangle [29, 46]. In each spot the capture probes share a barcode unique to that spot. During reverse transcription, the barcode is incorporated into the cDNA of all mRNAs captured on the spot (Fig 1D). This enables linking RNA-seq data to a spot on the array, and thereby the location of the cell from which the RNA originated. Each probe covalently immobilized on the glass surface has the following common 5′–3′ structure: 18-mer spot-unique positional barcode, 9-mer semi-

randomized UMI [47] and poly-20TVN capture region [48]. UMIs were utilized to identify unique transcripts per gene.

Cell cultivation

We chose three common plankton types, a diatom (*P.tricornutum* CCAP 1055/1), a dinoflagellate (*Heterocapsa sp.* CCAP 1125/4) and a ciliate (*Tetrahymena thermophila* 630/1M). Both *P. tricornutum* and *Heterocapsa* were purchased from the Culture Collection of Algae and Protozoa (CCAP) from the Scottish Association for Marine Science (SAMS). Cells were grown and kept in L medium (*Heterocapsa sp.*), f/2 medium amended with silica (*P. tricornutum*) and PPY medium (*T. thermophila*). For each cell type we estimated growth curves based on microscopic counts. Cells were harvested during the exponential phase and concentrated by centrifugation at 8000g for 5 min.

RNA extraction from cell cultures

For each cell type we collected samples for bulk RNA. Cells were harvested on 5 μm polycarbonate filters (Whatman). Filters were placed in cryoprotectant tubes with beads (Zirconium oxide beads, Precellys), amended with 600 μl of RLT buffer (Qiagen) which contained 1% of 2-Mercaptoethanol, shock frozen in liquid nitrogen and stored at -80°C until further processing. Prior to RNA extraction, samples were taken out of -80, thawed on ice and subjected to bead beating (1x 6000 rmp, Precellys Evolution homogenizer). *P. tricornutum* and *Heterocapsa* sp. samples were treated for 2 min and *T. thermophila* just vortexed vigorously. RNA from samples was further extracted with the RNeasy Qiagen kit (QIAGEN, Inc., Valencia, CA) following the kit instructions. RNA was re-eluted in 30 μl of nuclease free water.

Optimization microarray experiments

We conducted a series of different tests on optimization microarray slides (Fig 1E). The basic workflow of optimization microarray experiment is illustrated in Fig 1A. The individual protocol steps that were optimized include: cell fixation, cell attachment, staining, permeabilization, cell removal and fluorescent cDNA footprint Subsequently, library preparation microarray slides were used.

Cell fixation

Two different fixatives were tested (formaldehyde and methanol) and their impact on fluorescent cDNA synthesis and final library quality evaluated (S1 Fig). For this purpose, cells were chemically treated in three different fixatives during the exponential growth phase: fixation in formaldehyde, methanol and no fixation. For formaldehyde fixation, 2 ml of cell culture aseptically transferred into a microcentrifuge tube and supplemented with formaldehyde (36.5–38% (vol/vol); Sigma-Aldrich) to reach 2% final concentration. Cells were incubated at room temperature for 10 min, centrifuged at 8000g for 5 min, the pellets were washed twice with 1xPBS and then shock frozen in liquid nitrogen. For methanol fixation, 1 ml of cell culture was aseptically transferred in a microcentrifuge tube and 1 ml of ice-cold methanol was added very slowly, in drops. Tubes were placed at -20°C for a minimum of 30 minutes. Cells were washed in the same way as described for formaldehyde fixation. Additionally, unfixed cells were collected which were then fixated directly on the slide. Cells were washed twice with PBS as described above and pellets were shock frozen in liquid nitrogen (Fig 1A).

Freeze-thaw cycle

In order to improve permeabilization of *P. tricornutum* and *Heterocapsa sp*. a freeze-thaw cycle in liquid nitrogen and room temperature water was applied, which has previously been demonstrated to increase the RNA recovery [14]. Cells were collected, fixed and pelleted as described above. The pellet was first frozen in liquid nitrogen for 30 sec and dipped into room temperature water for 10 sec. The cycle was repeated 3 times and pellets stored at -80 until further processing. Freeze-thaw cycle was not used for *T. thermophila* since their cell membrane is not hard to permeabilized with enzymes.

Attachment, staining and imaging

Cells collected as described were placed on ice and diluted to a concentration of ∼1,000 cells μl^{-1} with cold 1xPBS. Before attaching the cells on the array, the whole slide was warmed at 37 ˚C for 1 min. Smearing was performed by taking 3 µl of the cell suspension and slowly pipetting the cells on the center of the array surface without touching the surface with the tip and maintaining a droplet by surface tension. Then, the cell solution was smeared with the side of a pipette tip (again careful not to touch the surface of the array) followed by attachment to the slide at 37 ˚C for 3 min (as detailed in [22]). Cells were afterwards stained with hematoxylin and eosin (H&E). Hematoxylin was added to cover cells attached to the slide (500 µl) and incubated for 4 min. Slide was washed by dipping in the beakers of nuclease-free water until the excess hematoxylin is washed away. Afterwards 500 µl of bluing buffer was added, incubated for 2 min and washed as described in the previous step. Finally, 500 µl of buffered eosine was added to the slide, incubated for 1 min and washed with nuclease free water. The slide that now contained stained cells was air dried and mounted with 1ml of 85% glycerol (Merck Millipore) and covered with a coverslip (Menzel-Glaser). Detailed staining protocol is described in Salmén et al. [49] Imaging was performed with Metafer Vslide scanning system (MetaSystems, Mannheim, Germany) installed on an Axio Imager Z2 LSM700 microscope (Carl Zeiss, Oberkochen, Germany). All images were taken with the 20 x Plan-Apochromat objective lens, and stitched using the VSlide software (MetaSystems Hard & Software GmbH, Altlussheim, Germany). After imaging, the coverslip was removed from the glass slide by dipping in nuclease free water and subsequently in 80% ethanol to remove glycerol.

Permeabilization, fluorescent cDNA synthesis and cell removal

Glass slides with arrays were put in a hybridization cassette (ArrayIT Corporation) with a rubber mask that separates each slide into six different subarrays to secure reaction chambers for each array and perform on-array reactions. Cells attached to each array were washed with 100 µl of 0.1 x saline-sodium citrate buffer (SCC) (Sigma-Aldrich) diluted with nuclease free water. Aliquots of 0.1% pepsin (Sigma-Aldrich) dissolved in 0.1 M HCl (Sigma-Aldrich) were pre-warmed at 37 ˚C for 5 min and then 70 µl of the solution was added into each array in order to permeabilize the cells. Cells of *P. tricornutum* and *Heterocapsa sp*. were permeabilized for 5 min and *T. thermophila* for 1 min. Pepsin was washed away with 100 µl of 0.1 x SCC buffer and 70 µl of reverse transcription (RT) mixture for fluorescent cDNA synthesis was added. The RT mixture contained: 1× First Strand Buffer (Invitrogen), 5 mM DTT (Invitrogen), 0.5 mM of each dNTP (Fisher Scientific) except dCTP which was 12.5 µM, and with additional inclusion of 25 µM Cy3-labeled dCTPs (PerkinElmer, Waltham, MA) in a 0.19 µg μl^{-1} BSA, 50 ng μl^{-1} actinomycin D (Sigma-Aldrich), 1% DMSO (Sigma-Aldrich), 20 U μl^{-1} Superscript III (Invitrogen) and 2 U μl^{-1} RNaseOUT (Invitrogen). The mixture was incubated overnight at 42˚C.

Each array was washed as described above and 70 μl of proteinase K (Qiagen) in PKD buffer (Qiagen) was added and the slide incubated at 56˚C for 1h and 30min with 300 rpm interval shake. After cell removal, slides were washed in 2xSCC which contained 0.1% SDS (Sigma Aldrich) at 50˚C for 10 min and subsequently washed with 0.2xSCC and 0.1xSCC for 1 min at room temperature. Arrays were mounted with 500 μl SlowFade Gold Antifade reagent (Invitrogen). Fluorescent footprints of the cell's cDNA were imaged as described in the section "*Attachment, staining and imaging*", only with fluorescent settings of the microscope (Cy3 filters). If the experiment was successful, fluorescence could be observed (Fig 1E).

Attachment test

In order to test if the cells attach on the slide without falling off, additional tests were performed (Fig 1E). Cells were attached, stained, imaged, washed, permeabilized as described above and imaged again in order to inspect for movement or cell loss. Two images were compared and analyzed with imageJ [50].

Permeabilization

Three different types of permeabilization methods and their outcome on fluorescent cDNA synthesis were tested, including pepsin [25], lysozyme, and 0.01M HCL (Fig 1E and S1 Table). By inspecting fluorescent footprint of cDNA, it was estimated that 5min of pepsin worked the best for *P. tricornutum* and *Heterocapsa sp*. while 1 min for *T. thermophila*.

Cell removal

Removing cells from the array of the slide is important in order to avoid false positive cDNA fluorescence footprint and contaminations during library preparation. To test cell removal and optimal concentration of proteinase K, a simple experiment was performed. Cells were attached, stained and imaged as described above and incubated with three concentrations of proteinase K in the PKD buffer (1:4, 1:5, and 1:6). The slides were imaged again to inspect if cell removal was successful (i.e., if the cells were not visible at the slide). Concentration of 1:4 and 1:5 successfully removed cells while 1:6 concentration did not remove all of the cells (results not shown).

Library preparation

Once the protocol was optimized on optimization microarray slides, we used library preparation slides to produce libraries for sequencing. One library from each cell type (*P. tricornutum* (Pha), *Heterocapsa* sp. (Het)) and two libraries for *T. thermophila* (Tet1 and Tet2) were prepared for sequencing. Additionally, extracted RNA from each cell type was smeared on the slide array and prepared together with the cells. Total extracted RNA from corresponding cultures was used as a positive control to single cells and it was estimated how many μl of RNA we need to smear on the slide to have the amount of ~ 2000 cells. During the optimization experiments methanol fixation provided the best outcome on library quality (S1 Fig), therefore this method was used for all subsequent experiments and library preparation.

Attachment, staining and imaging

Cells were attached, stained and imaged as described in the optimization microarray experiment.

Permeabilization, fluorescent cDNA synthesis and cell removal

Permeabilization mixture was prepared as described in the optimization microarray experiment. cDNA mixture for library preparation differed from optimization mixture in the way that it did not contain fluorescently labeled Cy3-dCTP. Mixture contained: 1× First Strand Buffer (Invitrogen), 5 mM DTT (Invitrogen), 0.5 mM of each dNTP (Fisher Scientific), 0.19 µg µl^{-1} BSA, 50 ng µl^{-1} actinomycin D (Sigma-Aldrich), 1% DMSO (Sigma-Aldrich), 20 U µl^{-1} Superscript III (Invitrogen) and 2 U µl^{-1} RNaseOUT (Invitrogen) and cDNA was synthesized over night at 42˚C. After the cDNA was synthesized, cells were removed with proteinase K in the PKD buffer (1:4 ratio) and subsequently washed as described in an optimization microarray experiment.

Probe release. Release of probes with mRNA–cDNA hybrids from the array surface was performed by adding 70 µl of release mix to the array chambers. Release mix was composed of 1× Second Strand Buffer (Invitrogen), 8.75 µM of each dNTP, 0.20 µg µl–1 BSA, and 0.1 U µl–1 USER enzyme (NEB). Incubation was performed at 37˚C for 2 hours with interval shaking and 300 r.p.m. Release mixture was collected and transferred to a 96 well plate and placed at -20˚C until further processing with an automatic robotic station.

Fluorescent imaging of spots

A fluorescence signal on library preparation arrays was obtained by adding 70 µl of reaction mixture containing 0.96× PBS, 0.2 µM Cy3-anti-A probe (Eurofins, [Cy3]AGATCGGAAGAG CGTCGTGT) and 0.2 µM Cy3-anti-frame probe (Eurofins, [Cy3]GGTACAGAAGCGCGATA GCAG). Incubation was performed at room temperature for 10 min. Arrays were washed with 2× SSC containing 0.1% SDS at 50˚C for 10 min, then with 0.2× SSC and 0.1× SSC at room temperature for 1 min, respectively. Subsequently, library preparation arrays were mounted with 500 µl of SlowFade Gold Antifade Reagent (Invitrogen) and covered with a coverslip. Fluorescence imaging, image stitching and extraction were performed following the procedure described in the section of optimization microarray experiment.

Image alignment

Image alignment of corresponding hematoxylin and eosin-stained cells and fluorescence images of spots were performed manually using Adobe Photoshop CC (Adobe).

Automated library processing

Libraries were prepared with Magnatrix 8000+ (Nordiag), an eight-channel robotic workstation capable of running custom made scripts for in-tip magnetic bead separations [51]. Released mRNA–cDNA hybrids were transferred to the robotic workstation where they underwent second strand synthesis, end repair and *in vitro* transcription. Following transcription, sequencing adaptors are ligated to the RNA, and another round of cDNA synthesis is performed. Each step in the process is followed by a reaction clean-up step. Detailed descriptions of every reaction happening in the robotic station are present in [25, 49, 51].

Final libraries amplification, indexing and sequencing

After clean-up, cDNA was amplified and indexed. The Optimum number of PCR cycles for final library amplification was estimated by qPCR in 10 µl of final volume. The qPCR reaction mixture contained 2 µl of purified sample, 1× KAPA HiFi HotStart Readymix (KAPA Biosystems), 0.5 µM PCR InPE1.0 primer (Eurofins, 5′–AATGATACGGCGACCACCGAGATCTACA CTCTTTCCCTACACGACGCT CTTCCGATCT–3′), 0.01 µM PCR InPE2.0 primer (Eurofins,

5′–GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT–3′), 0.5 μM PCR Index primer (Eurofins, 5′–CAAGCAGAAGACGGCATACGAGATXXXXXXGTGACTGGAGTTC–3′), and 1× EVA green (Biotium). The qPCR (Bio-Rad CFX Real-Time system) programme was as follows: 1×, 98˚C 3 min; 25×, 98˚C 20 s, 60˚C 30 s, 72˚C 30 s. The PCR was performed in a reaction volume of 25 μl with the above programme and the established optimal number of cycles per sample. A final extension at 72˚C for 5 min was included. Indexed libraries were purified by an automated MBS robot [52] and eluted in 20 μl of Elution Buffer (Qiagen). The library curves and average length were assessed using the DNA 1,000 kit (Agilent) on a 2,100 Bioanalyzer (Agilent), and the concentration was measured using the Qubit dsDNA HS Assay Kit (Life Technologies) following the manufacturer's instructions. Indexed libraries were diluted to 2 nM and sequenced using the Illumina NextSeq platform applying paired-end sequencing. The forward read (R1) was sequenced for 31 bases, and the reverse (R2) for 121 bases.

MASC-seq data analysis

Raw fastq files were processed through the ST pipeline [51], which involves removal of duplicate reads, homopolymer stretches, and reads with low quality. For the purposes of this study, we modified the original code slightly, making it possible for the user to decide which attribute tag in the GFF file to use when counting reads. The modified code is available at https://github.com/NBISweden/st_pipeline/. The ST pipeline uses the transcript information present in the reverse reads (R2) and the barcode information, unique to a certain spot, present in the forward reads (R1) together with the UMIs to generate a matrix of counts where the genes are represented as columns and the spatial barcodes are represented as rows, and each matrix cell then represents the expression level of a given gene in a given spot. ST pipeline requires two files as an input: indexed genome or transcriptome of the sequenced species and corresponding GFF or GTF file. Genomes of *P. tricornutum* [53] and *T. thermophila* [54] were previously sequenced and downloaded with corresponding GFF files (https://mycocosm.jgi.doe.gov/Phatr2/Phatr2.home.html; http://ciliate.org/index.php/home/downloads). A reference genome for *Heterocapsa* sp. is not available so we performed a de-novo assembly from bulk RNA-seq to obtain the transcriptome as a reference (see below). In order to make the analysis as comparable as possible between the three species, we predicted transcriptomes for *P. tricornutum* and *T. thermophila* using their GFF files and the program gffread (v0.12.1 [55]) with settings to exclude mRNAs that have no CDS features, and ran the ST pipeline in transcriptome mode for all three species. Ribosomal RNA (rRNA) sequences were identified in each reference transcriptome using barrnap (v0.9) with settings '—kingdom euk—reject 0.1', and were subsequently removed. The ST pipeline computes the expression level of a gene (reference transcript) as the sum of R2 reads mapping to the gene, but in order to remove duplicates stemming from amplification of the same starting molecule, it sorts transcripts with the same UMI by strand and start position, then groups them if they are within 250 bp of each other, allowing 1 mismatch in the UMI.

Top 10 most highly abundant genes were extracted in R studio [56] and BlastX was used to obtain gene annotations.

Bulk RNA sequencing of *Heterocapsa* sp.

RNA from *Heterocapsa* sp. Was extracted as described above and sent to SciLifeLab/NGI for library preparation and sequencing (Solna, Sweden). Libraries were prepared with Illumina TruSeq stranded mRNA kit with poly-A selection. Samples were sequenced on NovaSeq6000 with 151nt (Read1) -10nt (Index1)-10nt(Index2)-151nt(Read2) setup.

The bioinformatic analyses conducted on the bulk RNA-seq data were organized into a snakemake [57] workflow available at https://github.com/NBISweden/LTS-A_Andersson_2010-MASCSEQ. The workflow uses snakemake version 6.0.5.

Adapters were trimmed using cutadapt (v3.2) [58] using Illumina TruSeq adapters and minimum read length of 50. For sample P21817_104, the cutadapt maximum error rate was set to 0 in order to properly remove adapters. SortMERNA (v4.2.0) [59] was used to filter reads into rRNA and non_rRNA partitions and was run with default settings and all rRNA databases. Read quality was assessed with fastqc (v0.11.9) [60] and multiqc (v1.10) [61].

Transcriptome assembly was performed with Trinity (v2.12.0) [62] on the mRNA partition for each sample. Trinity was run in strand-specific mode ('—SS_lib_type FR') and default settings. Coding regions were predicted on transcripts using TransDecoder (v5.5.0) with default settings and the 'Universal' genetic code. The resulting gff file from TransDecoder was then used to only keep assembled transcripts with a predicted coding sequence.

Obtaining microarray spot coordinates for different groups of spots

In order to obtain the coordinates of the microarray spots, we used the program ST Spot Detector, that uses the bright field image of the array (that shows the hematoxylin and eosin-stained cells), in combination with a fluorescent image (where the array spots are highly visible since they contain Cy3 dye), for obtaining the array coordinates for each spot [63]. Spots with one cell (singles), two cells (doubles), >2 cells (clusters) and no cells (background) were manually selected. When selecting the background spots, only those within the boundary of the cell droplet are chosen to minimize any false positive differences between the background and other spot types. The output of the program is a file with the coordinates of selected spots that can be used together with a matrix of counts (output of the ST pipeline) to calculate gene expression coming from selected spots. The precision (as reported in S2 Table) was calculated as 'number of true positives/ (number of false positives + number of true positives)' by comparing the data for the zero cell (empty) spots with the data for the single-cell spots. The data in the single-cell spots are a mixture of data from the cell in the spot and from background, hence: true positives + false positives. We cannot distinguish the true positives in these spots from the false positives, but if we assume that spots with single-cells are as contaminated with sequences from other cells on the array as are the empty spots, we can assume that the counts of false positives in the single-cell spots equal the counts in the empty spots. And since the 'true positives = (true positives + false positives)—false positives' we can derive this as the difference between the signal in the single-cell spots minus the signal in the empty spots. Thus, if we use median values of the number of detected unique transcripts, and if SC and EC denotes this value for single-cell and empty spots, respectively, we get precision = $(SC—EC)/SC$

Downstream data analysis

Downstream data analysis and generation of graphs were conducted in R [56].

Supporting information

S1 Fig. Bioanalyzer traces of final libraries showing product length when cells were fixed with formaldehyde and methanol. Upper and lower columns correspond to the same cells, collected at the same time points.
(PDF)

S2 Fig. Cells of *Heterocapsa* sp. and *P. tricornutum* imaged after different ways of fixation.
(PDF)

S3 Fig. Bioanalyzer traces of final libraries before and after freeze-thaw cycling. Upper and lower columns correspond to the same cells, collected at the same time points.
(PDF)

S4 Fig. Examples of A) *T. thermophila* cells prefixed with methanol, attached to the slide and stained with hematoxylin and eosin; B) *T. thermophila* cells prefixed with methanol and then freeze-thawed in liquid nitrogen and room temperature water (3 times, attached on the slide and stained with hematotoxylin and eosin).
(PDF)

S1 Table. Reagents and combinations used to test lysis on optimization slides.
(PDF)

S2 Table. Mean and median values of expressed genes and uniquely mapped reads for background, single, double and cluster spots in each library.
(PDF)

Acknowledgments

Author Contributions

Conceptualization: Rachel A. Foster, Anders F. Andersson.

Data curation: Vesna Grujčić, John Sundh, Bengt Sennblad.

Formal analysis: Vesna Grujčić, John Sundh, Bengt Sennblad, Anders F. Andersson.

Funding acquisition: Rachel A. Foster, Anders F. Andersson.

Investigation: Vesna Grujčić, Sami Saarenpää, Benjamin Norgren.

Methodology: Vesna Grujčić, Anders F. Andersson.

Project administration: Anders F. Andersson.

Resources: Stefania Giacomello, Rachel A. Foster.

Software: John Sundh, Bengt Sennblad.

Supervision: Rachel A. Foster, Anders F. Andersson.

Validation: Stefania Giacomello, Anders F. Andersson.

Visualization: Vesna Grujčić, Meike Latz.

Writing – original draft: Vesna Grujčić, Anders F. Andersson.

Writing – review & editing: Vesna Grujčić, Sami Saarenpää, John Sundh, Bengt Sennblad, Meike Latz, Rachel A. Foster, Anders F. Andersson.

References

1. Boenigk J, Arndt H. Bacterivory by heterotrophic flagellates: community structure and feeding strategies. Antonie Van Leeuwenhoek. 2002; 81: 465–480. https://doi.org/10.1023/a:1020509305868 PMID: 12448743

2. Šimek K, Kasalický V, Jezbera J, Horňák K, Nedoma J, Hahn MW, et al. Differential freshwater flagellate community response to bacterial food quality with a focus on Limnohabitans bacteria. ISME J. 2013; 7: 1519–1530. https://doi.org/10.1038/ismej.2013.57 PMID: 23552621

3. Worden AZ, Follows MJ, Giovannoni SJ, Wilken S, Zimmerman AE, Keeling PJ. Rethinking the marine carbon cycle: factoring in the multifarious lifestyles of microbes. Science. 2015; 347: 1257594.

4. Weisse T, Anderson R, Arndt H, Calbet A, Hansen PJ, Montagnes DJS. Functional ecology of aquatic phagotrophic protists—Concepts, limitations, and perspectives. Eur J Protistol. 2016; 55: 50–74. https://doi.org/10.1016/j.ejop.2016.03.003 PMID: 27094869

5. Keeling PJ, Campo JD. Marine Protists Are Not Just Big Bacteria. Curr Biol. 2017; 27: R541–R549. https://doi.org/10.1016/j.cub.2017.03.075 PMID: 28586691

6. Sibbald SJ, Archibald JM. More protist genomes needed. Nat Ecol Evol. 2017; 1: 145. https://doi.org/10.1038/s41559-017-0145 PMID: 28812681

7. Miao W, Song L, Ba S, Zhang L, Guan G, Zhang Z, et al. Protist 10,000 Genomes Project. Innovation (Camb). 2020; 1: 100058. https://doi.org/10.1016/j.xinn.2020.100058 PMID: 34557722

8. Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol. 2017; 2: 1533–1542. https://doi.org/10.1038/s41564-017-0012-7 PMID: 28894102

9. Alneberg J, Bennke C, Beier S, Bunse C, Quince C, Ininbergs K, et al. Ecosystem-wide metagenomic binning enables prediction of ecological niches from genomes. Commun Biol. 2020; 3: 119. https://doi.org/10.1038/s42003-020-0856-x PMID: 32170201

10. Delmont TO, Gaia M, Hinsinger DD, Frémont P, Vanni C, Fernandez-Guerra A, et al. Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean. Cell Genomics. 2022; 2: 100123. https://doi.org/10.1016/j.xgen.2022.100123 PMID: 36778897

11. Rinke C, Lee J, Nath N, Goudeau D, Thompson B, Poulton N, et al. Obtaining genomes from uncultivated environmental microorganisms using FACS-based single-cell genomics. Nat Protoc. 2014; 9: 1038–1048. https://doi.org/10.1038/nprot.2014.067 PMID: 24722403

12. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods. 2009; 6: 377–382. https://doi.org/10.1038/nmeth.1315 PMID: 19349980

13. Shaw R, Tian X, Xu J. Single-Cell Transcriptome Analysis in Plants: Advances and Challenges. Mol Plant. 2021; 14: 115–126. https://doi.org/10.1016/j.molp.2020.10.012 PMID: 33152518

14. Onsbring H, Tice AK, Barton BT, Brown MW, Ettema TJG. An efficient single-cell transcriptomics workflow for microbial eukaryotes benchmarked on Giardia intestinalis cells. BMC Genomics. 2020; 21: 448. https://doi.org/10.1186/s12864-020-06858-7 PMID: 32600266

15. Urbonaite G, Lee JTH, Liu P, Parada GE, Hemberg M, Acar M. A yeast-optimized single-cell transcriptomics platform elucidates how mycophenolic acid and guanine alter global mRNA levels. Commun Biol. 2021; 4: 822. https://doi.org/10.1038/s42003-021-02320-w PMID: 34193958

16. Ku C, Sebé-Pedrós A. Using single-cell transcriptomics to understand functional states and interactions in microbial eukaryotes. Philos Trans R Soc Lond B Biol Sci. 2019; 374: 20190098. https://doi.org/10.1098/rstb.2019.0098 PMID: 31587645

17. Ku C, Sheyn U, Sebé-Pedrós A, Ben-Dor S, Schatz D, Tanay A, et al. A single-cell view on alga-virus interactions reveals sequential transcriptional programs and infection states. Science Advances. 2020; 6: eaba4137. https://doi.org/10.1126/sciadv.aba4137 PMID: 32490206

18. von Dassow P, Montresor M. Unveiling the mysteries of phytoplankton life cycles: patterns and opportunities behind complexity. J Plankton Res. 2011; 33: 3–12.

19. Brosnahan ML, Ralston DK, Fischer AD, Solow AR, Anderson DM. Bloom termination of the toxic dinoflagellate Alexandrium catenella: Vertical migration behavior, sediment infiltration, and benthic cyst yield. Limnol Oceanogr. 2017; 62: 2829–2849. https://doi.org/10.1002/lno.10664 PMID: 29263558

20. Yadav V, Sun S, Heitman J. On the evolution of variation in sexual reproduction through the prism of eukaryotic microbes. Proc Natl Acad Sci U S A. 2023; 120: e2219120120. https://doi.org/10.1073/pnas.2219120120 PMID: 36867686

21. Kolisko M, Boscaro V, Burki F, Lynn DH, Keeling PJ. Single-cell transcriptomics for microbial eukaryotes. Curr Biol. 2014; 24: R1081–2. https://doi.org/10.1016/j.cub.2014.10.026 PMID: 25458215

22. Ying Yan, Maurer-Alcalá Xyrus X., Knight Rob, Kosakovsky Pond Sergei L., Katz Laura A., Cavanaugh Colleen M. Single-Cell Transcriptomics Reveal a Correlation between Genome Architecture and Gene Family Evolution in Ciliates. MBio. 2019; 10: e02524–19. https://doi.org/10.1128/mBio.02524-19 PMID: 31874915

23. Liu Z, Hu SK, Campbell V, Tatters AO, Heidelberg KB, Caron DA. Single-cell transcriptomics of small microbial eukaryotes: limitations and potential. ISME J. 2017; 11: 1282–1285. https://doi.org/10.1038/ismej.2016.190 PMID: 28060364

24. Ma F, Salomé PA, Merchant SS, Pellegrini M. Single-cell RNA sequencing of batch Chlamydomonas cultures reveals heterogeneity in their diurnal cycle phase. Plant Cell. 2021; 33: 1042–1057. https://doi.org/10.1093/plcell/koab025 PMID: 33585940

25. Vickovic S, Ståhl PL, Salmén F, Giatrellis S, Westholm JO, Mollbrink A, et al. Massive and parallel expression profiling using microarrayed single-cell sequencing. Nat Commun. 2016; 7: 13182. https://doi.org/10.1038/ncomms13182 PMID: 27739429

26. Le Costaouëc T, Unamunzaga C, Mantecon L, Helbert W. New structural insights into the cell-wall polysaccharide of the diatom Phaeodactylum tricornutum. Algal Research. 2017; 26: 172–179.

27. Salas R, Tillmann U, Kavanagh S. Morphological and molecular characterization of the small armoured dinoflagellate Heterocapsa minima (Peridiniales, Dinophyceae). Eur J Phycol. 2014; 49: 413–428.

28. Nusblat AD, Bright LJ, Turkewitz AP. Conservation and innovation in Tetrahymena membrane traffic: proteins, lipids, and compartments. Methods Cell Biol. 2012; 109: 141–175. https://doi.org/10.1016/B978-0-12-385967-9.00006-2 PMID: 22444145

29. Giacomello S, Lundeberg J. Preparation of plant tissue to enable Spatial Transcriptomics profiling using barcoded microarrays. Nat Protoc. 2018; 13: 2425–2446. https://doi.org/10.1038/s41596-018-0046-1 PMID: 30353173

30. Piwosz K, Mukherjee I, Salcher MM, Grujčić V, Šimek K. CARD-FISH in the Sequencing Era: Opening a New Universe of Protistan Ecology. Front Microbiol. 2021; 12: 640066. https://doi.org/10.3389/fmicb.2021.640066 PMID: 33746931

31. Olenina I., Hajdu S., Edler L., Andersson A., Wasmund N., Busch S., et al. Biovolumes and size-classes of phytoplankton in the Baltic Sea. HELCOM BaltSea Environ Proc. 2006;No. 106,: 144pp.

32. Orias E, Cervantes MD, Hamilton EP. Tetrahymena thermophila, a unicellular eukaryote with separate germline and somatic genomes. Res Microbiol. 2011; 162: 578–586. https://doi.org/10.1016/j.resmic.2011.05.001 PMID: 21624459

33. Keeling PJ. Combining morphology, behaviour and genomics to understand the evolution and ecology of microbial eukaryotes. Philos Trans R Soc Lond B Biol Sci. 2019; 374: 20190085. https://doi.org/10.1098/rstb.2019.0085 PMID: 31587641

34. Mathur V, Wakeman KC, Keeling PJ. Parallel functional reduction in the mitochondria of apicomplexan parasites. Curr Biol. 2021; 31: 2920–2928.e4. https://doi.org/10.1016/j.cub.2021.04.028 PMID: 33974849

35. Mathur V, Kwong WK, Husnik F, Irwin NAT, Kristmundsson Á, Gestal C, et al. Phylogenomics Identifies a New Major Subgroup of Apicomplexans, Marosporida class nov., with Extreme Apicoplast Genome Reduction. Genome Biol Evol. 2021; 13. https://doi.org/10.1093/gbe/evaa244 PMID: 33566096

36. Zha Y, Chong H, Yang P, Ning K. Microbial dark matter: from discovery to applications. Genomics Proteomics Bioinformatics. 2022. https://doi.org/10.1016/j.gpb.2022.02.007 PMID: 35477055

37. Colin S, Coelho LP, Sunagawa S, Bowler C, Karsenti E, Bork P, et al. Quantitative 3D-imaging for cell biology and ecology of environmental microbial eukaryotes. Elife. 2017; 6. https://doi.org/10.7554/eLife.26066 PMID: 29087936

38. Cho C-S, Xi J, Si Y, Park S-R, Hsu J-E, Kim M, et al. Microscopic examination of spatial transcriptome using Seq-Scope. Cell. 2021; 184: 3559–3572.e22. https://doi.org/10.1016/j.cell.2021.05.010 PMID: 34115981

39. Chen A, Liao S, Cheng M, Ma K, Wu L, Lai Y, et al. Large field of view-spatially resolved transcriptomics at nanoscale resolution. bioRxiv. 2021. p. 2021.01.17.427004. https://doi.org/10.1101/2021.01.17.427004

40. Picelli S, Faridani OR, Björklund AK, Winberg G, Sagasser S, Sandberg R. Full-length RNA-seq from single cells using Smart-seq2. Nat Protoc. 2014; 9: 171–181. https://doi.org/10.1038/nprot.2014.006 PMID: 24385147

41. Vacek V, Novák LVF, Treitli SC, Táborský P, Cepicka I, Kolísko M, et al. Fe-S Cluster Assembly in Oxymonads and Related Protists. Mol Biol Evol. 2018; 35: 2712–2718. https://doi.org/10.1093/molbev/msy168 PMID: 30184127

42. Krabberød AK, Orr RJS, Bråte J, Kristensen T, Bjørklund KR, Shalchian-Tabrizi K. Single Cell Transcriptomics, Mega-Phylogeny, and the Genetic Basis of Morphological Innovations in Rhizaria. Mol Biol Evol. 2017; 34: 1557–1573. https://doi.org/10.1093/molbev/msx075 PMID: 28333264

43. Lax G, Kolisko M, Eglit Y, Lee WJ, Yubuki N, Karnkowska A, et al. Multigene phylogenetics of euglenids based on single-cell transcriptomics of diverse phagotrophs. Mol Phylogenet Evol. 2021; 159: 107088. https://doi.org/10.1016/j.ympev.2021.107088 PMID: 33545276

44. Dohn R, Xie B, Back R, Selewa A, Eckart H, Rao RP, et al. mDrop-Seq: Massively Parallel Single-Cell RNA-Seq of Saccharomyces cerevisiae and Candida albicans. Vaccines (Basel). 2021; 10. https://doi.org/10.3390/vaccines10010030 PMID: 35062691

45. Ståhl PL, Salmén F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science. 2016; 353: 78–82. https://doi.org/10.1126/science.aaf2403 PMID: 27365449

46. Ji AL, Rubin AJ, Thrane K, Jiang S, Reynolds DL, Meyers RM, et al. Multimodal Analysis of Composition and Spatial Architecture in Human Squamous Cell Carcinoma. Cell. 2020; 182: 497–514.e22. https://doi.org/10.1016/j.cell.2020.05.039 PMID: 32579974

47. Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, et al. Quantitative single-cell RNA-seq with unique molecular identifiers. Nat Methods. 2014; 11: 163–166. https://doi.org/10.1038/nmeth.2772 PMID: 24363023

48. Thrane K, Eriksson H, Maaskola J, Hansson J, Lundeberg J. Spatially Resolved Transcriptomics Enables Dissection of Genetic Heterogeneity in Stage III Cutaneous Malignant Melanoma. Cancer Res. 2018; 78: 5970–5979. https://doi.org/10.1158/0008-5472.CAN-18-0747 PMID: 30154148

49. Salmén F, Ståhl PL, Mollbrink A, Navarro JF, Vickovic S, Frisén J, et al. Barcoded solid-phase RNA capture for Spatial Transcriptomics profiling in mammalian tissue sections. Nat Protoc. 2018; 13: 2501–2534. https://doi.org/10.1038/s41596-018-0045-2 PMID: 30353172

50. Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. Nat Methods. 2012; 9: 671–675. https://doi.org/10.1038/nmeth.2089 PMID: 22930834

51. Jemt A, Salmén F, Lundmark A, Mollbrink A, Fernández Navarro J, Ståhl PL, et al. An automated approach to prepare tissue-derived spatially barcoded RNA-sequencing libraries. Sci Rep. 2016; 6: 37137. https://doi.org/10.1038/srep37137 PMID: 27849009

52. Lundin S, Stranneheim H, Pettersson E, Klevebring D, Lundeberg J. Increased throughput by parallelization of library preparation for massive sequencing. PLoS One. 2010; 5: e10029. https://doi.org/10.1371/journal.pone.0010029 PMID: 20386591

53. Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, et al. The Phaeodactylum genome reveals the evolutionary history of diatom genomes. Nature. 2008; 456: 239–244. https://doi.org/10.1038/nature07410 PMID: 18923393

54. Eisen JA, Coyne RS, Wu M, Wu D, Thiagarajan M, Wortman JR, et al. Macronuclear genome sequence of the ciliate Tetrahymena thermophila, a model eukaryote. PLoS Biol. 2006; 4: e286. https://doi.org/10.1371/journal.pbio.0040286 PMID: 16933976

55. Pertea G, Pertea M. GFF Utilities: GffRead and GffCompare. F1000Res. 2020; 9. https://doi.org/10.12688/f1000research.23297.2 PMID: 32489650

56. R Foundation for Statistical Computing, Vienna, Austria. R Core Team (2020). R: A language and environment for statistical computing. 2020 [cited 7 Aug 2022]. Available: https://www.r-project.org/

57. Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, et al. Sustainable data analysis with Snakemake. F1000Res. 2021; 10: 33. https://doi.org/10.12688/f1000research.29032.2 PMID: 34035898

58. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011; 17: 10–12.

59. Kopylova E, Noé L, Touzet H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics. 2012; 28: 3211–3217. https://doi.org/10.1093/bioinformatics/bts611 PMID: 23071270

60. Andrews S. FASTQC. A quality control tool for high throughput sequence data. 2010 [cited 7 Aug 2022]. Available: https://www.bibsonomy.org/bibtex/f230a919c34360709aa298734d63dca3

61. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016; 32: 3047–3048. https://doi.org/10.1093/bioinformatics/btw354 PMID: 27312411

62. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011; 29: 644–652. https://doi.org/10.1038/nbt.1883 PMID: 21572440

63. Wong K, Navarro JF, Bergenstråhle L, Ståhl PL, Lundeberg J. ST Spot Detector: a web-based application for automatic spot and tissue detection for spatial Transcriptomics image datasets. Bioinformatics. 2018; 34: 1966–1968. https://doi.org/10.1093/bioinformatics/bty030 PMID: 29360929