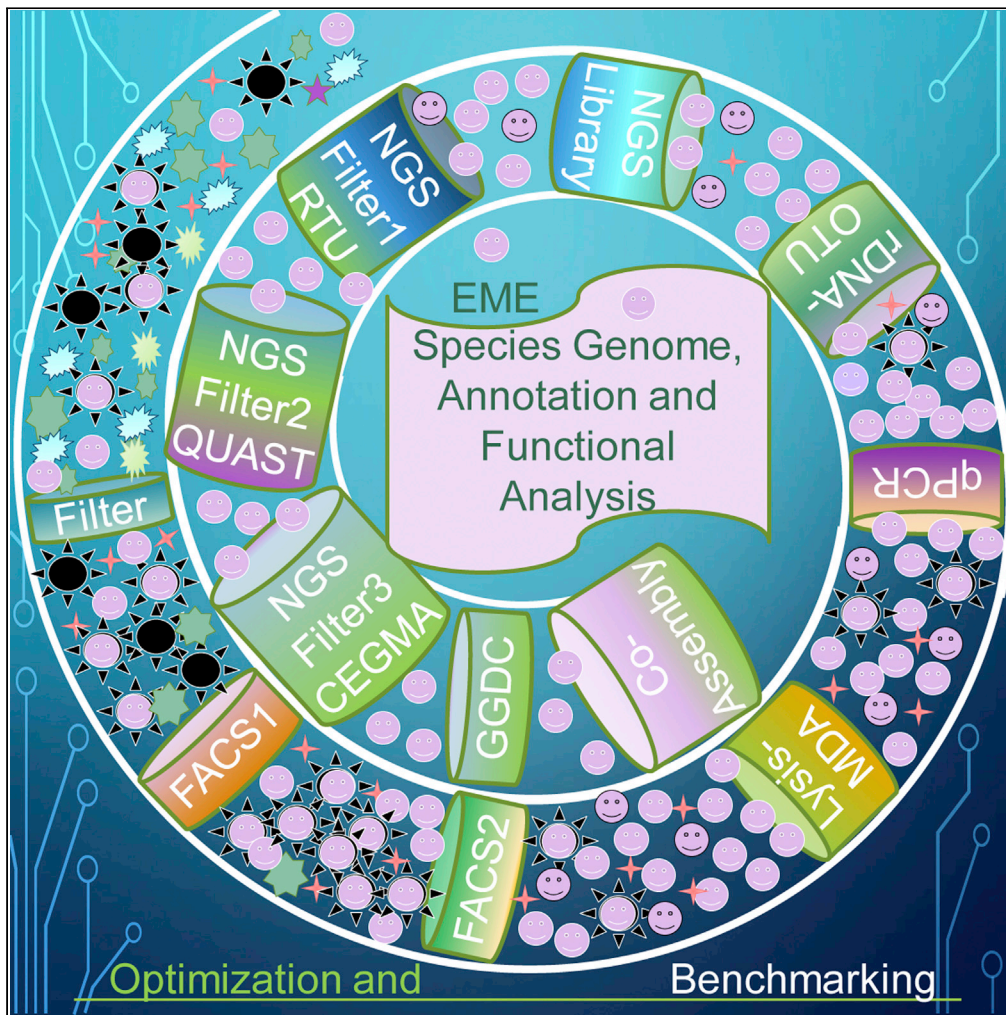


## Article

## A single-cell genomics pipeline for environmental microbial eukaryotes



Doina Ciobanu,  
Alicia Clum,  
Steven  
Ahrendt, ..., Igor V.  
Grigoriev, Timothy  
Y. James, Jan-  
Fang Cheng

dgciobanu@lbl.gov (D.C.)  
jfcheng@lbl.gov (J.-F.C.)

**Highlights**

We optimized single-cell methodology using a broad sample range, for EME

We combined bioinformatic and bench protocols into a concise workflow

We benchmarked the pipeline and used it on environmental samples

We selected a set of QC criteria for best genome quality prediction

## Article

## A single-cell genomics pipeline for environmental microbial eukaryotes

Doina Ciobanu,<sup>1,8,11,\*</sup> Alicia Clum,<sup>1,8</sup> Steven Ahrendt,<sup>1,2,8</sup> William B. Andreopoulos,<sup>1,8</sup> Asaf Salamov,<sup>1,8</sup> Sandy Chan,<sup>1,3</sup> C. Alisha Quandt,<sup>4,10</sup> Brian Foster,<sup>1</sup> Jan P. Meier-Kolthoff,<sup>5</sup> Yung Tsu Tang,<sup>6</sup> Patrick Schwientek,<sup>1,9</sup> Gerald L. Benny,<sup>7</sup> Matthew E. Smith,<sup>7</sup> Diane Bauer,<sup>1</sup> Shweta Deshpande,<sup>1</sup> Kerrie Barry,<sup>1</sup> Alex Copeland,<sup>1</sup> Steven W. Singer,<sup>6</sup> Tanja Woyke,<sup>1</sup> Igor V. Grigoriev,<sup>1,2,8</sup> Timothy Y. James,<sup>4,8</sup> and Jan-Fang Cheng<sup>1,8,\*</sup>

## SUMMARY

**Single-cell sequencing of environmental microorganisms is an essential component of the microbial ecology toolkit. However, large-scale targeted single-cell sequencing for the whole-genome recovery of uncultivated eukaryotes is lagging. The key challenges are low abundance in environmental communities, large complex genomes, and cell walls that are difficult to break. We describe a pipeline composed of state-of-the-art single-cell genomics tools and protocols optimized for poorly studied and uncultivated eukaryotic microorganisms that are found at low abundance. This pipeline consists of seven distinct steps, beginning with sample collection and ending with genome annotation, each equipped with quality review steps to ensure high genome quality at low cost. We tested and evaluated each step on environmental samples and cultures of early-diverging lineages of fungi and Chromista/SAR. We show that genomes produced using this pipeline are almost as good as complete reference genomes for functional and comparative genomics for environmental microbial eukaryotes.**

## INTRODUCTION

Single-cell genomics has significantly advanced mammalian and prokaryote studies related to human health, revisions of the tree of life, and biotechnology, as well as enhanced our understanding of the roles that microbes play in ecosystems (Gawad et al., 2016 and references inside; Linnarsson and Teichmann, 2016; Macaulay and Voet, 2014; Neu et al., 2017; Rinke et al., 2013; Stepanauskas, 2012). In some microbial eukaryotes, such as fungi and protozoa, single-cell genomics has been used to reveal the ecological and biological functions of some uncultured species (Berbee et al., 2017; Lin et al., 2014; Roy et al., 2014; Yoon et al., 2011 and based on this pipeline, Ahrendt et al., 2018). Nevertheless, to date, the majority of microbial eukaryote species are not considered feasible targets for genomic environmental studies (metagenomics and single-cell studies). In particular, genomes of uncultivated environmental microbial eukaryotes (EMEs) remain largely unavailable for the currently available genomic technology (Hyde, 2001; Lazarus and James, 2015; Sibbald and Archibald, 2017). Most studies that have recovered single-cell eukaryote genomes were performed on organisms that are abundant, have been successfully cultivated, or have a well-described molecular physiology (Gawryluk et al., 2016; Lin et al., 2014; Roy et al., 2014; Troell et al., 2016; Yoon et al., 2011; Zhang et al., 2017). The genomes of a diversity of low-abundance (at a concentration below 5% to as low as 0.01% in environments) EMEs (Tkacz et al., 2018; Wurzbacher et al., 2017 and present study target enrichment estimates) are therefore unexplored.

A quick look at fungi and Chromista shows the extent of skewed representation of annotated genomes and the knowledge gap in current genome biology. For example, in the fungal genome database MycoCosm (Grigoriev et al., 2014), seven of nine fungal phyla are early diverging lineages with more than 2,000 species described (Blackwell, 2011; Stajich et al., 2009), whereas the 173 sequenced and annotated genomes of early diverging fungi (EDF) represent only  $\leq 10\%$  of the total number of fungal genomes and are heavily skewed to the derived group Dikarya (1,626). Likewise, in Chromista (Ruggiero et al., 2015) or supergroup SAR (Burki et al., 2019), the phylum Ciliophora has sequenced genomes for only 4 of 11 classes. Furthermore, 20 of 23 sequenced and annotated genomes of Ciliophora species available in NCBI belong to two classes, and the majority of these genomes are incompletely annotated. The full list of unrepresented

<sup>1</sup>US Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory Berkeley, Berkeley, CA, USA

<sup>2</sup>Department of Plant and Microbial Biology, University of California Berkeley, Berkeley, CA 94720, USA

<sup>3</sup>Geisel School of Medicine at Dartmouth, Hanover, NH 03755, USA

<sup>4</sup>Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109, USA

<sup>5</sup>Department of Bioinformatics and Databases, Leibniz Institute DSMZ - German Collection of Microorganisms and Cell Cultures, Inhoffenstrasse 7B, 38124 Braunschweig, Germany

<sup>6</sup>Joint BioEnergy Institute, Emeryville, CA 94608, USA

<sup>7</sup>Department of Plant Pathology, University of Florida, Gainesville, FL 32611, USA

<sup>8</sup>These authors contributed equally

<sup>9</sup>Present address: Oralta, 479 Jessie St, San Francisco, CA 94103, USA

<sup>10</sup>Present address: Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, CO, 80309, USA

<sup>11</sup>Lead contact

\*Correspondence: dciobanu@lbl.gov (D.C.), jfcheng@lbl.gov (J.-F.C.)  
<https://doi.org/10.1016/j.isci.2021.102290>



phyla and classes from these two kingdoms is much longer. However, rDNA operational taxonomic unit (OTU) screening, environmental observations, and biochemical studies suggest that many of the understudied lineages are ubiquitous, extremely diverse, and highly adaptable (Alexander et al., 2016; Berbee et al., 2017; Blackwell, 2011; Foissner, 1999, 2009; Hyde, 2001; Lazarus and James, 2015; Sibbald and Archibald, 2017; Stajich et al., 2009). Consequently, there is a need for broader sampling of eukaryotic genome diversity (Sibbald and Archibald, 2017).

In addition to developing affordable tools for deeper and broader phylogenetic sampling targeting low-abundance EMEs in their native habitats, assessing the level of genome completeness is critical for understanding EME function and evolution. We explored the latter question in our recent publication (Ahrendt et al., 2018) using the pipeline that we describe in this study. Among the large number of publications dedicated to single-cell genomics (Arriola et al., 2007; Chen et al., 2017; Clingenpeel et al., 2014, 2015; Ellegaard et al., 2013; Garvin et al., 2015; Gawad et al., 2016; Gawryluk et al., 2016; Lan et al., 2017; Lin et al., 2014; Linnarsson and Teichmann, 2016; Macaulay and Voet, 2014; Neu et al., 2017; Rinke et al., 2013, 2014; Roy et al., 2014; Spits et al., 2006; Stepanauskas, 2012; Troell et al., 2016; Zhang et al., 2017), none has addressed the specific challenges associated with obtaining high-quality annotated *de novo* assembled genomes of poorly studied, low-abundance EMEs from a wide range of environmental samples.

Here, we explore the possibilities and current limitations of eliminating bias in genome representation by adapting current single-cell genomics methods into a pipeline for studying and annotating the genomes of EMEs with low and ultralow abundance to produce a broader representation of eukaryotic organisms in genomic studies. We explore, step by step, the potential critical impact of each challenge on both the quality of the recovered genomes and the cost of the study and the characteristics that set the genomics of single-cell EMEs apart from those of bacteria, archaea, and abundant or cultivated eukaryotes. The analyzed set of target species spans five fungal clades: four EDF (7 species with approximately 70 genomes), one Dikarya (1 species with 6 single-cell genomes), and one Chromista (Ciliophora with 7 single-cell genomes belonging to one unknown species). To date, this is the widest set of single-cell genomics methods, benchmarked against isolate genomes and tested on a wide range of sample types (isogenic cultivated, heterogenic co-cultivated, various environmental) spanning a wide phylogenetic range of target species.

Multiple displacement amplification (MDA) is the most widely used single-cell genome amplification method. MDA-associated genome amplification bias (GAB) has been shown to affect the quantification of copy number variation, SNP, and chimera formation (Garvin et al., 2015; Hou et al., 2015; Lasken and Stockwell, 2007; Zong et al., 2012). For mammalian cells, it has been shown that whole-genome amplification (WGA) methods other than MDA have lower GAB (Foissner, 2009; Gawad et al., 2016; Hou et al., 2015; Ning et al., 2015). Nevertheless, when appropriate genome assembly algorithms are applied, MDA-generated high-molecular-weight DNA fragments are of paramount value for *de novo* assemblies compared with linear amplification methods involving PCR (Spits et al., 2006).

Cell wall lysis (CWL) is another common challenge for single-cell genomics of the environmental microbiome (Brown and Audet, 2008; Tighe et al., 2017), but it is not a problem for mammalian species or species lacking cell walls, for which most single-cell genomics methods have been developed. Coupling CWL with downstream genome amplification has been explored for bacterial single-cell genomics methods (Clingenpeel et al., 2014, 2015; Rinke et al., 2014). Inadequate CWL and DNA contamination have been found to affect genome completeness in bacteria and perhaps lead to amplification bias in bacterial genomes (Clingenpeel et al., 2014, 2015). The same authors observed that an early start of genome amplification (SGA) correlated with larger genome assemblies, presumably due to less biased amplification, suggesting that early SGA occurs when lysis is complete (Clingenpeel et al., 2014, 2015). Only one study has explored a number of protist species using the same lysis method (Yoon et al., 2011). In other studies of single-cell fungal, plant, or protist genomes, CWL has been tailored for a specific target (Gawryluk et al., 2016; Lin et al., 2014; Roy et al., 2014; Troell et al., 2016). To date, a universal CWL, compatible with same reaction MDA was explored only for single-cell prokaryote and mammalian genomes (reviewed in Gawad et al., 2016 and citations). The use of either low-volume same-tube reactions or microfluidics has been shown to significantly reduce both amplification bias and contamination levels and to result in more complete genomes (Rinke et al., 2014, reviewed in Gawad et al., 2016). However, existing microfluidics devices (Gawad et al., 2016; Lan et al., 2017) do not accommodate the diverse size and shape of environmental microorganisms and/or cell lysis requirements. There is a need for methods that can

accommodate lysis conditions that work for a wide range of unknown organisms without having to purify nucleic acids and without inhibiting subsequent molecular reactions.

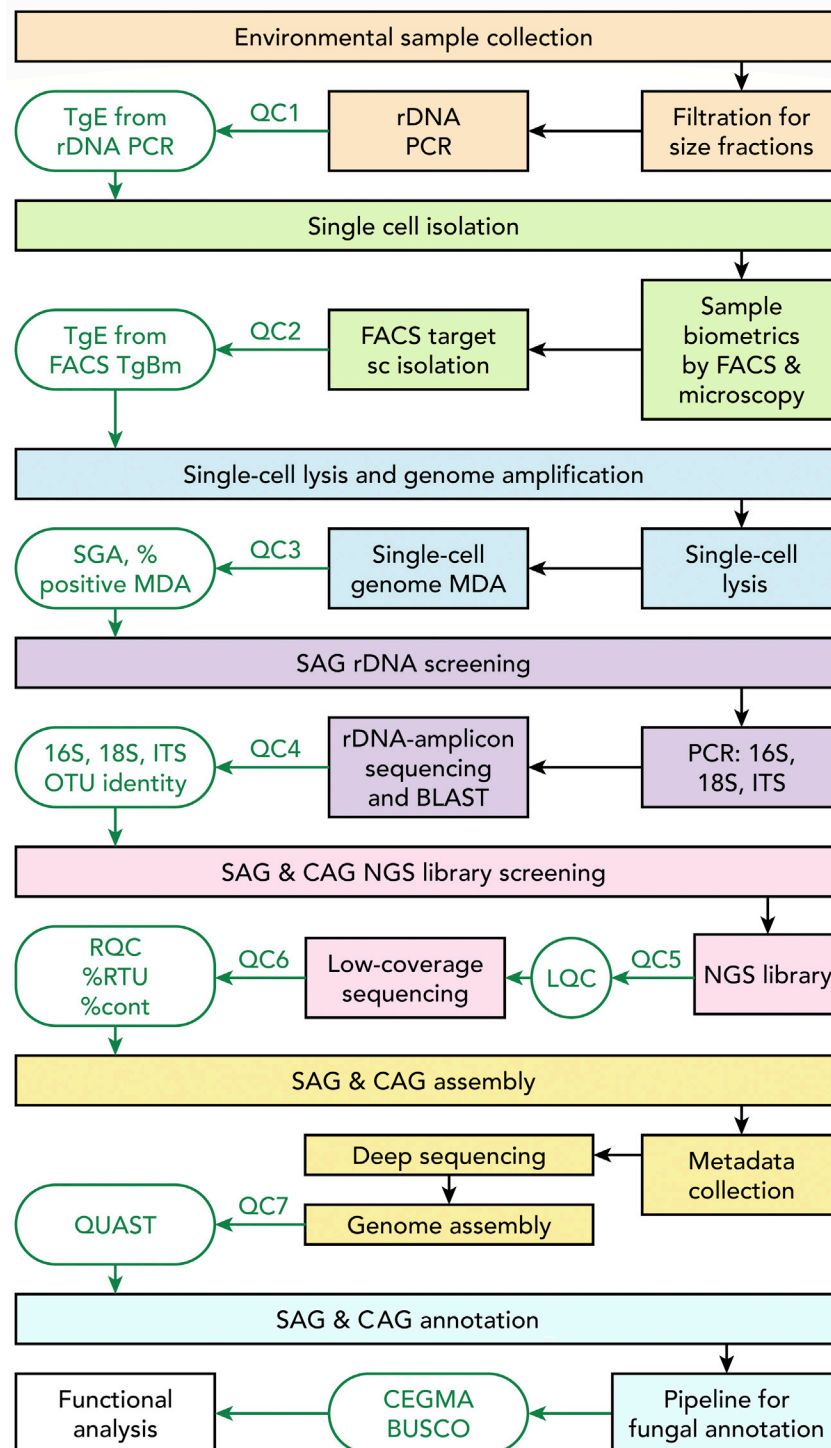
Specifically, we explored (1) the accurate single-cell isolation of the target species at a one-in-millions representation in the environmental sample by testing fluorescent-activated cell sorting (FACS) and microfluidic methods; (2) the suitability of various new and known (Brown and Audet, 2008; Clingenpeel et al., 2014, 2015; Rinke et al., 2014; Tighe et al., 2017; Yoon et al., 2011) CWL methods for a broad range of organisms and their compatibility with downstream high-throughput semiautomated processes; (3) all factors known to affect genome completeness, such as GAB due to MDA (Hou et al., 2015) and genome size, previously reported in mammalian cells and prokaryotes (Garvin et al., 2015; Hou et al., 2015; Lasken and Stockwell, 2007; Zong et al., 2012) but never explored in EMEs; (4) GC%, cell wall complexity, and genome structure as factors that have not been shown to cause MDA-GAB or genome incompleteness; (5) the costs associated with a high sequencing capacity, which has not been previously discussed in attempts for a broader and deeper phylogenomic mining of EMEs (unlike environmental prokaryotes with genomes that are 100-fold smaller than those of eukaryotes on average, mass sequencing of poor-quality or nontarget EME genomes can become prohibitively costly, even when using NovaSeq); (6) methods using shallow sequencing to evaluate genome quality before deep sequencing (Daley and Smith, 2014), which is particularly important for high throughput; (7) genome assembly tools (Bankevich et al., 2012; Butler et al., 2008; Peng et al., 2012) for accurate *de novo* reconstruction of single-cell EME genomes from novel lineages that suffered partial GAB and chimerization during amplification; (8) bioinformatics tools (Auch et al., 2010; Han et al., 2016; Meier-Kolthoff et al., 2013) to accurately predict single-cell intra- and interspecific genome distance for EMEs—a prerogative for assembling a genome by coassembling or selecting single-cell or multiple-cell genomes with similar and low genome distances that are characteristic of intraspecific phylogeny (coassemblies provide a higher level of genome completeness [Kogawa et al., 2018]); (9) genome completeness assessment tools (Parra et al., 2007, supplementary reference 4); and (10) correct gene structure prediction and annotation from fragmented and/or incomplete genomes.

## RESULTS

### Pipeline synopsis

We developed and benchmarked a 7-step pipeline for EME genome recovery, as illustrated in Figure 1. Brief description of our pipeline:

- 1. Environmental sample collection (Step 1), shipping, storage, and target species enrichment evaluation (QC1):** Step 1 included sample collection, storage, and shipping. When possible and necessary, filtration steps were used for target population enrichment. Enriched filters in suitable media were used to suspend cells before storage and shipment. QC1 included target organism enrichment evaluation (TgE) via rDNA profiling, microscopic identification and counting, or both. **Optimizations** involved enrichment and storage methods.
- 2. Single-cell isolation (Step 2) and target species enrichment evaluation (QC2):** During step 2, samples were visualized using a microscope and a FACS instrument; target populations were identified and sorted in bulk into large tubes in 0.02  $\mu\text{m}$  filtered original media. After the first round of FACS enrichment of the target population, a second round of sorting (as needed to reduce carryover contaminants) was used before single-set sorting. FACS-enriched target populations were used for FACS of single cells or batches of 10–100 cells into 384-well plates and immediately frozen on dry ice. QC2 included TgE via FACS with or without microscopic validation before single-cell isolation. **Optimizations** involved cell staining protocols and target cell isolation methods.
- 3. Single-cell lysis and genome amplification (Step 3) and lysis-MDA efficiency evaluation (QC3).** During step 3, isolated single cells (or batches of 10–100 cells) were lysed and amplified via MDA in the same well. Single-cell amplified genomes or multiple-cell amplified genomes were quantified in real time and at the endpoint. QC3 criteria included real-time-monitored SGA, percent positive MDA, fold genome amplification (FGA), and other criteria described in transparent methods step 3 and later. **Optimizations** involved the efficiency and compatibility of a number of lysis-MDA protocols.
- 4. Confirmation of species identity, contaminant, or symbiont occurrences in single and multiple amplified genomes (Step 4) and OTU screening (QC4).** During step 4, an aliquot of single and multiple amplified genomes was subjected to qPCR for rDNA genes followed by Sanger sequencing and



**Figure 1. Pipeline schematics for environmental microbial eukaryotic single-cell whole-genome recovery, *de novo* assembly, and annotation**

Square boxes show the pipeline steps and components. QC1 through 7 and green ovals show quality check steps and the main criteria used in this study, described in the brief pipeline overview. TgE, target organism enrichment; rDNA, ribosomal DNA; FACS, fluorescent-activated cell sorting; TgBm, target organism biometrics; sc, single cell; MDA, multiple displacement amplification; SGA, start of genome amplification; SAG, single amplified genome; CAG, composite amplified genome; OTU, operational taxonomic unit; BLAST, basic local alignment search tool

**Figure 1. Continued**


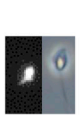
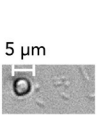

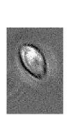

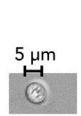
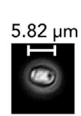
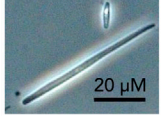
(Altschul et al., 1990); NGS, next-generation sequencing; LQC, NGS library quality check; RQC, Illumina sequencing read quality check; RTU, random 20-mer uniqueness; cont, contaminant identified by BLAST on NGS read; QUAST, quality assessment tool for genome assemblies (Lazarus et al., 2017); CEGMA, core eukaryotic genes mapping approach (Parra et al., 2007); BUSCO, (Simão et al., 2015). See also [Data S2](#) and [Table S10](#).

BLAST against NCBI and AFTOL (Celio et al., 2006) databases. QC4 included OTU identification, evaluation of Step 1 through Step 3 process efficiency, and selection of target genomes for sequencing. **Optimizations for step 4:** a range of 18S, 16S, 28S, and ITS rDNA primers and regions were tested for suitability with a broad phylogenetic range. **Optimizations for merging step 4 with step 5:** various rDNA assembly methods from shallow short-read sequencing and various next-generation sequencing (NGS) library protocols for high-throughput shallow sequencing were tested for suitability for rDNA assembly methods.

- 5. NGS library construction and shallow sequencing for single and multiple amplified genomes (Step 5) and library (QC5) and genome (QC6) quality predictions.** In step 5, genomic DNA was fragmented, and NGS libraries were constructed and pooled for shallow sequencing following Illumina guidelines. QC5 included library quality evaluation following Illumina recommendations. QC6 included genome quality prediction based on criteria such as percent contaminant, random 20-mer uniqueness (RTU) and other NGS-read-QC metrics described in supplemental section step 5. **Optimizations:** We tested various fragmentation protocols, read length and library construction protocols, and the use of various NGS-read-QC metrics to identify libraries with the following: (1) high GAB, (2) high chimerization rate, and (3) contamination and incomplete prediction of the genome assembly outcome by eliminating poor-quality libraries.
- 6. Deep sequencing for single and multiple amplified genome assembly, single amplified genome phylogenomic distance estimation, and coassembly of multiple single amplified genomes into a single-species genome (Step 6), single amplified genome assembly quality evaluation, and annotation quality prediction (QC7).** During step 6, we assembled genomes from deep-sequenced libraries that passed QC6 evaluation. We found that single and multiple amplified genome distances and their coassembled genomes had the same intraspecific distance. For genome distance calculation for species coassembled from single and multiple amplified genome, we used Genome-to-Genome Distance Calculator (GGDC) formula 2 described in Meier-Kolthoff et al. (2013). For QC7, we used quality assessment tool for genome assemblies (QUAST) (Lazarus et al., 2017) criteria and tested the core eukaryotic gene mapping approach (CEGMA) values (Parra et al., 2007). **Optimizations:** We tested all 20 QUAST criteria for predictability of assembly quality for high-quality annotations and further functional predictions. We reduced the number of necessary QUAST criteria to a few essential criteria. We tested CEGMA values as a predictor of genome completeness before annotation. We tested various genome distance estimation methods for suitability with large eukaryotic genomes that were amplified, fragmented, and incomplete.
- 7. Annotation and functional predictions (Step 7) and genome completeness evaluation (QC8).** In step 7, assembled genomes were annotated using the MycoCosm pipeline toolsets (Grigoriev et al., 2014). QC8 included CEGMA values (Parra et al., 2007) to assess genome completeness (using the presence of conserved core eukaryotic genes) before functional and phylogenomic analysis (Ahrendt et al., 2018). **Optimizations:** We tested both CEGMA and BUSCO values for genome completeness prediction; manual curation was necessary for species that had no close genome annotations.

**Sample selection for EME single-cell genomics pipeline benchmarking**

We tested this pipeline on a range of samples shown and described in [Figure 2](#), [Tables 1](#) and [2](#), and [Data S1](#). Samples ranged from pure environmental ([Data S1RR–S1U](#)) to dual cultures mimicking natural conditions ([Data S1A–S1P](#)). Samples had various levels of complexity ([Table 2](#) and [Figure 2](#)). The complexity level was estimated based on a range of factors described in [Table 2](#). Environmental samples ranged from high complexity (ultralow abundance of target species among high-abundance nontarget organisms in the original environment, see [Data S1T](#) and [S1U](#)) to medium complexity (see [Data S1RS](#) and [S1S](#) and [Table 2](#)). Dual cultures ranged from low ([Data S1G–LM](#)) and medium ([Data S1E](#), [S1F](#), [S1M–S1P](#)) to highly enriched for target species (see [Data S1A–S1D](#)). We examined the effect of organisms' shapes, sizes, and motility on single-cell isolation success and the range of enrichment levels relative to other taxa. The latter was

Target phylum	Ciliophora	Cryptomycota	Chytridiomycota	Chytridiomycota	Zoopagomycota*	Zoopagomycota	Zoopagomycota	Zoopagomycota	Ascomycota
Target species	Unknown ciliate (CiPr)	<i>Rozella allomycis</i> (R.a)	<i>Blyttiomycetes helicis</i> (B.h)	<i>Caulochytrium protostelioides</i> (C.p)	<i>Dimargaris cristalligena</i> (D.c)	<i>Piptocephalis cylindrospora</i> (P.c)	<i>Thamnocephalis sphaerospora</i> (T.s)	<i>Syncephalis pseudoplumigaleata</i> (S.p)	<i>Metschnikowia bicuspidata</i> (M.b) yeast (y) ascospore (a)
Biometric properties									
Lifestyle	Free living	Parasite	Saprobe	Parasite	Parasite	Parasite	Parasite	Parasite	Parasite
Host	NA	Fungi	Plant	Fungi	Fungi	Fungi	Fungi	Fungi	Crustacea
Genome GC, % <sup>#</sup>	38	35	54	65	49	51	55	56	51
Genome Size, Mb <sup>\$</sup>	120	12	48	13	31	12.6	16	13.9	10
TgE, %	1.8	83	10	30	31	32	14	67	2.7
Tgwidth, μm	20	5	5	3	5	2	6	3	2.5
Tglength, μm	20	7	5	5	8	5	6	6	13
SCL	10	1	8	2	5	3	7	5	9

**Figure 2. Main target single-cell diversity used for pipeline evaluation**

Shown here are nine of the eleven samples used. For the other two samples, see [Table 1](#). Pictures for the first through eighth species are sized relative the 5 μm scale bar. Heatmap colors reflect the spectrum of values: red, highest; yellow, lowest; and green, average. TgE is the FACS target enrichment estimated in step 2 of the pipeline (for step 1 TgE, shape and other details, see [Table 1](#) and [Video S1](#)). SCL is the sample complexity level (for details see [Table 2](#)). The genome size (\$<sup>§</sup>) was not known for any of the target organisms before assembly and was therefore estimated based on the assembly size and genome completeness. The genome average GC% (#) was not known before genome assembly and was therefore predicted based on the GC% of the existing genes or by estimation using the nearest phylogenetic group. \*The phylum Zoopagomycota was established by ([Spatafora et al., 2016](#)) in part using data obtained from these four single-cell genomes. Before this study, the phylogenetic data available for this group were limited. See also [Data S1](#).

<2%–85%; shapes were spherical, oval, cylindrical, trapezoidal, and needle-shaped; sizes were from 2 to 100 μm; motility modes were sessile or flagellated ([Table 1](#), [Figure 2](#), and [Data S1](#)).

For each sample and each target species, multiple genomes were isolated and analyzed. However, only the highest quality genomes were assembled and annotated. Pipeline optimizations are described in the [transparent methods](#) and [supplemental information](#) and were performed using separate sorting of 1, 10, 30, 50, and 100 cells per reaction and the set of QC criteria described in [Data S2](#). The results of the pipeline QC evaluation are illustrated in [Figure 3](#) and [Data S2](#), and they are explored in greater depth further in the article and in the [transparent methods](#) in [supplemental information](#). Specifically, single-cell isolation methods were tested, further developed, and optimized in steps 1 and 2 on 11 samples (see [Table 1](#), [Data S1](#) and [transparent methods](#) steps 1, 2). Steps 3 through 7 were tested, developed, and optimized on the 9 species described in [Figure 2](#) and [transparent methods](#) steps 3–7). We tested and optimized the merging of steps 4 and 5 using all 11 samples. We benchmarked the methods used in these steps, against the unamplified genomes of two fungal species (*Rozella allomycis* [Cryptomycota] and *Caulochytrium protostelioides* [Chytridiomycota]). Their unamplified genomes were isolated from the same highly enriched dual cultures of parasitic target fungi with their fungal host and microbiome used for single-cell isolation. Genome size and GC% variations of the known target species were used to test amplification efficiency and bias. The presence or absence of the cell wall in the target species was used to test lysis efficiency.

The EME pipeline single-cell genome recovery lowest efficiency threshold was identified as 1 target in 500 amplified genomes or 0%–2% rDNA OTU during step 1 screening, using three environmental samples: two aimed at fungal phyla (Cryptomycota and Chytridiomycota) with ultralow abundance of unknown target species and one aimed at one SAR phyla (Ciliophora) with unknown OTU (see details in [Data S1R–S1U](#) and [Table 1](#)). To refine the target enrichment and single-cell isolation methods, we used eight more samples with one known target species per sample ([Figure 2](#) and [Table 2](#)). Six of these samples harbored different mycoparasite target fungal species with different hosts, all dual non-axenic cultures with different microbiome diversity and target concentration levels (for details see [Data S1A–S1D](#) and [S1G–S1P](#)). Two other samples re-created the natural environment of the target species and were available in very low amount, one a symbiont of pollen ([Data S1E](#) and [S1F](#)) and the other a parasite of crustacean ([Data S1O](#)

**Table 1. Combination of critical factors affecting FACS single-cell isolation success of microbial eukaryotes from diverse samples**

Sample target organism name	Microscope evaluation of target population concentration in original sample, %	FACS assessment of target population concentration in original sample, %	target organism Shape	other organisms Shape	target size, $\mu\text{m}$	other organisms size, $\mu\text{m}$	target motility before FACS	Sort type	FACS target enrichment prior single cell sorting, %	total MDA positives, %	rDNA target positives, %
R.a – <i>Rozella allomyces</i>	100	80	tapered oval with flagella	NA	5x7 body with 10 flagella	< 3	Zoospore	double clean	95	92.7	27.04
S.p – <i>Syncephalis pseudoplumigaleata</i>	90	66	ellipse	round	3x6	<4	Autospore	double clean	90	6.9	0.46
P.c – <i>Piptocephalis cyclindrospora</i>	50	32/90	cylindrical, ellipse	round, cylindrical	2x5	<2	Autospore	double clean	99	45.13	3.1
D.c – <i>Dimargaris crystalligena</i>	90	30	tapered oval elongated	round	5x8	<5	Autospore	double clean	90	15.27	5.2
C.p – <i>Caulochytrium protostelioides</i>	100	30	tapered oval with flagella	NA	3x5 body with 20 flagella	0.5 to 20	Zoospore	double clean	90	29	11.45
T.s – <i>Thamnocephalis sphaerospora</i>	80	14	round		6	<6	Autospore	direct	14	32.98	3.64
B.h – <i>Blyttomyces helicus</i>	5	14	round with flagella	cylindrical, round, pear	5 body with 30 flagella	<3	Zoospore	direct	10	6.25	8.33
M.b, ascospore – <i>Metchnikowia bicuspidata</i>	10	3.5	needle	round	2.5x60	<13	Autospore	direct	3.5	18.4	1.1
M.b, yeast	10	2.6	elongated	round	2.5x13	<5	Autospore	direct	2.7	33.33	7.29
CIPr – Ciliate protist, 99% identity to 18S <i>Platyophrya</i> sp.	10	0.7	sphere and tapered oval	round, amorphous	12x12,20x20,30x30,50x50,50x100,	1,2,5,10,11	Cysts, trophonts	double clean	38	55.2	12
Gosling lake sample, target Cryptomycota new species	NA, rDNA - 0.1	FACS+itag: 0.0	Round, oblong	Round, oblong	1x1, 2x2	1-20	Zoospore, and unknown	double clean	30 *	50	0.02
Gosling lake sample, target Chytridiomycota new species	NA, rDNA – 2.2	FACS+itag: 0.0	Round, oblong	Round, oblong	1x1, 2x2	1-20	Zoospore, and unknown	double clean	30 *	50	0.02
Third Sister Lake, target Cryptomycota new species	NA, rDNA - 0.0	FACS+itag: 0.0	Round, oblong	Round, oblong	1x1, 2x2	1-20	Zoospore, and unknown	double clean	30 *	50	0.02
Third Sister Lake, target Chytridiomycota new species	NA, rDNA – 0.1	FACS+itag: 0.0	Round, oblong	Round, oblong	1x1, 2x2	1-20	Zoospore, and unknown	double clean	30 *	50	0.02

Color coding of the samples: green, environmental; light green, laboratory rec-created environmental sample without artificial substrate; blue, non-axenic heterogeneous co-cultures; white, highly enriched non-axenic co-cultures.

\* This is the concentration of an enriched FACS population containing target species.

and S1P). Both had distinct target species cell shapes albeit high content of contaminating organisms. All these samples were useful for studying the impact of different sample features on single-cell isolation efficiency (Table 1).

### EME single-cell genomics pipeline step-by-step bottlenecks and optimizations

We found that the following combination of factors prevented successful target single-cell FACS isolation: very low concentration of target coupled with one or two of the following: size above 30  $\mu\text{m}$ , presence of non-target organisms with highly similar biometric properties (e.g., flagella, cilia, shape, size, and light refraction index), and very low volume of sample (Tables 1 and 2). On the other hand, a two-step FACS isolation, whenever possible, improved target single-cell recovery, shown in Figure S1 and described in Optimization in step 2 in transparent methods.

This finding was further supported by the lack of strong correlation between target concentration in original sample and total number of positive MDA reactions, as well as target single-cell OTU (rDNA-PCR confirmed) (see Figure S4). However, there was a strong correlation between MDA-amplified single-cell and rDNA identified target OTU, shown in Figure S4, indicating that FACS enrichment step before single-cell sorting, and cell lysis, had a significant impact on the number of recovered target genomes.

We performed an extensive optimization of the single-cell lysis methods compatible with the same-well MDA, described in step 3 of the transparent methods and Data S3. We found one lysis method that allowed



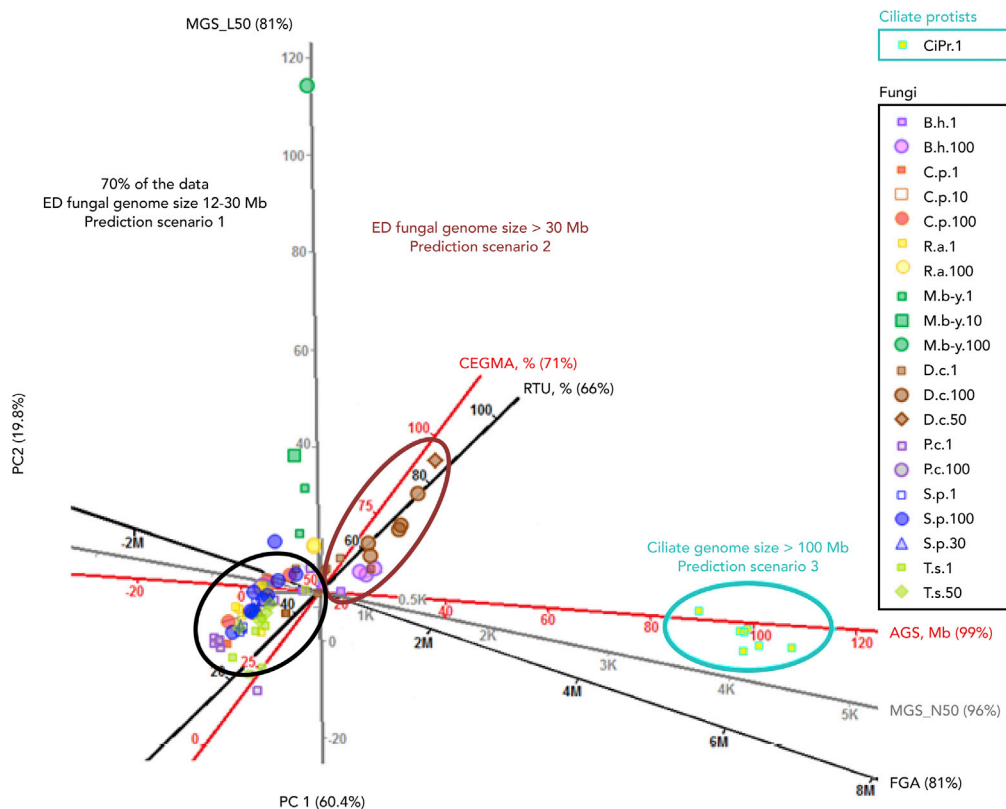
**Table 2. Range of criteria useful for sample complexity level prediction for EME single-cell genome recovery.**

Tested Samples in this study and their level of complexity.	R.a	C.p	P.c.	No sample in this study	D.c,S.p	No sample in this study	T.s	B.h	M.b.	Compost ciliate, Gosling and Third Sister lakes Cryptomycota and Chytridiomycota unknown species
Complexity Level	1	2	3	4	5	6	7	8	9	10
target cell abundance (concentration)	80–100%	30–100%	50–90%	30–80%	30–70%	30–50%	10–20%	1–15%	1–5%	≤1
target cell abundance (amount)	≥10 <sup>2</sup>	≥10 <sup>3</sup>	10 <sup>4</sup> – 10 <sup>5</sup>	10 <sup>4</sup> – 10 <sup>5</sup>	10 <sup>4</sup> – 10 <sup>6</sup>	10 <sup>4</sup> – 10 <sup>6</sup>	10 <sup>4</sup> – 10 <sup>6</sup>	10 <sup>4</sup> – 10 <sup>6</sup>	10 <sup>3</sup> – 10 <sup>6</sup>	≤1000
diversity of organisms	≤2	≤2	≤5	10≥5	50≥5	50≥10	50≥5	10≥2	100≥2	1000≥5
presence of 'competing' cells (with similar biometric characteristics)	0	0	0	≤2%	≤3%	≤4%	≤5%	≥10%	≥10%	≥10%
shape of the target cell	Spheroid, oval, cylindrical	Spheroid, oval, cylindrical	Spheroid, oval, cylindrical	Spheroid, oval, cylindrical	Spheroid, oval, trapezoidal, cylindrical	Needle-like, trapezoidal, Spheroid, oval, cylindrical	Needle-like, trapezoidal, Spheroid, oval, cylindrical	Needle-like, trapezoidal, Spheroid, oval, cylindrical	Needle-like, trapezoidal, Spheroid, oval, cylindrical	Needle-like, trapezoidal, Spheroid, oval, cylindrical
Size of the target cell	1–20uM	1–20uM	1–20uM	1–25uM	1–25uM	1–25uM	1–30uM	0.5 –50uM	0.5 –60uM	0.5 –80uM
target cell wall complexity (layers)	≤1	≤1	≤1	2	2	0–2	0–2	0–2	0–2	0–3

for the shortest protocol and reduction of the reagent carry-over contaminating DNA (transparent methods Step 3 optimization). Although this method worked well across all species, we observe some heterogeneity between species (Figure S2) and some, but less, difference between single-cell level and multiple-cell level within one species, when start of MDA was used as a proxy for lysis-MDA efficiency.

Further optimization of the OTU screening (see Table S1 for best rDNA primers) of the amplified EME genomes led to improved quality of the single-cell genome recovery, as well as reduction of costs, as described in detail in transparent methods Optimization step 4. Merging of Steps 4 and 5 of the pipeline offered an additional benefit of cost reduction and identification of EMEs with symbionts, which otherwise are screened out in Step 4. The merging of these steps relies on rDNA assembly tool capable of correctly assembling it from MDA-NGS reads. We optimized this step, described in transparent methods Optimization step 5 and the results of the rDNA assembly tools testing are shown in Figure S5.

Another improvement in single-cell genome quality during step 5 was the use of RTU measure. We found that libraries with high RTU (above 60%) produced low-biased genomes used later for co-assembly and annotation here, as opposed to the highly biased genomes with low 20-mer uniqueness, most screened



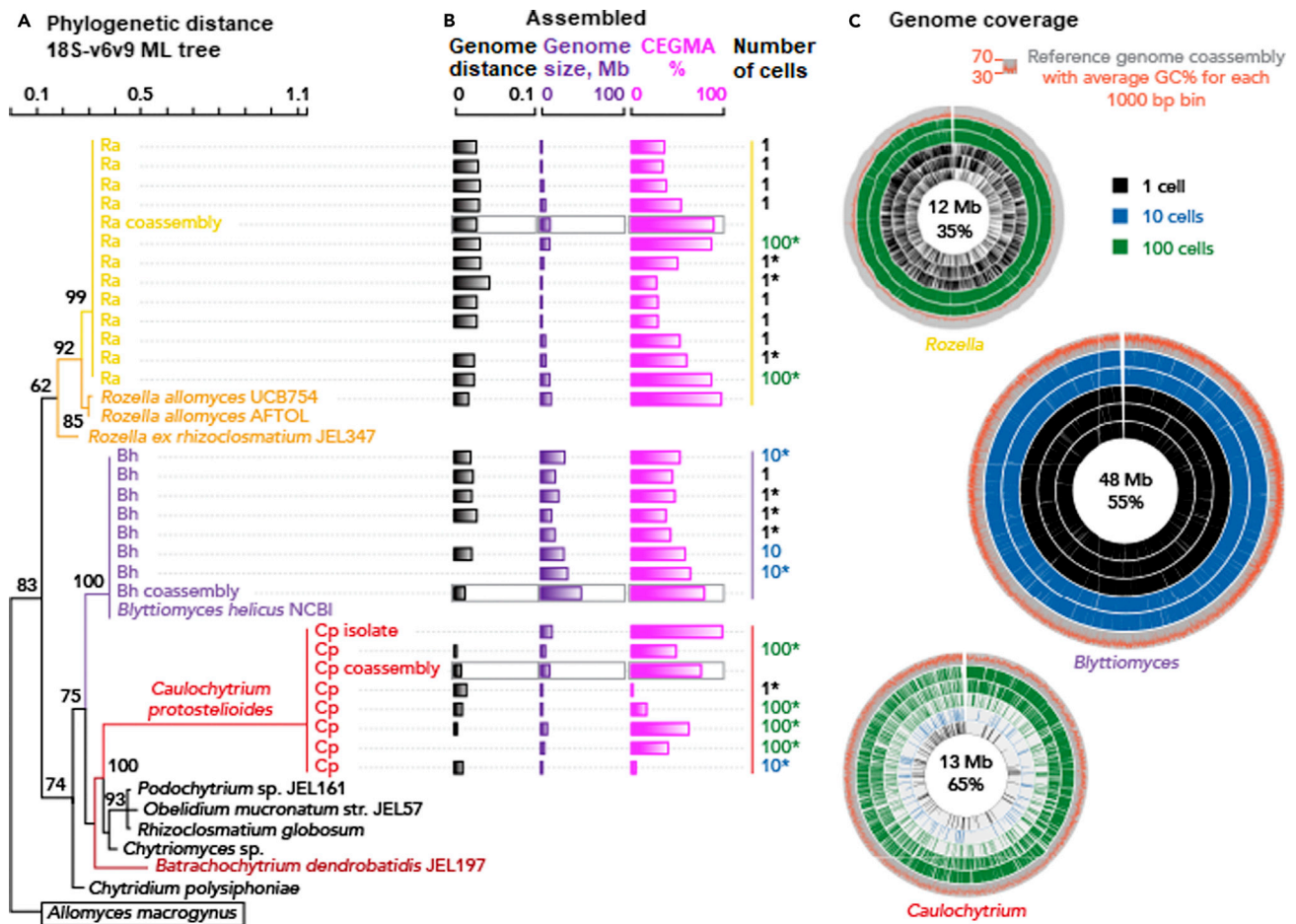
**Figure 3. Predictive value and applicability of the used QC criteria for a wide phylogenetic group**

Twenty QC criteria examined (see [Data S2](#)) can be reduced to six shown here. Axes color: black, pre-assembly criteria; gray, assembly metrics; red, pre-annotation criteria. Gower & Hand PCA biplot represents similarity between data points, with smaller distance higher similarity. Shown plot explains 80.2% of the variability for fungi plus ciliate protists group. Any point on the plot projected orthogonally onto the axes will show the approximate value of the variable. Percent at the end of the axes labels indicates predictability value of the axes. Species full names are given in [Figures 2](#) and [S1](#). ED, early diverging. See also [Table S10](#).

out during optimization step and some shown in [Figure 5](#). The correlation between RTU and genome completeness as CEGMA is very high, as illustrated in [Figure 3](#) and [Data S2c](#). Similarly, *de novo* assembly tools and type of NGS had a significant impact on the genome quality ([Tables S2–S4](#) and [Figure S7](#)), described in [transparent methods](#) optimization step 6.

Measuring genome completeness for *de novo* assemblies is an imperative requirement for quality evaluation. However, only approximate estimates could be obtained using mathematical algorithms. Two tools developed for eukaryotic genomes looked most promising: CEGMA ([Parra et al., 2007](#)) and BUSCO ([Simão et al., 2015](#)). We used CEGMA for our pipeline evaluation and later tested the newly developed BUSCO. Unexpectedly BUSCO did perform worse (detected less genes) than CEGMA for the EDF. BUSCO's inaccurate performance for EDF could be due to lower availability of a statistically significant number of EDF of a specific phylum and high diversity within phylum. We decided not to use this engine until a larger database of EDF annotated genomes is acquired and suggest periodic evaluation of BUSCO versus CEGMA for future EMEs studies.

Using CEGMA as a proxy for genome completeness, we evaluated the minimum number of single-cell amplified genomes necessary for each species to achieve a near-complete coassembly with quality comparable to isolated genomes obtained from unamplified DNA obtained from thousands of cells. To coassemble one species genome from multiple single-cell libraries, we tested various approaches for determining genome-to-genome distance ([Figures 4](#) and [6](#) and [Table 3](#) and [Table S8](#)). Once again, we found that only one of the three known approaches, GGDC tool, formula 2, performed the best for the



**Figure 4. Intra- and interspecific variabilities**

(A) Cryptomycota and Chytridiomycota 18S rDNA (region v6 to v9) ML tree based on the HKY85 nucleotide substitution model with bootstrap values shown above 60%.

(B) Assembled: Genome distance was calculated using GGDC formula 2, designed for incomplete isolated genomes (Auch et al., 2010; Meier-Kolthoff et al., 2013); the genome size shows the degree of variation in genome recovery between single-cell and multiple-cell sorts, and the core eukaryotic gene mapping approach (CEGMA) value reflects genome completeness. \*Assemblies used for the genome coverage Circos plots. For all the other species, see Figure 5.

(C) Genome coverage shows mapping in 1,000-bp bins from individual select single-cell or multiple-cell libraries to the reference coassembled species genome.

See also Figure S6.

widest range of species, described in the next section and technical details provided in transparent methods steps 4, 5, and 6: Phylogenetic and phylogenomic calculations. Owing to the random nature of the amplification bias in all but one species (*C. protostelioides*) with high GC%, to produce an improved species genome coassembly from single-cell assemblies, on average 3 single cells are necessary (Table S5); however, co-assemblies from best single- or multiple-cells genomes led to significantly improved results (Figures 4 and 6). Overall, we recovered a range of 5%–95% CEGMA from a single cell (median = 60%) and 60%–98% for coassemblies. The obtained single-cell and 10- to 100-cell genomes and coassemblies are examined for functional predictability in the final section.

### QC criteria as predictors of genome quality and EME single-cell genomics pipeline efficiency

We examined all QC criteria described in Data S2 for the potential to predict genome quality (fragmentation, completeness, functional prediction power) and process efficiency (time plus reagent and sequencing cost). We divided the pipeline QC process into three parts: preassembly, assembly, and postassembly. For

**Table 3. Genome-to-genome distance estimates for eukaryotic genera**

Fungi	Distance: Scale: 0 to 1	Distance: Scale: 0 to 1
Ascomycota interspecific distance	<i>Metschnikowia bicuspidata</i> (crab parasite)	<i>Metschnikowia fruticola</i>
<i>Metschnikowia bicuspidata</i> yeast cell (Mby) ( <i>Daphnia pulex</i> parasite used in this study)	0.191	0.168
Zoopagomycota intergeneric distance	<i>Piptocephalis cylindrospora</i>	<i>Dimargaris cristalligena</i>
<i>Dimargaris cristalligena</i>	0.163	
<i>Syncephalis pseudoplumigaleata</i>	0.140	0.047
<i>Thamnocephalis sphaerospora</i>	0.160	0.167
<i>Piptocephalis cylindrospora</i>		0.163
Chytridiomycota Intergeneric distance	<i>Cpi</i>	<i>Batrachochytrium dendrobatidis</i>
Caulochytrium protostelioides isolate (Cpi)		0.172
<i>Blyttomyces helices</i>	0.159	0.120
Interphylum distance	<i>Cpi</i> (Chytridiomycota)	<i>Batrachochytrium dendrobatidis</i> (Chytridiomycota)
<i>Rozella allomycis</i> (Cryptomycota)	1.000	0.145
Mby ( <i>Daphnia pulex</i> parasite used in this study) (Ascomycota)	0.317	0.139
<i>Dimargaris cristalligena</i> (Zoopagomycota)	0.217	0.131
<i>Thamnocephalis sphaerospora</i> (Zoopagomycota)	0.200	0.160
<i>Piptocephalis cylindrospora</i> (Zoopagomycota)	0.200	0.141
<i>Syncephalis pseudoplumigaleata</i> (Zoopagomycota)	0.199	0.127
Ciliate protists	Distance: scale: 0 to 1	
CiPr_NSBU	0.013	
CiPr_NSBW	0.013	
CiPr_NS BX	0.012	
CiPr_NS BY	0.012	
CiPr_NSCG	0.012	
CiPr_NSCA	0.012	
CiPr_Co-assembly	0.038	
Astramina rara GCA_000211355.2_ASM21135v2_genomic	0.086	
Param_tetraurelia GCF_000165425.1_ASM16542v1_genomic	0.134	
Param_biaurelia GCA_000733385.1_ASM73338v1_genomic	0.136	
Param_sexauurelia GCA_000733375.1_ASM73337v1_genomic	0.136	
Sterkiella GCA_001273305.2_ASM127330v2_genomic	0.142	
Sphaeroforma GCF_001186125.1_Spha_arctica_JP610_V1_genomic	0.158	
Reticulomyxa GCA_000512085.1_Reti_assembly1.0_genomic	0.161	
<i>Ichthyophthirius multifiliis</i> GCF_000220395.1_JCVI-IMG1-V.1_genomic	0.162	
<i>Naegleria fowleri</i> _1.0 GCA_000499105.1_genomic	0.168	
Tetra_borealis_V1 GCA_000260095.1_genomic	0.178	
Param_caudatum GCA_000715435.1_43c3d_assembly_v1_genomic	0.179	
<i>Euglena gracilis</i> GCA_001638955.1_Euglena_mito_Newbler_genomic	0.180	
Tetrahymena_thermoph GCF_000189635.1_JCVI-TTA1-2.2_genomic	0.181	
Trypanosoma GCA_000227375.1_ASM22737v1_genomic	0.190	

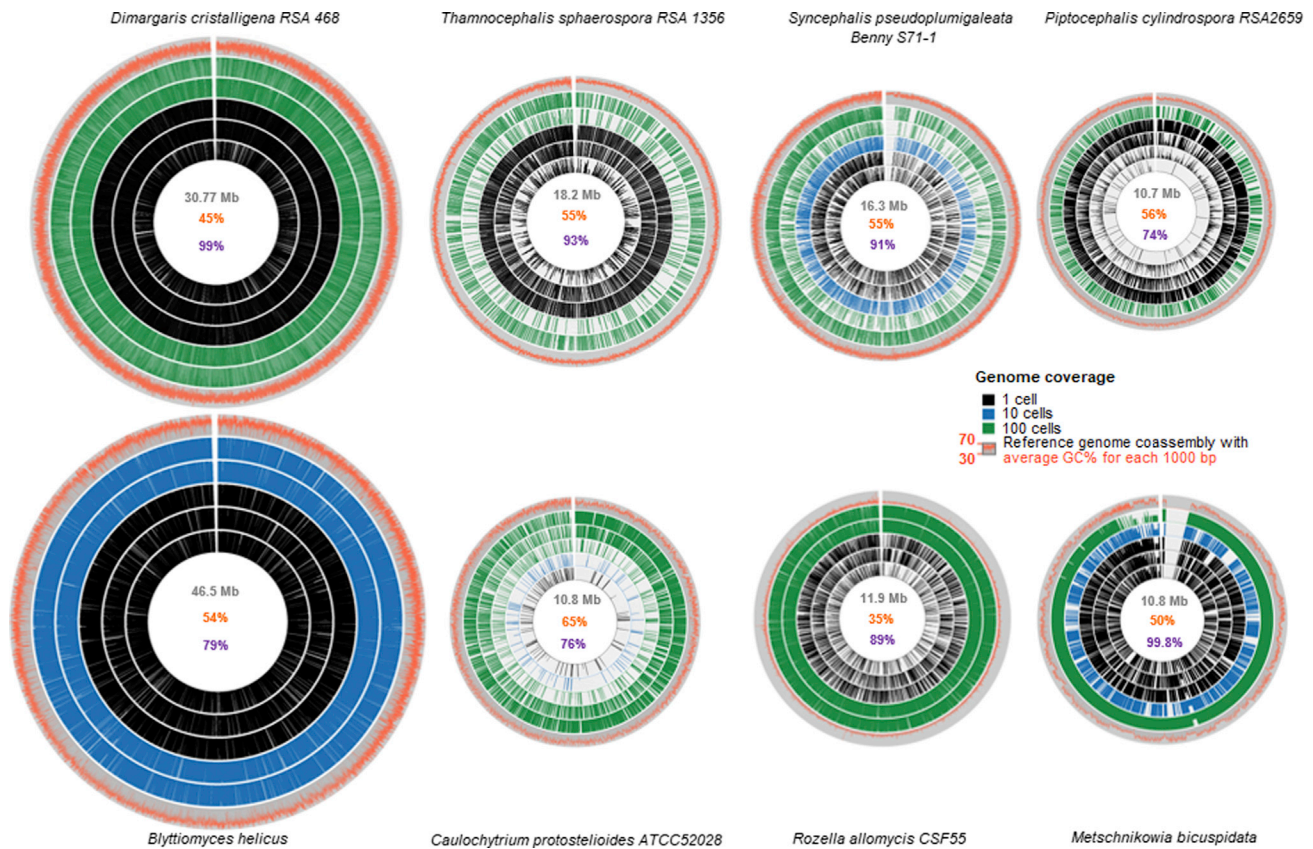
Whole-genome distance was estimated using genome-to-genome distance calculator (GGDC), formula 2 (Meier-Kolthoff et al., 2013). For the intraspecific distance estimates see Figure 6. Species names: CiPr, ciliate protist from compost followed by unique single-cell genome identifier. Ciliate genomes retrieved from NCBI have their NCBI genome identifier after species name.

each part, we eliminated redundant and non-predictive criteria and obtained a nonredundant predictive criteria list (Figure 3).

For the preassembly, QC1 and QC2 criteria were assessed for their prediction power in target single-cell recovery (number of cells versus efficiency). QC3 criteria were examined as predictors for single-cell lysis efficiency and GAB, later found to be better predicted by QC6 criteria. QC4 criteria were used for examining contamination and target single-cell OTU confirmation. We found that the QC1 and QC2 TgE criteria were critical for the prediction of target-species single-cell recovery, but these criteria were not as critical for genome quality prediction. Based on the genome recovery rate correlation with TgE, we established three success categories. In the first category, TgE1 (target concentration in original sample) or TgE2 (post FACS enrichment) estimated at  $\geq 50\%$ , guaranteed a successful and cost-efficient recovery of multiple single-cell genomes. In the second group, with TgE between 2.5% and 50%, genome recovery was reliant on morphological differences between target and nontarget cells as well as lysis-amplification efficiency. In the third group, with TgE below 2%, a wider range of factors alone or in combination (including sample volume, target cell size, shape, refractive index, viability rate, and morphological difference with nontarget cells) proved to be important for the genome recovery rate. The QC3, SGA, criterion was predictive of lysis efficiency but not amplification bias (Data S3, Figures S2 and S6). Thus the correlation between SGA and CEGMA was found to be weak, ranging from 0.1 to 0.3 depending on the species (Figure S6 and transparent methods step 3). rDNA OTU screening at QC4 or QC6 proved to be indispensable for reducing the sequencing costs by excluding nontarget genomes and genomes with contamination levels affecting genome assembly and genome completeness. See transparent methods step 4 for more details.

For the prediction of the successful outcome of the assembly phase, criteria of the QC5 (NGS library quality) and QC6 (shallow NGS read quality) steps were examined as predictors of quality of the *de novo* assembly of the amplified single-cell genomes. In addition, the QC6 criteria were explored for OTU identification of targets and of contaminants or symbionts (transparent methods step 5). The QC5 and QC6 criteria were found to be good predictors of genome quality. The QC5 library insert size reflected sequencing success with an impact on assembly quality, with reads in the range of 300–500 bp providing the best results. The RTU value of QC6 was found to be the best predictor of amplification bias, % read contaminant for process contamination, and a number of Illumina read QC metrics for genome fragmentation and CEGMA value (transparent methods step 5 and Data S2). For fungal species with genome sizes from 10 to 30 Mb, the RTU correlation with CEGMA was positive and negative with FGA (Data S2), supporting its predictive value. Conversely, protist single-cell genomes with sizes 80–110 Mb and a very high RTU value had a strong negative correlation between RTU and CEGMA, due to the low variation and high values of CEGMA for each genome. In addition, we examined the correlation between shallow NGS read GC% and CEGMA and assembled genome size (AGS). For the whole set of species, no correlation between GC% and CEGMA or between GC% and AGS was found (Data S2). Nevertheless, in five fungal species, correlation between read GC% and CEGMA was positive, and the correlation between GC% and AGS was positive in four fungal species and the ciliate. Similarly, the correlation between GC% and FGA or between GC% and SGA in QC3 was low for the whole species set (Data S2).

Based on the observed heterogeneity between species, we investigated the role of the genome GC% in amplification bias, as well as the impact of amplification bias on the completeness of genome recovery. We compared genome assembly quality from bulk isolate unamplified DNA and single- and multiple-cell-sort amplified DNA from *C. protostelioides* and *R. allomyces*. These species had similar genome sizes and lysis-MDA efficiencies but the highest (68%) and lowest (35%) GC%, respectively. Genome assemblies for the isolate unamplified DNA from *C. protostelioides* had 55%–65% higher RTU than single- and multiple-cell amplified DNA (Table S9); similarly, the isolate genome size was 50% higher and CEGMA completeness was 20% higher. For the *R. allomyces*, RTU was 2.13% lower for the unamplified isolate genome and genome size was 26% percent larger than the MDA-amplified genome coassembly (11 Mb versus 9 Mb), whereas CEGMA was similar between amplified and unamplified genomes (Table S9). In *C. protostelioides*, amplification bias correlated with genome GC% (Figure S6 and Table S6). The average genome assembly GC% was 10%–15% higher for the unamplified than MDA-amplified genomes. A close-up examination of the areas that did not amplify revealed an average GC% of 68%, a number similar to the average GC% of the unamplified DNA genome assembly and 3% higher than the amplified coassembled genome (Figure S6 and Table S9). In *R. allomyces*, the GC% for the unamplified genome and MDA-amplified genome assembly was similar (34.5%–35%) (Table S9) and very low amplification bias (Figure 5). Both



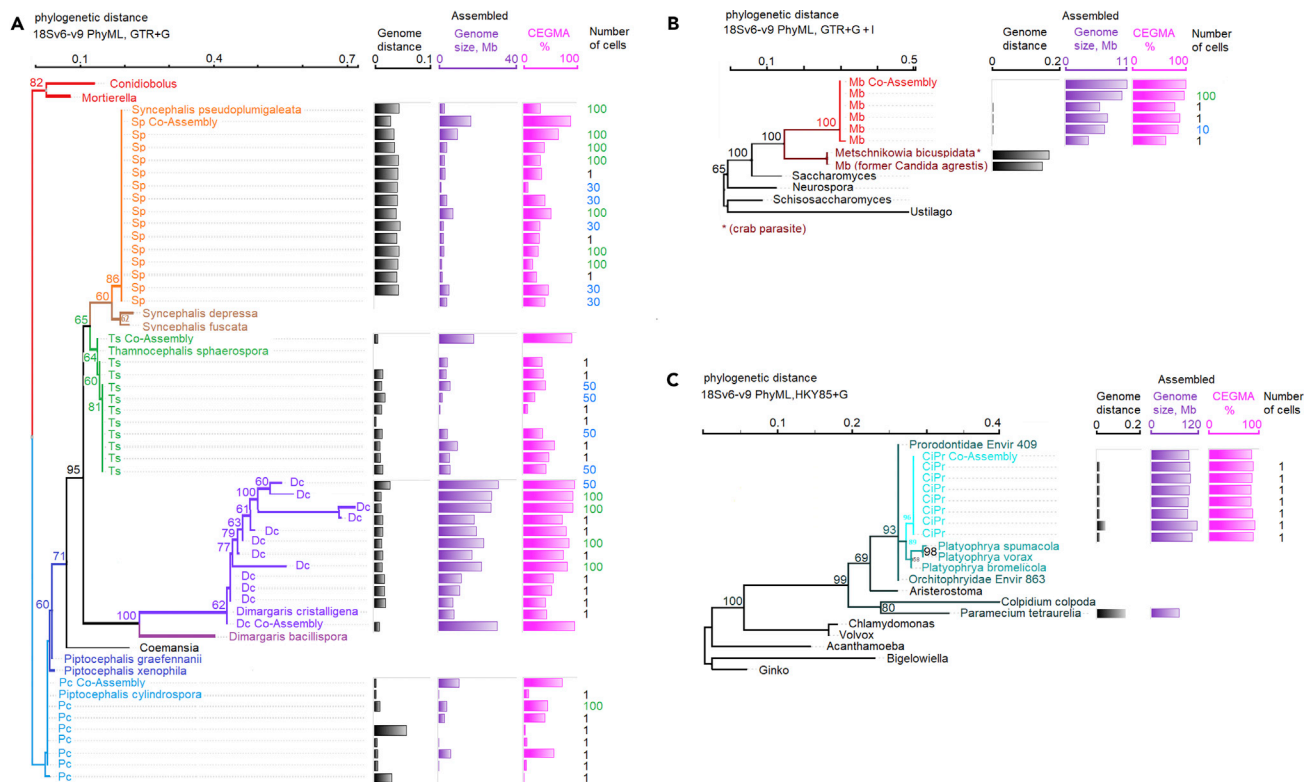
**Figure 5. Intra- and interspecific genome coverage variability**

Each species Circos map is scaled relative to the largest genome (*Blyttomyces helicus*) true to size. Reference genome (coassembly) is shown as the outer gray circle. Coassembly GC% plotted as a red line over the gray circle. 1,000-bp bins were plotted against reference co-assembly genome and scaled proportionally to the co-assembly genome size. Five representative libraries from single-cell (black); 10-cell, 30-cell, or 50-cell depending on the species (blue); and 100-cell sorts were chosen for each species. For each cell sort category one worst case, one average case, and one best case were picked when available. In the middle of the plot numbers are coassembly genome in gray, GC% in red, and CEGMA completeness in purple. See also [Figures S6](#) and [S8](#).

*C. protostelioides* and *R. allomycis* had a similar genome size, lysis-MDA efficiency, and final amount of amplified DNA, which suggests that FGA was not the cause of bias. Genome bias in *C. protostelioides* occurred during MDA amplification of the high GC regions (see the detailed description in [transparent methods](#) step 5 optimization and [Figure S6](#) and [Table S6](#)).

For the other species, we could not isolate enough pure material to make libraries from unamplified DNA. Instead, we compared single-cell genome assemblies against the coassemblies with with the largest genome completeness scores. These species had a moderate GC% (38%–55%), and the difference between coassemblies and single-cell assemblies was insignificant. We observed in five of nine species that the single-cell amplified genomes coassembly had a much higher CEGMA than the best assemblies from either single- or multiple-cell amplified genomes (30–100 cells) ([Figures 4](#), [6](#), [7](#) and [S3](#)). The average increase in CEGMA was 12% (range 0.2%–33%). In the case of *Dimargaris cristalligena*, CEGMA from a 50-cell amplified genome was already 98.7%, only 0.2% lower than that of the coassembly ([Figure 7](#)). Based on CEGMA analysis, completeness of the coassemblies or best individual assemblies for fungal genomes was estimated at an average of 91%, with a minimum of 75.5% and maximum of 98.9%; for the ciliate, the average was 94.3%. For genomes with a moderate GC% (38%–55%), FGA had the largest negative impact on genome coverage bias ([Figure 7](#)).

For the prediction of the postassembly outcome, we explored each of the QAST (QC7) criteria and fold amplification and assembled genome size ([Data S2](#)). Genome quality was measured by CEGMA ([Figures 4](#), [6](#) and [7](#)) and the continuity of the assembled genomes by mapping each individual assembly to the reference genome shown in [Figure 5](#). We found that the QAST criteria MGS\_N50 and MGS\_L50 directly



**Figure 6. Intraspecific single-cell genome variability and phylogenetic placement of single- and multi-cell genomes**

Phylogenetic distance was estimated based on the 18S rDNA region v6 through v9 using PhyML package (Guindon et al., 2010). Genome distance was estimated using Genome-to-Genome Distance Calculator (GGDC), formula 2 (Meier-Kolthoff et al., 2013). See Table 3 for Genome-to-Genome Distance between genera.

(A) Zoopagomycota phylogenetic tree and GGDC, genome size, and completeness. Best nucleotide substitution model estimated HKY85, random starting tree, estimated best tree with bootstrap analysis, bootstrap shown values above 60%. Tree: Branches are shown as: Dc, *Dimargaris cristalligena* RSA 468; Ts, *Thamnocephalis sphaerospora*, Sp, *Syncephalis pseudoplumigaleata*; Pc, *Piptocephalis cylindrospora* RSA2659.

(B) Ascomycota single-cell phylogenetic tree and GGDC, genome size, and completeness. Best nucleotide substitution model estimated GTR+G+I, random starting tree, estimated best tree with bootstrap analysis, bootstrap shown values above 60%. Tree: Branches are shown as Mb, *Metschnikowia bicuspidata* in red with closest species in dark red.

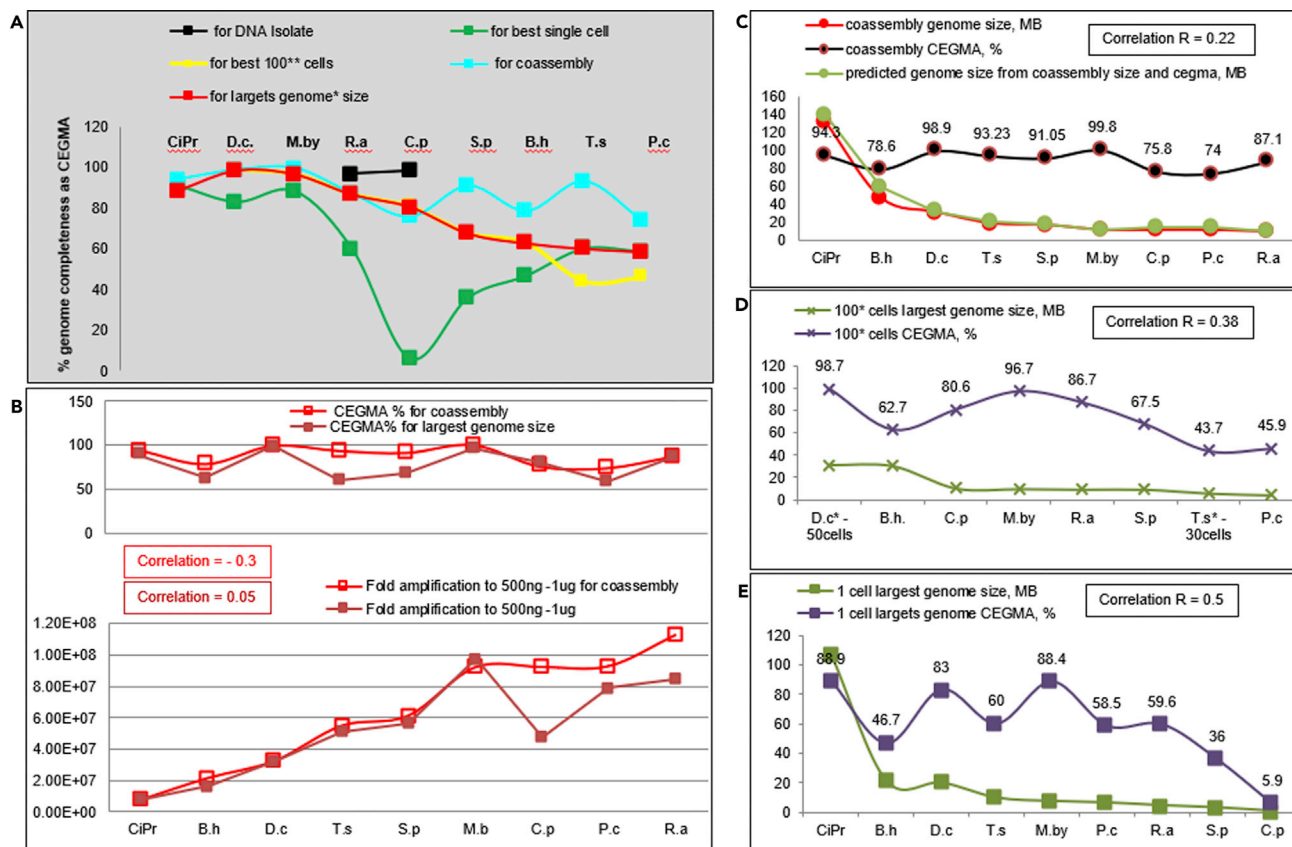
(C) Compost ciliate single-cell phylogenetic tree and GGDC, genome size and completeness. Best nucleotide substitution model estimated HKY85+G, with bootstrap analysis, shown values above 60%. Branches for the single-cells are shown with: aqua, ciliate Protist (CiPr). Species with closest 18S are shown in dark teal. Non-Alveolata branches are shown in black.

correlate with the CEGMA value (Data S2 and Tables S2–S4). The correlation between AGS and CEGMA within and between each species was positive, supported by the negative correlation between FGA and CEGMA (Data S2). However, we found a weak correlation between these values in the best single-cell or coassembled genomes across all species (Data S2). The highest genome fragmentation bias was observed in genomes with the highest amplification bias and did not correlate directly with any criterion. In summary, the pass and fail value of the QC criteria for *de novo* single-species genome assembly is shown in Table S10.

### Intra- and interspecific variabilities in single-cell genomics and the impact on genome distance estimates for species coassembly

Comparative analysis of single-cell genome variability within and between studied species revealed that variability between species was higher than within species (Figures 4, 5, and 6). Among all impact factors, we found that high GC% (>63%) (Figure S6), poor lysis of cell walls, and small genome size each provided a basis for amplification bias and variability in genome coverage between single-cells (Figures 4, 5, and 6).

For example, species with similar, good lysis efficiency and GC%, but smaller genome size experienced higher GAB (Figure 5). Similarly, higher GC% in species with the same size and lysis efficiency led to higher genome fragmentation and poor recovery (Figure 4). Species with medium GC% and genome size, but



**Figure 7. Amplification bias for coassemblies and largest assembled genomes (from 1 or 100 cells)**

(A) Genome completeness as CEGMA across species for best genomes.

(B) Correlation between fold amplification and CEGMA for coassembly and for largest genome.

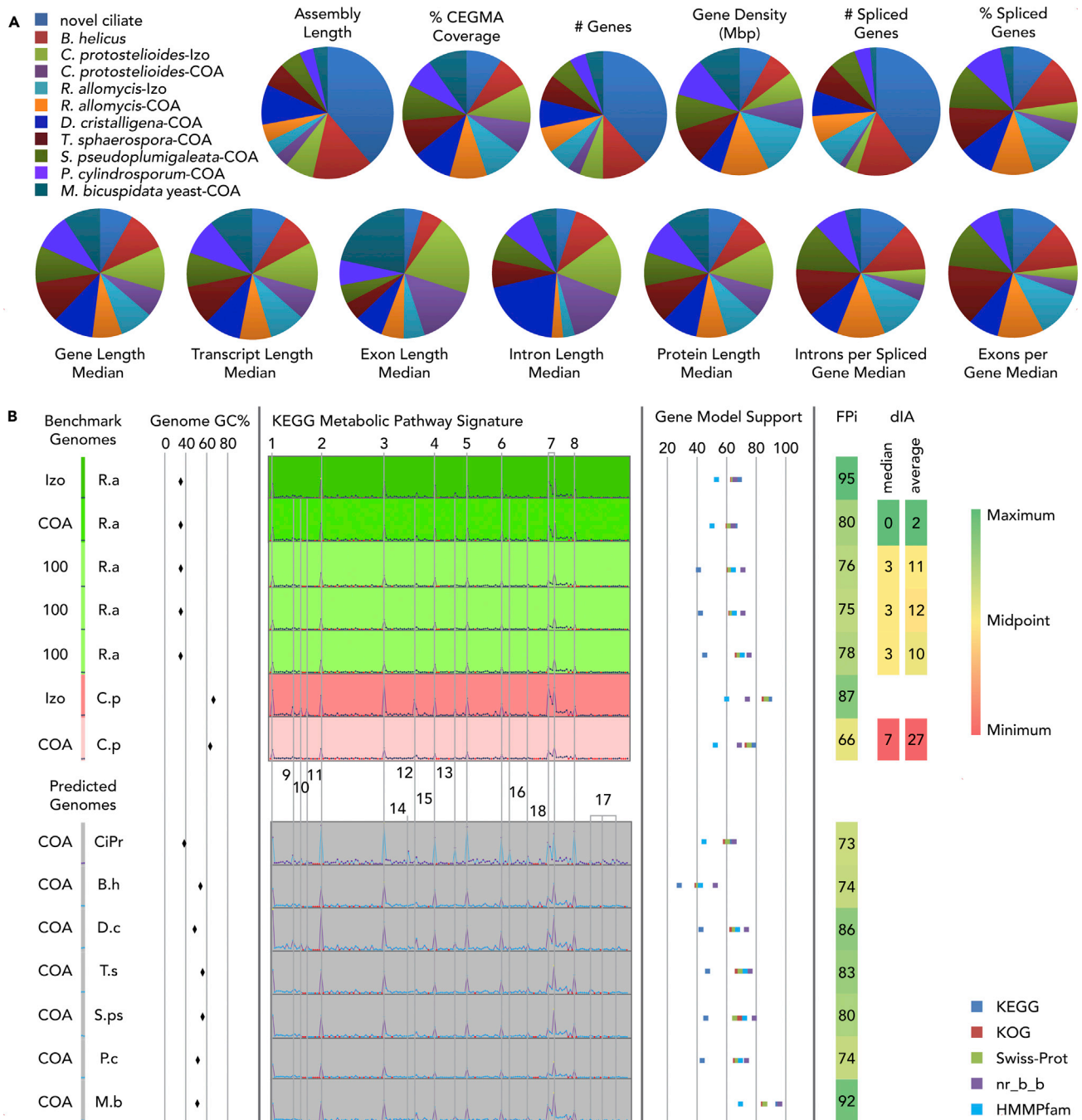
(C–E) Correlation between assembly size and CEGMA. Correlation is Pearson value (R) for genome size and CEGMA. (C) For coassembly. (D) For 100-cells genomes where available. (E) For single-cell genomes.

See also [Figure S4](#).

poor lysis efficiency (*M. bicuspidata*) had fewer genomes recovered, but less GAB (Figures 5 and 6). Intra-specific variability was observed mostly between assemblies from different cell sorts (1 versus 10 versus 100) and supported a random intraspecific variation in genome coverage. Respectively, assemblies from multiple-cell libraries had higher CEGMA and lower fragmentation than did those from single-cell libraries. The observed intraspecific random variability had a beneficial impact on coassembly quality, resulting in a much higher species genome quality (Figures 4, 5, 6, and 8). However, for *C. protostelioides*, the variability between multiple-cell (100-cell) assemblies was as high as the observed variability between single-cell assemblies in other species (Figures 4, 5, and 6 and Table S9).

A requirement for creating species-level genome assemblies from multiple individual genome assemblies is determining the phylogenetic identity of individual genomes before coassembly. We compared the use of rDNA, ANI, and GGDC for intraspecific distance estimation for the purpose of creating species coassemblies (Figures 4 and 6 and Tables S3 and S8; for detailed description see [transparent methods](#) step 6 Optimization and steps 4, 5, and 6: Phylogenetic and phylogenomic calculations). Our tests found that GGDC formula 2 (Auch et al., 2010; Meier-Kolthoff et al., 2013) was the most reliable tool for establishing intraspecific distance for incomplete genomes obtained from EME single-cell amplified DNA, even though it was originally designed and tested only on unamplified DNA bulk isolate genomes. The most striking proof of formula 2's ability to correctly predict the genome distance for this group of eukaryotes is correctly placed *C. protostelioides* single-cell and 10-cell genomes, which had the same intraspecific distance as other species' genomes despite having much lower CEGMA values (Figures 4 and S8). Our results showed that the minimum CEGMA value necessary for accurate genome-to-genome distance prediction was lower by 10%,





**Figure 8. Single-cell genome coassembly quality assessment for functional genomics studies**

(A) Comparative analysis of annotated genomes. See also Table S7.

(B) Functional prediction value assessment. Izo, isolate unamplified genome assembly; COA, coassembly of several single- and/or multiple-cell assemblies, and 100-100 cell-sort genome assembly. For the fungi, the scale for the KEGG metabolic pathway signature was the same (0–678). For the CiPr (ciliate protist), the scale was 0–1252. FPI – functional genomics prediction index, where  $i$  = geomean % complete genes and % CEGMA coverage; dIA, absent entries in the amplified genome compared with the isolate genome. Abbreviations for species names are explained in Table 1. Detailed information for each KEGG entry is available in Table S9. KEGG peak numbers: 1 through 8 are category total, 9 through 18 are specific enrichments in subgroups of the respective category. 1. Amino acid metabolism, 2. Biosynthesis of secondary metabolites, 3. Carbohydrate metabolism, 4. Glycan biosynthesis and metabolism, 5. Lipid metabolism, 6. Metabolism of cofactors and vitamins, 7. Nucleotide metabolism and overview of biosynthesis of alkaloids and hormones, 8. Xenobiotic

**Figure 8. Continued**

biodegradation and metabolism, 9. Tryptophan metabolism, 10. Biosynthesis of polyketides and nonribosomal peptides, 11. Biosynthesis of siderophore group nonribosomal peptides, 12. Starch and sucrose metabolism, 13. Lipopolysaccharide biosynthesis, 14. Pentose phosphate pathway, 15. Energy metabolism, 16. Nicotinate and nicotinamide metabolism (cyt p450), 17. Benzoate degradation via CoA ligation, drug metabolism cytochrome p450, gamma-hexachlorocyclohexane degradation, metabolism of xenobiotics by cytochrome p450, 18. Metabolism of other amino acids.

from the original publication (Auch et al., 2010; Meier-Kolthoff et al., 2013). Our GGDC formula 2 estimates were supported by phylogenomic placement of the *C. protostelioides* single-cell genome Figure S8 and discussed further in Ahrendt et al. (2018). We observed no negative impact of single-cell genome quality variability on genome distance estimates.

**Genomes obtained via single-cell amplification are suitable for functional analyses**

To determine if the assemblies produced by this pipeline were suitable for functional analysis, we annotated them and performed a comparative structural analysis and a functional prediction analysis (for technical details see transparent methods step 7). We annotated the best single- and multiple-cell assemblies and coassembled fungal and protist genomes using the MycoCosm annotation pipeline with manual curation when necessary (Grigoriev et al., 2014).

A comparative analysis of the genome structure, e.g., number of genes, gene density and intron/exon structure, and transcript and protein length, for each species revealed an average 372 genes/Mbp density, with small variability for the analyzed set of genomes (Table S7 and Figure 8). Highest gene density was observed in *R. allomycis* (497 and 535 gwnws/Mbp for coassembly and isolate, respectively). The lowest gene density was seen in *D. cristalligena* (242 genes/Mbp), the second largest fungal genome with the greatest intron length (Figure 8). Both are mycoparasites. The highest number of genes (12,167) was in the largest fungal genome *Blyttiomycetes helicus*, whereas the average among fungi was 6,422 genes. The seven ciliate protist single-cell genomes were 2.5×–10× larger than the spread of fungal genomes used for the study (Figure 8 and Table S7). As expected, for this ciliate the number of genes was much larger adding up to 40,072 gene models (Tables S7 and S9). Nevertheless, gene density (331.5) was below the fungal average with slightly smaller transcript (880 bp), exon (144 bp), intron (49 bp), and protein (282aa) median length (Table S7). For fungi, average transcript length was 1,101 bp with little fluctuation between species. Average of the exon and intron length were 297 and 93 bp, respectively, ranging between 675 and 141 bp for exon median; intron median varied between 29 and 197 bp (Table S7). The longest exon median was observed in the smallest fungal genome, the yeast *Metschnikowia bicuspidata*. In the second and third largest fungal genomes (*D. cristalligena*, *C. protostelioides*) the intron length was the first and second highest and exon was the third and second highest. Most of the EDF had, on average, 75.5% spliced transcripts, and the median per spliced gene for introns and exons was 2.7% and 2.8%, respectively. *Caulochytrium* made an exception and was more alike to the yeast *Metschnikowia* with only 38% and 24% spliced genes for each species with median one intron and exon per spliced genes. The ciliate and five of the fungi had equal and the highest numbers of introns per spliced gene, and four of these fungi and the ciliate had equal but fewer exons per gene. The percent of the spliced genes was similar between the ciliate and the five fungi, whereas the number of spliced genes was significantly smaller in fungi, somewhat correlating with their genome size. In spite of the aforementioned variability, protein length median average was 305 aa, with little variation between species.

We used the two fungal species unamplified “isolate” genomes (Cpi: *C. protostelioides* isolate, and Rai: *R. allomycis* isolate) to benchmark and evaluate the amplified single- and multiple-cell genomes, lacking isolate references, for suitability for comparative genomic analysis. We observed that both *C. protostelioides* and *R. allomycis* had a significant percentage of incomplete genes in the coassembly and single- or multiple-cell assemblies relative to their isolate genome (Tables S7 and S9). However, coassembly and multiple- or single-cell assemblies had fewer genes and exon, transcript, and protein lengths for *C. protostelioides* but not for *R. allomycis* (Tables S7 and S9). On the other hand, the gene density was the same in the isolated genome and the coassembly genome for *C. protostelioides* and differed between those for *R. allomycis* (Figure 8 and Table S7). This variation perhaps influenced the estimation of the % spliced genes between isolates and coassemblies.

Gene model support showed that for the fungi and ciliate protists, on average 60.2% and 65% of models had homology to KEGG database proteins and 68% and 61% to Swiss-Prot proteins, respectively (Figure 8 and Table S9). The highest number of supported models was observed in *Metschnikowia* (95% KEGG and

86% Swiss-Prot) and the lowest in *Blyttiomycetes helicus* (28% KEGGs and 42% Swiss-Prot) (Figure 8). In fungi the number of complete genes (from start codon to stop codon) was higher than in ciliate protists (average 74% and 56%, respectively). However, *Caulochytrium* had the lowest number of complete genes for the amplified genome (55%) despite similar number of hidden Markov models-supported Pfam with *Rozella*, *Blyttiomycetes*, and the ciliate protists (Table S9 and Figure 8).

We used the KEGG database, gene model support and completeness, and CEGMA values (Table S9) to create a functional genomics prediction index (FPI) for this purpose (Figure 8 and Table S9). Furthermore, we assessed the quality of the coassemblies relative to the isolate annotation for the benchmarked genomes using the presence/absence score (dIA) for each KEGG category. The dIA was calculated by subtracting the number of models for each KEGG -EC (enzyme commission) pathway map of the amplified genome from the isolated genome. An average and a mean value were obtained for the entire set of KEGG ECs for this calculation (Table S9).

Our analysis of the functional genomics prediction power, using the FPI and dIA, showed (Figure 8B) that a conservative 80% cutoff for FPI value characterizes a genome nearly identical to the isolated genome in terms of KEGG values (profile and dIA). We observed a significant alteration of the KEGG profile and dIA values when FPI dropped to 66% or less. As shown in Figure 8, *Rozella* coassembly and multiple-cell assemblies resulted in annotations that produced a significantly lower dIA score compared with the *Caulochytrium* coassembly, supporting our FPI score cutoff. We observed that when the FPI value was between 70% and 80%, KEGG gene counts were reduced without changing the pattern qualitatively (e.g., presence-absence) (Figure 8 and Table S9). For example, for the most biased amplified genome (*C. protostelioides*), the KEGG number of genes from the coassembly was significantly lower than in the isolate, making KEGG inaccurate for gene expansion-reduction analysis. In contrast, for *Rozella*, the KEGG pattern was nearly identical between the coassembly and isolate. In *Rozella*, nearly identical numbers between the coassembly and isolate for KEGG were accompanied by similar CEGMA completeness and low GC% (34%), as well as the number of complete genes in *Rozella* was higher than that in the *Caulochytrium* coassembly (67% versus 54%).

We found that the combination of <60% complete genes, lower than 90% CEGMA, and high average GC%, as observed for the coassembly of *Caulochytrium*, led to lower-than-acceptable scores for reliable quantitative functional predictions (at least for KEGG-based gene counts).

Four fungal single-cell genome coassemblies without an isolate (e.g., *Metschnikowia*, *Dimargaris*, *Thamnocephalis*, and *Syncephalis*) had an FPI score above 80%. They had nearly complete CEGMA values and average GC%, further supporting FPI-inferred predictions for highly accurate KEGG profiles. In the three other species, *Piptocephalis*, *Blyttiomycetes*, and the ciliate protist, a lower FPI (70%–80%) and either a lower CEGMA value (fungi) or lower percent of complete genes (protist) predicted the identified KEGG numbers to be underrepresented. However, given that the FPI for these three species was within 10% of the reliable interval (70%–80%), qualitative but not quantitative analysis would reflect the same functional predictions as their respective isolated genomes.

As an example of the types of functional analyses possible with our genomes, we investigated expansions of KEGG groups. In two fungal species (*Dimargaris* and *Caulochytrium*) and the ciliate, we found expansions in tryptophan metabolism, biosynthesis of polyketides and nonribosomal peptides, and biosynthesis of siderophore group nonribosomal peptides (Figure 8B, peaks 9, 10, and 11, respectively). In addition, *Caulochytrium* showed an expansion in the starch and sucrose metabolism category, whereas several other fungi had an increased count of genes from the energy metabolism category. An in-depth examination of the gene expansion and losses and functional implications for the lifestyles of these single-cell-derived eight fungal genomes was given in our recent article (Ahrendt et al., 2018).

Unlike fungi, ciliate protist genome displayed an expansion in the number of genes involved in xenobiotic biodegradation and metabolism, specifically benzoate degradation via CoA ligation, drug metabolism-cytochrome P450, gamma-hexachlorocyclohexane degradation, and metabolism of xenobiotics by cytochrome P450 (Figures 8B and Table S9). The ciliate displays a unique expansion in the pentose phosphate pathway and the lipopolysaccharide biosynthesis categories. An in-depth functional analysis of the ciliate protist comparative genomics in the context of a group of 180 species was developed in a separate manuscript by Ciobanu et al., (in preparation).

## DISCUSSION

In this article we review, test, and optimize single-cell genomics approaches for studying microbial eukaryotes in their natural environment. As a result, we developed a single-cell pipeline for mining the genomes of EMEs by combining known, optimized, and novel wet-bench approaches with bioinformatics tools. Our results empowered large-scale functional genomics studies that shed light on the ecological niches of EMEs and uncovered their functional plasticity (Ahrendt et al., 2018 and submission ciliate).

We benchmarked the pipeline against two fungal isolate genomes (*R. allomyces* and *C. protostelioides*) using a set of eight known fungal species, three environmental samples with unknown EME species, and a set of 20 quality check criteria. We evaluated (1) the steps that had the largest impact on genome completeness, (2) the criteria that were most predictive of the best genome quality and completeness, and (3) the correlation between genome completeness and each evaluation criterion used. As a proof of concept, we used this pipeline for genome recovery of unknown EME Chromista/SAR (Ciliophora) and fungi (Cryptomycota and Chytridiomycota) species. Neither the rDNA nor the genome of the ciliate could be recovered using standard metagenome sequencing and assembly methods in previous studies (Eichorst et al., 2013; Luo et al., 2012). The fungal EME samples had an rDNA phyla profile but were microscopically undistinguishable from other non-target organisms and were at an ultra-low concentration. Complex genomes from novel lineages that are subjected to partial chimerization during amplification require correct bioinformatics tools for genome assembly, genome-to-genome similarity estimation, genome completeness assessment, correct gene structure prediction, and annotation from fragmented and/or incomplete genomes. We tested a number of bioinformatics tools (Auch et al., 2010; Bankevich et al., 2012; Butler et al., 2008; Han et al., 2016; Meier-Kolthoff et al., 2013; Peng et al., 2012) for this purpose and, when necessary, adjusted the most suitable to achieve the aforementioned results.

The pipeline allowed recovery of high-quality genomes from individual cells with a CEGMA median genome completeness of 60% (range 5%–90%). We found that the EME target single-cell recovery rate and genome quality increased when an enrichment FACS step was performed before single-cell isolation, during presequencing part of the pipeline. For sequencing, one simple but critical improvement was the implementation of a shallow sequencing step before the deep sequencing step required for *de novo* assembly. Shallow sequencing for amplification bias estimation has been proposed previously (Daley and Smith, 2014). Our pipeline, unlike the suggested method in Daley and Smith (2014), used the read-QC-pipeline RTU metric to estimate amplification bias as well as the contamination level and a number of other parameters indicative of the sequence quality. We found that the best predictor for amplification bias and the highest measure of genome completeness out of all criteria used during this step was RTU. In addition, we tested the use of shallow sequencing reads for rDNA assembly followed by OTU screening. This eliminated issues related to PCR primer bias (Lazarus et al., 2017) or Sanger sequencing and allowed us to identify cases where symbiotic organisms were present despite any issues of rDNA divergence or contaminating OTUs. Several of the commonly used criteria (Gurevich et al., 2013) describing genome assembly quality (number of scaffolds in the range of 2–10 kb, 10–25 kb, and 25–50 kb, main genome scaffold\_N50) correlated well with CEGMA and genome size.

We identified critical factors affecting the EME recovery rate using a range of QC and optimization steps. The main bottleneck for successful enrichment was a combination of extreme cell size and cell shape (narrow elliptical,  $2 \times 50 \mu\text{M}$ ) along with a low sample volume (1–2 mL) or when nontarget organisms with similar morphology to the target were present at a higher concentration in the sample (e.g., more Cryptophyta (flagellate algae) than Cryptomycota (flagellate fungi), along with other organisms with the same size). Poor lysis efficiency was found to affect the number but not the quality of amplified target genomes (e.g., *M. bicuspidata*, *B. helicus*, *T. sphaerospora*, and *S. pseudoplumigaleata*). The most striking supporting example was zoospores of *C. protostelioides*, where despite the high number of successfully lysed and amplified cells, we observed the highest amplification bias, likely due to the higher-than-average GC (68%). In contrast, *M. bicuspidata* with yeast cell walls was lysed at a significantly lower percentage but had 90% higher CEGMA values for single-cell amplified genomes and 30% higher values for 100-cell sorting than *C. protostelioides* zoospores. Both species had a small genome, 13 and 11 Mbp, but *M. bicuspidata* had 10%–15% lower GC% than *C. protostelioides*, contributing toward resulting genome quality. Although a universally efficient method for opening cell walls in a single-cell reaction remains to be found for maximizing the number of EMEs recovered from environmental samples, our work found that the main culprit for lower genome quality was amplification bias. Amplification bias in smaller genomes has been reported

previously (Dean et al., 2002; Gawad et al., 2016 and references), but the mildly high (68%) GC of the genome causing extreme MDA bias was unexpected. Several articles have indicated a mild bias for the MDA reactions caused by higher GC% (Garvin et al., 2015; Xu et al., 2014), whereas others did not find a similar correlation (Ellegaard et al., 2013). Given the strand displacement ability of the phi29 DNA polymerase that allows unwinding of DNA without nucleotide bias, it is possible that structural features (e.g., the presence of more regulatory or chromatin organization protein complexes in the higher GC% regions) were the true cause of the amplification bias observed in *C. protostelioides*. The genomes with an average GC% of 35%–55% and larger genomes had low amplification bias and the highest completeness.

Overall, we looked at 20 potentially predictive criteria for the pipeline outcome. Majority of criteria were highly predictive, and we reduced the redundant ones. Several were weakly predictive: The SGA and FGA (reflecting the duration of amplification and DNA amount) were weakly inversely correlated with genome completeness (assessed by CEGMA); the strength of predictability of the latter two was lower than that of the RTU. Although RTU was a good predictor of genome quality for most species, we observed that for genome size >60 Mb with RTU >80%, the RTU did not correlate with genome completeness, thus setting up the highest threshold for the RTU predictability. The largest amplification bias was observed for genomes ranging in size between 10 and 30 Mb. These genomes would benefit the most from significantly reduced amplification times. For larger genomes (e.g., 30–50 Mb), this trend was also observed but was not as strong. Supporting this observation, for the seven >100-Mb protist genomes, the RTU was very high and not correlated with CEGMA. For a generalized prediction, we speculate that the RTU reflects amplification bias for genome sizes smaller than 60 Mb using our MDA protocol (3–4 h amplification), whereas for genomes larger than 60 Mb, this is not the case. Based on this study, we suggest that for unicellular protists and genomes with size >60 Mb, a larger study is necessary to include a broader spectrum of genome average GC% to understand the impact of GC% on amplification bias.

To streamline our QC process, we reduced the set of 20 criteria to six that span the preassembly, assembly, and post-assembly pipeline steps. Based on the set of criteria with the highest predictive value, we found that EMEs with genome sizes ranging from 10 to 30 Mb clustered around 40% genome completeness. This group was the largest of the dataset and had the most statistical support (70%) for genome completeness prediction power. Genomes from 30 to 60 Mb clustered at approximately 70% genome completeness, and genomes larger than 100 Mb clustered at approximately 90% completeness. However, the proportion of organisms with larger genomes was smaller in our dataset, and therefore, the statistical support for genome completeness prediction power was lower (25%) for them (Figure 3). Several outliers supported the observation that specific combinations of factors could affect genome completeness. (1) The yeast *M. bicuspidata* had an extremely high CEGMA value despite its small genome size and poor lysis, perhaps due to low GC%. (2) *B. helicus* and (3) *D. cristalligena* are partial outlier species from the other EDF. Both of them, regardless of the poor start of amplification, produced the largest fungal genome sizes of the dataset with very high CEGMA values. We conclude that for smaller genomes, reducing genome amplification time paired with amplified Illumina libraries would improve genome recovery.

It has been shown (Kogawa et al., 2018) and we confirm here that completion of a species genome could be achieved via coassembly of individual genomes of the same species. A prerogative for species-specific coassembly for environmental single-cell genomics is establishing the correct genome-to-genome taxonomic distance. No tools for EME genome-to-genome similarity have been evaluated to date. We found that the best tool for intraspecific and intragenic genome-to-genome distance calculation was the GGDC from DSMZ (Auch et al., 2010; Meier-Kolthoff et al., 2013). Originally developed for unamplified prokaryotic genomes, this tool set provides several formulas suitable for various levels of genome completeness. Our tests showed that formula 2 was highly suitable for amplified eukaryotic genomes. Several authors (Auch et al., 2010; Meier-Kolthoff et al., 2013) have reported that this formula performed well with prokaryotic genomes with as low as 20% completeness, and in our study, we found that formula 2 performed accurately with amplified single-cell fungal genomes with as low as 5.9% completeness. We found that, for the most part, the quality of the coassemblies and, in some cases, single-cell as well as 10- to 100-cell genomes was comparable to that of unamplified genomes (derived from millions of cells) and could be used for comparative genomics.

Reaching genome completeness for single-cell genomes whose quality is close to that of isolated DNA genomes empowers comparative genomics studies to make meaningful functional predictions. Using our

benchmarking species with isolated genomes, we proposed a new criterion called FPI to be able to estimate functional prediction value. This criterion is based on three separate genome quality standards for *de novo* assembled single-cell genomes. We conclude that for EME single-cell genomes that suffered serious amplification bias, it was necessary to coassemble multiple partial genomes to achieve a meaningful functional prediction score. We found that for genomes with high amplification bias, coassembling three 100-cell sorts was necessary; for moderate amplification bias, three to five single-cell sorts were sufficient; and for those with low amplification bias, single-cell assemblies were sufficient. The example of a broad and in-depth analysis of the functional value of single-cell genomics based on this fungal set of genomes was presented in our complementary article (Ahrendt et al., 2018).

A challenge not discussed so far anywhere is the costs associated with the high sequencing capacity necessary for EMEs. Given that EMEs are significantly underrepresented in the environment compared with their prokaryotic neighbors (Wurzbacher et al., 2017), mining EMEs with existing methods would require significantly more resources, starting with sample volumes and ending with computational resources, including all the steps in between discussed in this article. For example, unlike environmental prokaryotes with genomes 100-fold smaller than those of eukaryotes on average, mass sequencing of poor-quality genomes or nontarget genomes for EMEs can become prohibitively costly, even on NovaSeq when attempting broader and deeper phylogenomic mining. Therefore, establishing methods for evaluating the quality of the genome before deep sequencing is critical for affordable high-throughput EME genomics. We provide several optimizations that allow affordable exploration of EME genomics in the context of their ecological niche. We provide analytical methods to explore factors and predictors of successful genome recovery across broad-spectrum taxa of diverse genome size, GC%, cell wall composition, and phylogenetic origin. Following the general idea of using shallow sequencing as a prediction tool at an earlier step, we found a new highly predictive metric for GAB and made several highly effective changes to the amplification and screening process of single-cell genomes before the deep sequencing step, which allowed us to reduce costs and significantly improve the genome quality of the EMEs. For example, just implementing the screening after the shallow sequencing step, using HighSeq Illumina technology for deep sequencing, savings are 10-fold for smaller genomes or more for larger genomes in this study; for NovaSeq Illumina technology savings are less, but still over 10-fold for larger genomes. Considering that for a good genome *de novo* assembly from amplified reads we needed 50 million reads or more for 60-Mb genomes, blind sequencing on NovaSeq a full MDA plate (288 cells, or 60% of this, if taking only positive MDAs) is still costly, whereas implementing rDNA-OTU, RTU, and contaminant screening step reduced the high-quality targets to 4–6 per plate and on average 30 lower quality targets per plate. Although in this study we did not analyze the non-target EME genomes from the three environmental samples, in a broader targeted study the outcome of the pipeline would be higher. Another application of this pipeline, especially the use of shallow sequencing step rDNA or other marker genes assembly, would be larger phylogenetic studies of the tightly associated community of the target EMEs. Thus, our study offers an avenue to increase the resolution of microbial communities and allow for functional prediction of the role of EMEs in their environment at the next level of depth and breadth.

### Limitations of the study

We established the target single-cell EME genome recovery limits for this pipeline. The most critical limitation is the ultra-low concentration of the target EME and the same size of the most abundant non-target prokaryote and eukaryote organisms in the same sample. The second most impactful limitation is caused by amplification bias of the high GC% genomes (here above 65%). The third less impactful limitation is amplification bias caused by fold amplification of the small genomes, which can be reduced by overall reduction of the amplification time in step 3 for all EME genomes, regardless of the knowledge about their genome size. The fourth limitation is the organisms that have a size exceeding 100  $\mu\text{m}$  diameter due to the FACS limitation. This can be managed by replacing FACS with LCM or micromanipulation, but reduces the high-throughput aspect.

### Resource availability

#### Lead contact

Doina Ciobanu, [dgociobanu@lbl.gov](mailto:dgociobanu@lbl.gov).

#### Materials availability

No new unique reagents were generated in this study. Reagents sources are listed in [transparent methods](#).

### Data and code availability

The coassembled genomes and annotations of the target species reported in this paper are available through "MycoCosm:<https://genome.jgi.doe.gov/fungi>" using the following "MycoCosm:URLs" and "GenBank: accession numbers" reported in this paper are respectively: *R. allomyces* CSF55 single-cell "[https://genome.jgi.doe.gov/Rozal\\_SC1](https://genome.jgi.doe.gov/Rozal_SC1)"; "GenBank: QUVT000000000", *B. helicus* Perch Fen single-cell "<https://genome.jgi.doe.gov/Blyhe1>"; "GenBank:QPFV000000000", *C. protostelioides* ATCC 52028 single-cell "[https://genome.jgi.doe.gov/Caupr\\_SCcomb](https://genome.jgi.doe.gov/Caupr_SCcomb)"; "GenBank:QUVS000000000", *D. cristalligena* RSA 468 single-cell "<https://genome.jgi.doe.gov/DimcrSC1>"; "GenBank:QRFA000000000", *P. cylindrospora* RSA 2659 single-cell "[https://genome.jgi.doe.gov/Pipcy3\\_1](https://genome.jgi.doe.gov/Pipcy3_1)"; "GenBank:QPFT000000000", *T. sphaerospora* RSA 1356 single-cell "<https://genome.jgi.doe.gov/Thasp1>"; "GenBank:QUVU000000000", *S. pseudoplumigaleata* Benny S71-1 single-cell "<https://genome.jgi.doe.gov/Synps1>"; "GenBank:-QUVV000000000" and *M. bicuspidata* single-cell "[https://genome.jgi.doe.gov/Metbi\\_SCcomb](https://genome.jgi.doe.gov/Metbi_SCcomb)"; "GenBank:QUVR000000000". The whole-genome sequence for the non-single-cell isolate *C. protostelioides* ATCC 52028 is available through "MycoCosm:<https://genome.jgi.doe.gov/Caupr1>" and "GenBank: QAJV000000000". The whole-genome sequences for the non-single-cell isolate of *R. allomyces* CSF55 was not determined in this study and is available through "MycoCosm:[https://genome.jgi.doe.gov/Rozal1\\_1](https://genome.jgi.doe.gov/Rozal1_1)") and "GenBank:ATJD000000000".

## METHODS

All methods can be found in the accompanying [transparent methods supplemental file](#).

## SUPPLEMENTAL INFORMATION

Supplemental information can be found online at <https://doi.org/10.1016/j.isci.2021.102290>.

## ACKNOWLEDGMENTS

We would like to thank Meghan Dufy for the *Metschnikowia* sample; Joyce E. Longcore for the *Blyttomyces helicus* sample; Christopher Daum and Mathew Zane for providing valuable information for sequencing optimization; Eugene Goltsman for valuable discussion and suggestions for data analysis and visualization; Catherine Adam and Yi Peng Lee for library qPCR; Laura Sandor, Jenifer Kaplan, and Jennifer Chiniquy for running Illumina and PacBio sequencing instruments at JGI during this project; Bryce Foster for running the JGI RQC pipeline; and the NERSC facility for providing excellent service.

The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. S.R.A., I.V.G., T.Y.J., and C.A.Q. are supported by NSF grant DEB-1354625. G.L.B., M.E.S., and T.Y.J. were supported by NSF grant DEB-1441677.

## AUTHOR CONTRIBUTIONS

D.C. wrote the manuscript with valuable input from T.Y.J., T.W., I.V.G., A. Copeland, S.A., and C.A.Q.

T.Y.J. and I.V.G. initiated this study.

J.-F.C., D.C., T.Y.J., and I.V.G. designed the study.

J.-F.C., D.C., I.V.G., T.Y.J., A. Copeland, and A. Clum coordinated the implementation of pipeline segments into one process.

D.C. and S.C. designed and performed wet-bench protocols and pipeline evaluation.

A. Clum, A. Copeland, W.B.A., S.A., and B.F. designed and performed genome assembly tool testing, optimizations, final assemblies, and assembly QC.

W.A. and C.A.Q. performed rDNA assembly tool testing and optimization.

J.P.M.-K. and D.C. performed GGDC tool testing for protists and fungi.

P.S. and W.A. performed ANI tool testing.

A.S. and S.A. performed annotation tool testing, annotations and QC.

Y.T.T., G.L.B., M.E.S., T.Y.J., S.W.S., and C.A.Q. performed sample collection and preservation.

K.B. and S.D. performed data tracking.

## DECLARATIONS OF INTERESTS

The authors declare no competing interests.

Received: September 1, 2020

Revised: February 12, 2021

Accepted: March 4, 2021

Published: April 23, 2021

## REFERENCES

- Ahrendt, S.R., Quandt, C.A., Ciobanu, D., Clum, A., Salamov, A., Andreopoulos, B., Cheng, J.F., Woyke, T., Pelin, A., Henrissat, B., et al. (2018). Leveraging single-cell genomics to expand the fungal tree of life. *Nat. Microbiol.* **3**, 1417–1428.
- Alexander, W.G., Wisecaver, J.H., Rokas, A., and Hittinger, C.T. (2016). Horizontally acquired genes in early-diverging pathogenic fungi enable the use of host nucleosides and nucleotides. *Proc. Natl. Acad. Sci. U S A* **113**, 4116–4121.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.
- Arriola, E., Lambros, M.B., Jones, C., Dexter, T., Mackay, A., Tan, D.S., Tamber, N., Fenwick, K., Ashworth, A., Dowsett, M., et al. (2007). Evaluation of Phi29-based whole-genome amplification for microarray-based comparative genomic hybridisation. *Lab. Invest.* **87**, 75–83.
- Auch, A.F., von Jan, M., Klenk, H.P., and Göker, M. (2010). Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. *Stand. Genomic Sci.* **2**, 117–134.
- Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Pribelski, A.D., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477.
- Berbee, M.L., James, T.Y., and Strullu-Derrien, C. (2017). Early diverging fungi: diversity and impact at the dawn of terrestrial life. *Annu. Rev. Microbiol.* **71**, 41–60.
- Blackwell, M. (2011). The fungi: 1, 2, 3... 5.1 million species? *Am. J. Bot.* **98**, 426–438.
- Brown, R.B., and Audet, J. (2008). Current techniques for single-cell lysis. *J. R. Soc. Interfaces* **5**, S131–S138.
- Burki, F., Roger, A., Brown, M., and Simpson, A. (2019). The new tree of eukaryotes. *Trends Ecol. Evol.* **35**, 43–55.
- Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I.A., Belmonte, M.K., Lander, E.S., Nusbaum, C., and Jaffe, D.B. (2008). ALLPATHS: *de novo* assembly of whole-genome shotgun microreads. *Genome Res.* **18**, 810–820.
- Celio, G.J., Padamsee, M., Dentinger, B.T.M., Bauer, R., and McLaughlin, D.J. (2006). Assembling the fungal tree of life: constructing the structural and biochemical database. *Mycologia* **98**, 850–859.
- Chen, C., Xing, D., Tan, L., Li, H., Zhou, G., Huang, L., and Xie, X.S. (2017). Single-cell whole-genome analyses by linear amplification via transposon insertion (LIANTI). *Science* **356**, 189–194.
- Clingenpeel, S., Schwientek, P., Hugenholtz, P., and Woyke, T. (2014). Effects of sample treatments on genome recovery via single-cell genomics. *ISME J.* **8**, 2546–2549.
- Clingenpeel, S., Clum, A., Schwientek, P., Rinke, C., and Woyke, T. (2015). Reconstructing each cell's genome within complex microbial communities—dream or reality? *Front. Microbiol.* **5**, 771.
- Daley, T., and Smith, A.D. (2014). Modeling genome coverage in single-cell sequencing. *Bioinformatics* **30**, 3159–3165.
- Dean, F.B., Hosono, S., Fang, L., Wu, X., Faruqi, A.F., Bray-Ward, P., Sun, Z., Zong, Q., Du, Y., Du, J., et al. (2002). Comprehensive human genome amplification using multiple displacement amplification. *Proc. Natl. Acad. Sci. U S A* **99**, 5261–5266.
- Eichorst, S.A., Varanasi, P., Stavila, V., Zemla, M., Auer, M., Singh, S., Simmons, B.A., and Singer, S.W. (2013). Community dynamics of cellulose-adapted thermophilic bacterial consortia. *Environ. Microbiol.* **15**, 2573–2587.
- Ellegaard, K.M., Klasson, L., and Andersson, S.G.E. (2013). Testing the reproducibility of multiple displacement amplification on genomes of clonal endosymbiont populations. *PLoS One* **8**, e82319.
- Foissner, W. (1999). Protist diversity: estimates of the near-imponderable. *Protist* **150**, 363–368.
- Foissner, W. (2009). Protist diversity and distribution: some basic considerations. In *Protist Diversity and Geographical Distribution*, W. Foissner and D.L. Hawksworth, eds. (Springer Netherlands), pp. 1–8.
- Garvin, T., Aboukhalil, R., Kendall, J., Baslan, T., Atwal, G.S., Hicks, J., Wigler, M., and Schatz, M.C. (2015). Interactive analysis and assessment of single-cell copy-number variations. *Nat. Methods* **12**, 1058–1060.
- Gawad, C., Koh, W., and Quake, S.R. (2016). Single-cell genome sequencing: current state of the science. *Nat. Rev. Genet.* **17**, 175–188.
- Gawryluk, R.M.R., Del Campo, J., Okamoto, N., Strasser, J.F.H., Lukeš, J., Richards, T.A., Worden, A.Z., Santoro, A.E., and Keeling, P.J. (2016). Morphological identification and single-cell genomics of marine diplomonads. *Curr. Biol.* **26**, 3053–3059.
- Grigoriev, I.V., Nikitin, R., Haridas, S., Kuo, A., Ohm, R., Otiillar, R., Riley, R., Salamov, A., Zhao, X., Korzeniewski, F., et al. (2014). MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Res.* **42**, D699–D704.
- Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321.
- Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075.
- Han, N., Qiang, Y., and Zhang, W. (2016). ANItools web: a web tool for fast genome comparison within multiple bacterial strains. *Database (Oxford)* **2016**, baw084.
- Hou, Y., Wu, K., Shi, X., Li, F., Song, L., Wu, H., Dean, M., Li, G., Tsang, S., Jiang, R., et al. (2015). Comparison of variations detection between whole-genome amplification methods used in single-cell resequencing. *GigaScience* **4**, 37.
- Hyde, K.D. (2001). Where are the missing fungi? *Mycol. Res.* **105**, 1409–1412.



- Kogawa, M., Hosokawa, M., Nishikawa, Y., Mori, K., and Takeyama, H. (2018). Obtaining high-quality draft genomes from uncultured microbes by cleaning and co-assembly of single-cell amplified genomes. *Sci. Rep.* **8**, 2059.
- Lan, F., Demaree, B., Ahmed, N., and Abate, A.R. (2017). Single-cell genome sequencing at ultra-high-throughput with microfluidic droplet barcoding. *Nat. Biotechnol.* **35**, 640–646.
- Lasken, R.S., and Stockwell, T.B. (2007). Mechanism of chimera formation during the multiple displacement amplification reaction. *BMC Biotechnol.* **7**, 19.
- Lazarus, K.L., and James, T.Y. (2015). Surveying the biodiversity of the Cryptomycota using a targeted PCR approach. *Fungal Ecol.* **14**, 62–70.
- Lazarus, K.L., Benny, G.L., Ho, H.M., and Smith, M.E. (2017). Phylogenetic systematics of *Syncephalis* (Zoopagales, Zoopagomycotina), a genus of ubiquitous mycoparasites. *Mycologia* **109**, 333–349.
- Linnarsson, S., and Teichmann, S.A. (2016). Single-cell genomics: coming of age. *Genome Biol.* **17**, 97.
- Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., et al. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1**, 18.
- Macaulay, I.C., and Voet, T. (2014). Single cell genomics: advances and future perspectives. *PLoS Genet.* **10**, e1004126.
- Meier-Kolthoff, J.P., Auch, A.F., Klenk, H.P., and Göker, M. (2013). Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* **14**, 60.
- Neu, K.E., Tang, Q., Wilson, P.C., and Khan, A.A. (2017). Single-cell genomics: approaches and utility in immunology. *Trends Immunol.* **38**, 140–149.
- Ning, L., Li, Z., Wang, G., Hu, W., Hou, Q., Tong, Y., Zhang, M., Chen, Y., Qin, L., Chen, X., et al. (2015). Quantitative assessment of single-cell whole genome amplification methods for detecting copy number variation using hippocampal neurons. *Sci. Rep.* **5**, 11415.
- Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067.
- Peng, Y., Leung, H.C., Yiu, S.M., and Chin, F.Y. (2012). IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428.
- Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.-F., Darling, A., Malfatti, S., Swan, B.K., Gies, E.A., et al. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437.
- Rinke, C., Lee, J., Nath, N., Goudeau, D., Thompson, B., Poulton, N., Dmitrieff, E., Malmstrom, R., Stepanauskas, R., and Woyke, T. (2014). Obtaining genomes from uncultivated environmental microorganisms using FACS-based single-cell genomics. *Nat. Protoc.* **9**, 1038–1048.
- Roy, R.S., Price, D.C., Schliep, A., Cai, G., Korobeynikov, A., Yoon, H.S., Yang, E.C., and Bhattacharya, D. (2014). Single cell genome analysis of an uncultured heterotrophic stramenopile. *Sci. Rep.* **4**, 4780.
- Ruggiero, M.A., Gordon, D.P., Orrel, T.M., Bailly, N., Bourgoin, T., Brusca, R.C., Cavalier-Smith, T., Guiry, M.D., and Kirk, P.M. (2015). A higher level classification of all living organisms. *PLoS One* **10**, e0119248.
- Sibbald, S.J., and Archibald, J.M. (2017). More protist genomes needed. *Nat. Ecol. Evol.* **1**, 145.
- Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212.
- Spatafora, J.W., Chang, Y., Benny, G.L., Lazarus, K., Smith, M.E., Berbee, M.L., Bonito, G., Corradi, N., Grigoriev, I., Gryganskyi, A., et al. (2016). A phylum-level phylogenetic classification of zygomycete fungi based on genome-scale data. *Mycologia* **108**, 1028–1046.
- Spits, C., Le Caignec, C., De Rycke, M., Van Haute, L., Van Steirteghem, A., Liebaers, I., and Sermon, K. (2006). Whole-genome multiple displacement amplification from single cells. *Nat. Protoc.* **1**, 1965–1970.
- Stajich, J.E., Berbee, M.L., Blackwell, M., Hibbett, D.S., James, T.Y., Spatafora, J.W., and Taylor, J.W. (2009). The fungi. *Curr. Biol.* **19**, PR840–PR845.
- Stepanauskas, R. (2012). Single cell genomics: an individual look at microbes. *Curr. Opin. Microbiol.* **15**, 613–620.
- Tighe, S., Afshinnekoo, E., Rock, T.M., McGrath, K., Alexander, N., McIntyre, A., Ahsanuddin, S., Bezdán, D., Green, S.J., Joye, S., et al. (2017). Genomic methods and microbiological technologies for profiling novel and extreme environments for the extreme microbiome project (XMP). *J. Biomol. Tech.* **28**, 31–39.
- Tkacz, A., Hortala, M., and Poole, P.S. (2018). Absolute quantitation of microbiota abundance in environmental samples. *Microbiome* **6**, 110.
- Troell, K., Hallström, B., Divne, A.-M., Alsmark, C., Arrighi, R., Huss, M., Beser, J., and Bertilsson, S. (2016). *Cryptosporidium* as a testbed for single cell genome characterization of unicellular eukaryotes. *BMC Genomics* **17**, 471.
- Wurzbacher, C., Nilsson, R.H., Rautio, M., and Peura, S. (2017). Poorly known microbial taxa dominate the microbiome of permafrost thaw ponds. *ISME J.* **11**, 1938–1941.
- Xu, B., Li, T., Luo, Y., Xu, R., and Cai, H. (2014). An empirical algorithm for bias correction based on GC estimation for single cell sequencing. In *Trends and Applications in Knowledge Discovery and Data Mining*, W.-C. Peng, H. Wang, J. Bailey, V.S. Tseng, T.B. Ho, Z.-H. Zhou, and A.L.P. Chen, eds. (Springer International Publishing), pp. 15–21.
- Yoon, H.S., Price, D.C., Stepanauskas, R., Rajah, V.D., Sieracki, M.E., Wilson, W.H., Yang, E.C., Duffy, S., and Bhattacharya, D. (2011). Single-cell genomics reveals organismal interactions in uncultivated marine protists. *Science* **332**, 714–717.
- Zhang, Q., Wang, T., Zhou, Q., Zhang, P., Gong, Y., Gou, H., Xu, J., and Ma, B. (2017). Development of a facile droplet-based single-cell isolation platform for cultivation and genomic analysis in microorganisms. *Sci. Rep.* **7**, 41192.
- Zong, C., Lu, S., Chapman, A.R., and Xie, X.S. (2012). Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* **338**, 1622–1626.
- Lin, K., Limpens, E., Zhang, Z., Ivanov, S., Saunders, D.G., Mu, D., Pang, E., Cao, H., Cha, H., Lin, T., et al. (2014). Single nucleus genome sequencing reveals high similarity among nuclei of an endomycorrhizal fungus. *PLoS Genet.* **10**, e1004078.

## **Supplemental information**

### **A single-cell genomics pipeline for environmental microbial eukaryotes**

**Doina Ciobanu, Alicia Clum, Steven Ahrendt, William B. Andreopoulos, Asaf Salamov, Sandy Chan, C. Alisha Quandt, Brian Foster, Jan P. Meier-Kolthoff, Yung Tsu Tang, Patrick Schwientek, Gerald L. Benny, Matthew E. Smith, Diane Bauer, Shweta Deshpande, Kerrie Barry, Alex Copeland, Steven W. Singer, Tanja Woyke, Igor V. Grigoriev, Timothy Y. James, and Jan-Fang Cheng**

## Supplemental Information

### Supplemental Data, Figures and Tables Legend:

**Figure S1. Target EME Single-Cell Isolation from Environmental Samples. Related to Figure 1, step 2. A.** Schematics showing FACS target enrichment procedure (two-step FACS). FACS schematics was adapted from BD Influx™ Cell Sorter User's Guide. **B.** Target (here Protist +Fungi) recovery efficiency (numbers show % of total sorted cells) between direct-sort and two-step FACS depicted in panel A. Note: bacteria category, in this case, could also represent more than one organism, tightly associated groups, as well as target contamination via surface carryover. This does not exclude the presence of the target organism, but rather points out to the presence of undesirable organisms if the goal is a clean one-species genome assembly. Two-step FACS is recommended also for endosymbiont recovery, as it reduces significantly surface associated organisms.

**Figure S2. MDA start time comparison for confirmed target species. Related to Figure 1, step 3. and Data S3.** Top to bottom – first three plots show raw amplification start times for individual wells sorted with respective number of cells. Bar-graph shows normalized single-cell to 100 cells and 10 cells MDA start time. **A. Chytrid Fungi. B. Zoopagomycotina fungi. C. Kickxellomycotina and Ascomycota fungi.**

**Figure S3. Correlation between MDA start time, genome size and genome completeness. Related to Figure 1, 3, 7 and Data S2.** Top three plots: Graphic representation of correlation between MDA start time, assembled genome size and CEGMA estimated genome completeness, plotted for individual sorted wells. Well ID shown on the x-axis, for species where single-cell sorts were not enough for meaningful statistics, multiple cell sorts were included and are shown next to the plate well code; for the rest of the species only single-cell sorts are shown. Bottom table shows numerical correlation for these criteria. **A.** Chytrid fungi, **B.** Zoopagomycotina fungi, **C.** Ascomycota (*M.bicuspidata*) and Kickxellomycotina (*D. cristalligena*) fungi.

**Figure S4. Target Single-Cell Isolation Success from Environmental Samples. Related to Figure 7.A.** Relationship between FACS estimated target concentration in original sample (red) and total amplified single-cells (blue) and rDNA-PCR-sequencing confirmed target single-cells (purple). Samples on the plot are arranged from high to low target concentration in original sample based on FACS estimation. Polynomial trend curve is the best fitting trend. **B.** Pearson correlation (R) between FACS estimated target concentration in original sample and total MDA amplified single-cells, confirmed target single cells identified using rDNA-PCR-sequencing, as well as total amplified genomes and rDNA-PCR confirmed target OTU. Heat map: negative-red, no correlation –yellow, positive correlation – green. **C.** Percent amplified target genomes relationship with other metrics. % positive MDAs - % positive multiple displacement reactions; % positive PCRs - % positive PCR reactions for 16S, 18S, ITS rRNA regions; %positive Sanger - % PCR amplified, Sanger sequenced and BLAST confirmed rRNA for target species.

**Figure S5. rDNA assembly and OTU identification tools evaluation. Related to Figure 1, step 4.** Shown results are average for 8 fungal species (over 80 libraries) with standard deviation between species.

**Figure S6. *Caulochytrium protostelioides* single-cell genome coverage bias. Related to Figure 4 and 5.** Note: Average genome GC% for isolate was 65%, co-assembly regions with coverage was 50%, regions with no coverage was 68.99% +/- 0.0566%, see Table S6 for the no coverage regions. **A.** Whole genome mapped to the isolate genome assembly: purple: six single libraries individual genome assemblies. black: six single libraries individual genome assemblies and their co-assembly. Note that the

read coverage for assemblies was: isolate genome = 25X $\pm$  53; co-assembly of the six libraries = 55x $\pm$  88 of the normalized clean reads from merged fastq set. **B.** Zoomed into the genome locations 10000-11000 bp: **C.** six single libraries only. **D.** six single libraries individual genome assemblies and their co-assembly. Note that the read coverage for assemblies was: isolate genome = 25X $\pm$  53; co-assembly of the six libraries = 55x $\pm$  88 of the normalized clean reads from merged fastq set. **C.** Genome coverage over the coding regions, see **Table S6** for the list of genes with zero coverage.

**Figure S7. Long Read technology for MDA amplified genomes. Related to Figure 1, step 5. A.**

Illumina long read CLRS library, average Insert size 2500 bp. Inward and same direction reads are chimeric reads. Outward reads may contain partial chimera, identifiable after assembly. **B.** PacBio, 8 SMRT cells each library, average: read length 2900bp, PF Mb/cell: 85.8, PF reads/cell: 29,200, PF RQ: 84.50%. For 100 single cells Raw PacBio reads cover 98% of the reference at least 1x, for 1 single cell Raw PacBio reads cover 23% of the reference at least 1x.

**Figure S8. Phylogenomic placement of partial genomes. Related to Figure 4.**

RaxML trees with bootstrap values. Phyla names are on the right side of the color-coded vertical bars. **A.** *C. protostelioides* single-cell with lowest completeness (marked by sc) alone. **B.** *C. protostelioides* single- and multiple-cell amplified genomes assemblies with various degree of completeness (marked by sc). Co-assembly is marked by SC\_comb. Isolate unamplified genome is marked by 1. **C.** *D. cristalligena* single-cell or multiple-cell amplified genome assemblies with various degree of completeness (marked by sc).

**Table S1. rDNA qPCR primers used for OTU identification. Related to Figure 1, step 3 and Figure S5.**

Pairs are designated by the same color. Superscript refer to the original source: 1 [https://sites.duke.edu/vilgalyslab/rdna\\_primers\\_for\\_fungi/](https://sites.duke.edu/vilgalyslab/rdna_primers_for_fungi/) 2- Lazarus, et al., 2017, 3 - Dawson and Pace, 2002. These rDNA qPCR primers were selected and established and most reliable for a wide range of eukaryotes after testing the full list from source 1.

**Table S2. Four assemblers performance comparison for single-cell microbial eukaryotes with large genomes. Related to Figure 1, step 5.** Shown are top five assembly quality metrics that reflect the degree of fragmentation and completeness relative estimated genome size. For the test where used 51mln 2x150 bp Illumina raw normalized reads from three MiSeq ciliate protist libraries. Sag pipeline is the standardized production pipeline for prokaryote single-cell amplified genomes and consists of IDBA plus Allpaths, metagenome pipeline is SOAP.

**Table S3. Individual single-cell genome library assembly statistics for the metagenome pipeline.**

**Related to Figure 1, step 5.** Assembly metrics for HiSeq 27-30x read coverage for 7 libraries, after normalization, based on 100 MB genome size.

**Table S4. Four assemblers performance comparison for single-cell microbial eukaryotes with small genomes. Related to Figure 1, step 5.**

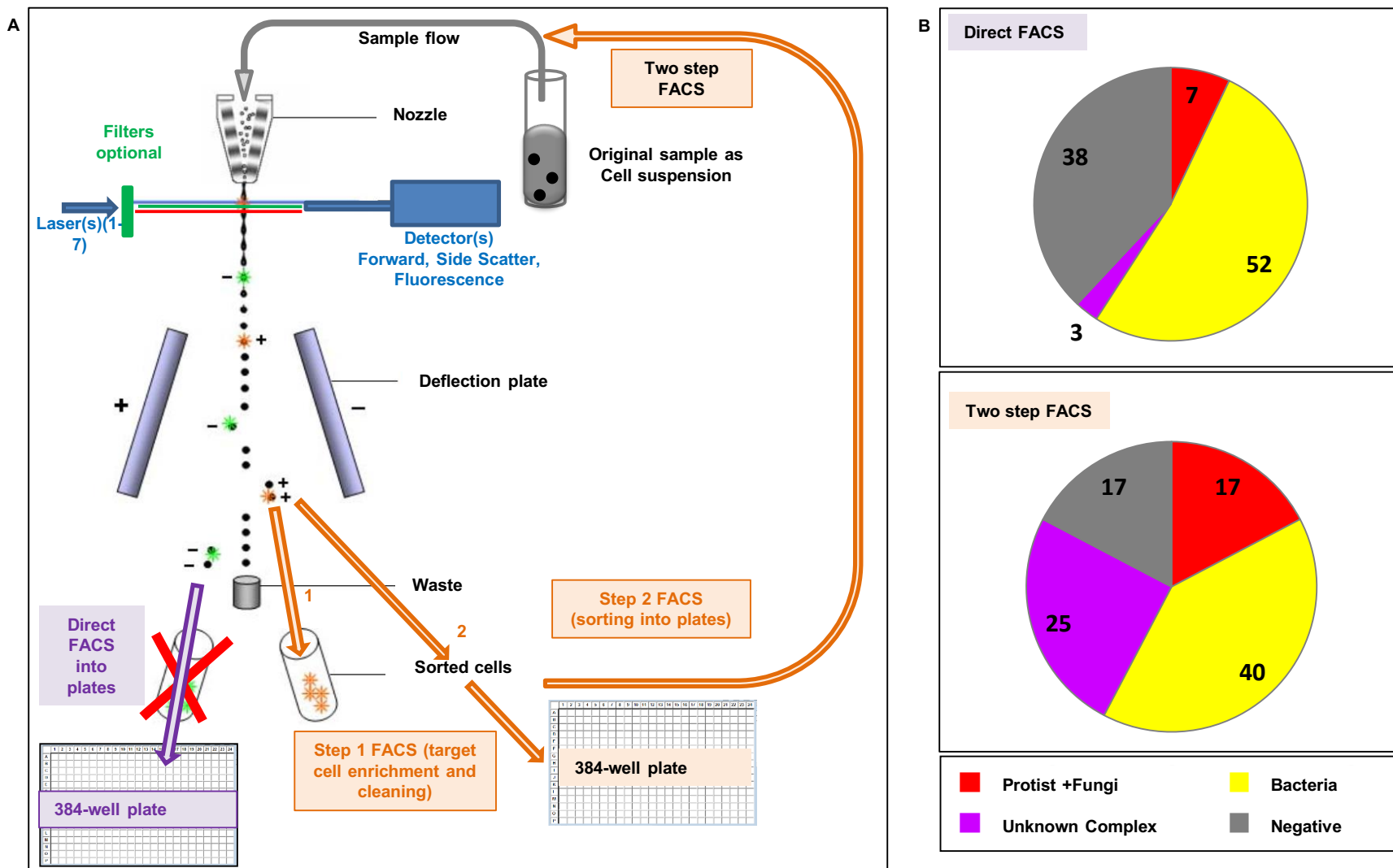
Assemblers were tested using three *P. cylindrospora* single-cell pooled libraries. Note: IDBA-UD and sag pipeline failed to complete individual assemblies for various reasons. \*Failed to run for co-assembly, but run for individual assemblies.

**Table S7. Annotation pipeline statistics for gene structure of the co-assembled species single-cell genomes. Related to Figure 8.**

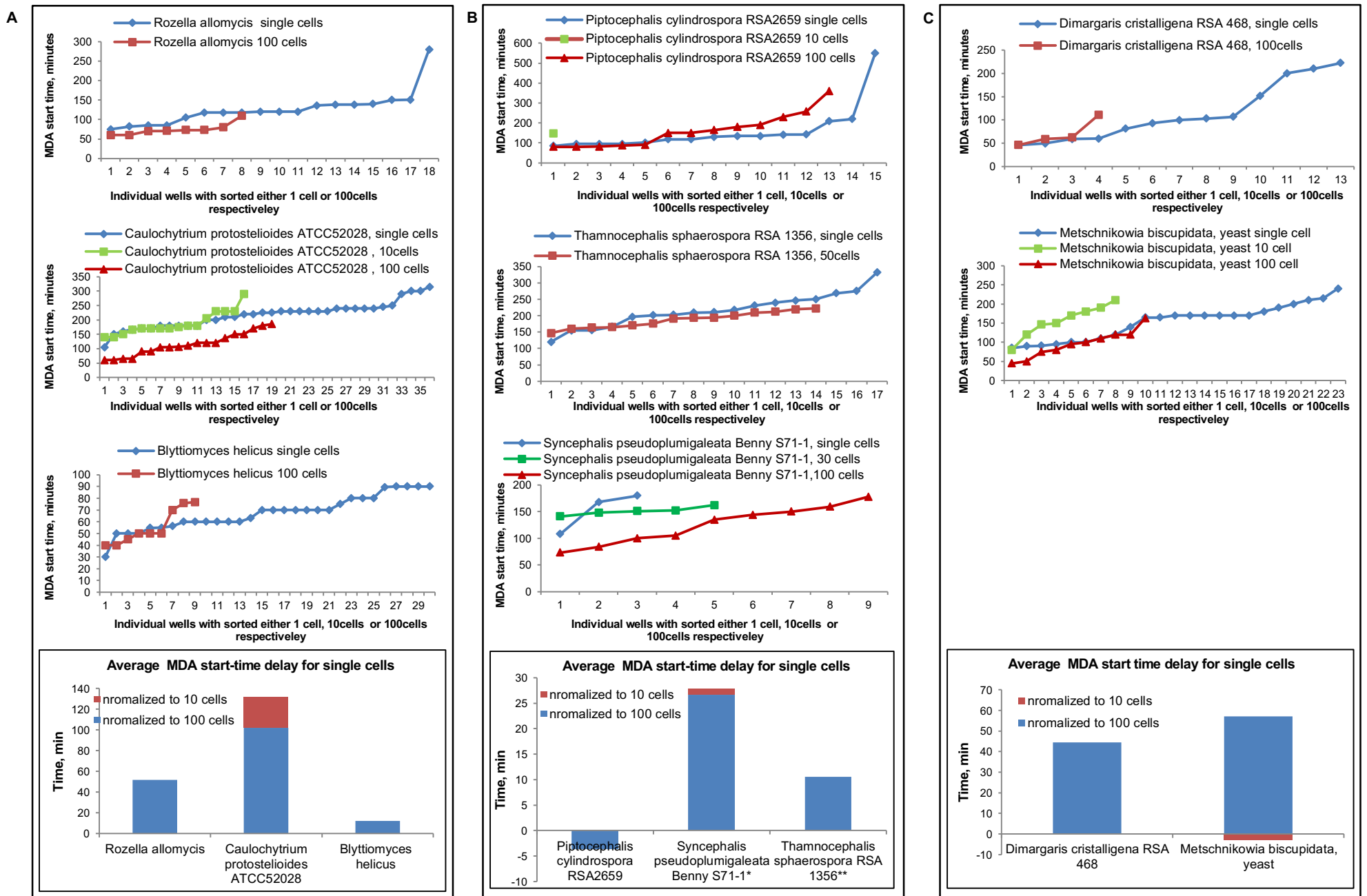
**Table S10. QC criteria pass and fail value for de novo clean species genome assembly. Related to Figure 1, 3 and Table 1, 2, 3.**

TgE – target enrichment. \*Exception are samples that pass biometric difference, for them pass value is 0.2% and fail value is 0.01%, in between more data is needed. The criteria recommended to be replaced by the following criterion are grey filled. RTU- random twentymer uniqueness. \*\* The values are shown for smallest genomes here (11Mb-13Mb), for larger genomes see

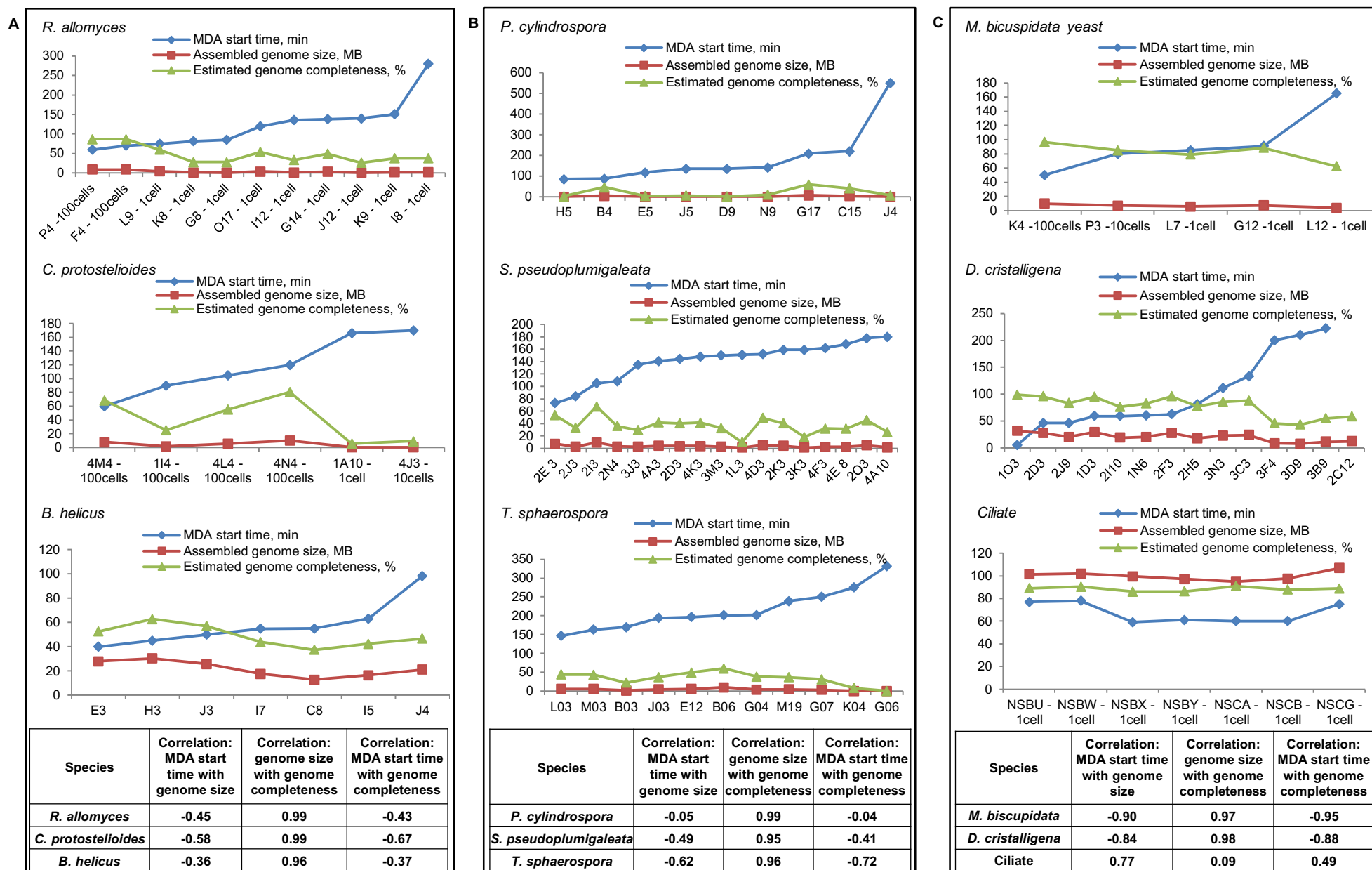
Figure 3.\*\*\*For phylogenomic, non-functional analysis the fail value is <5%. \$ BUSCO can be used, but a comparative assessment of BUSCO value with CEGMA is recommended.



**Figure S1. Target EME Single-Cell Isolation from Environmental Samples. Related to Figure 1, step 2. A.** Schematics showing FACS target enrichment procedure (two-step FACS). FACS schematics was adapted from BD Influx™ Cell Sorter User's Guide. **B.** Target (here Protist +Fungi) recovery efficiency (numbers show % of total sorted cells) between direct-sort and two-step FACS depicted in panel A. Note: bacteria category, in this case, could also represent more than one organism, tightly associated groups, as well as target contamination via surface carryover. This does not exclude the presence of the target organism, but rather points out to the presence of undesirable organisms if the goal is a clean one-species genome assembly. Two-step FACS is recommended also for endosymbiont recovery, as it reduces significantly surface associated organisms.

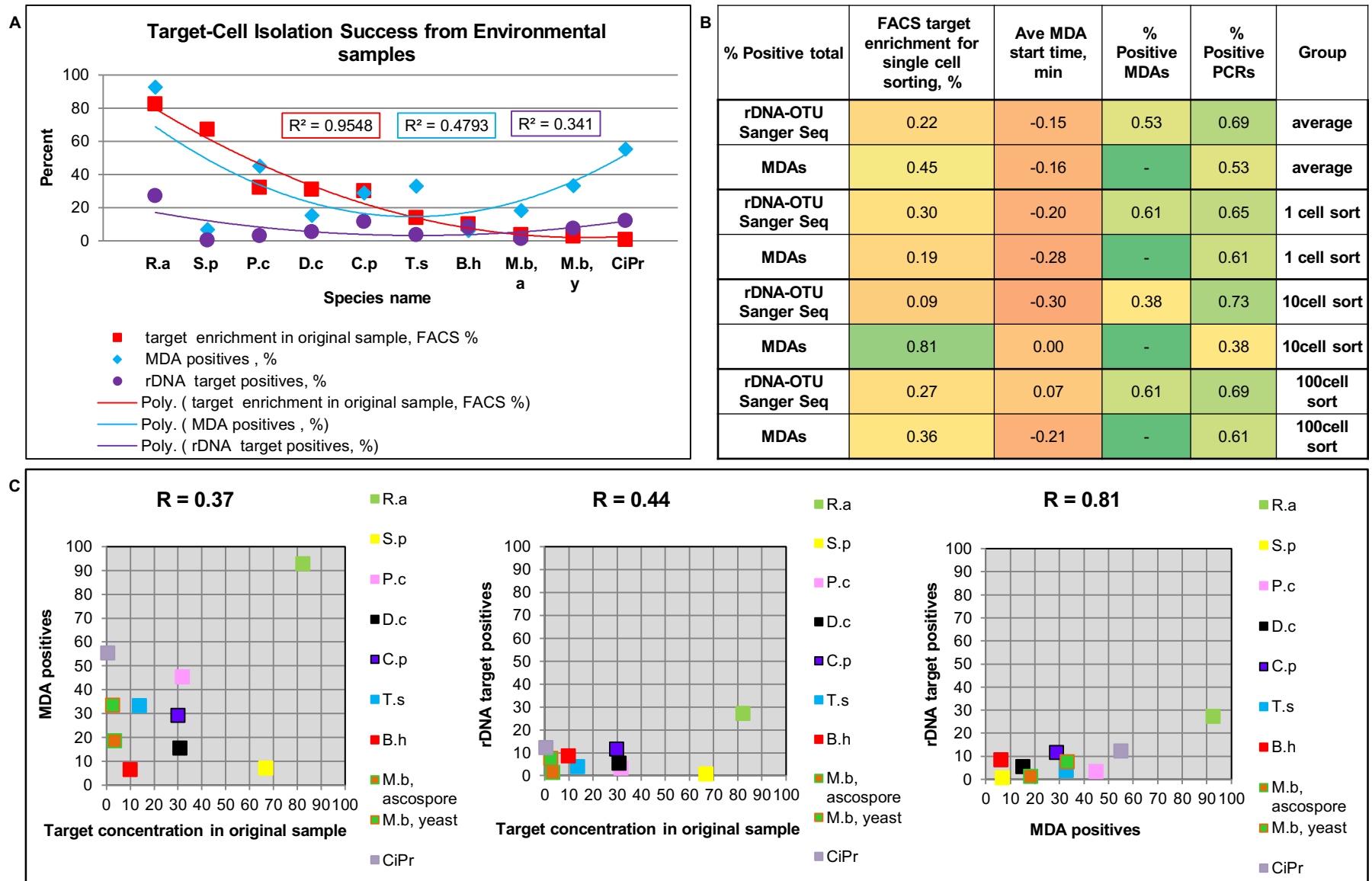


**Figure S2. MDA start time comparison for confirmed target species. Related to Figure 1, step 3. and Data S4. Top to bottom – first three plots show raw amplification start times for individual wells sorted with respective number of cells. Bar-graph shows normalized single-cell to 100 cells and 10 cells MDA start time. A. Chytrid Fungi. B. Zoopagomycotina fungi. C. Kickxellomycotina and Ascomycota fungi.**

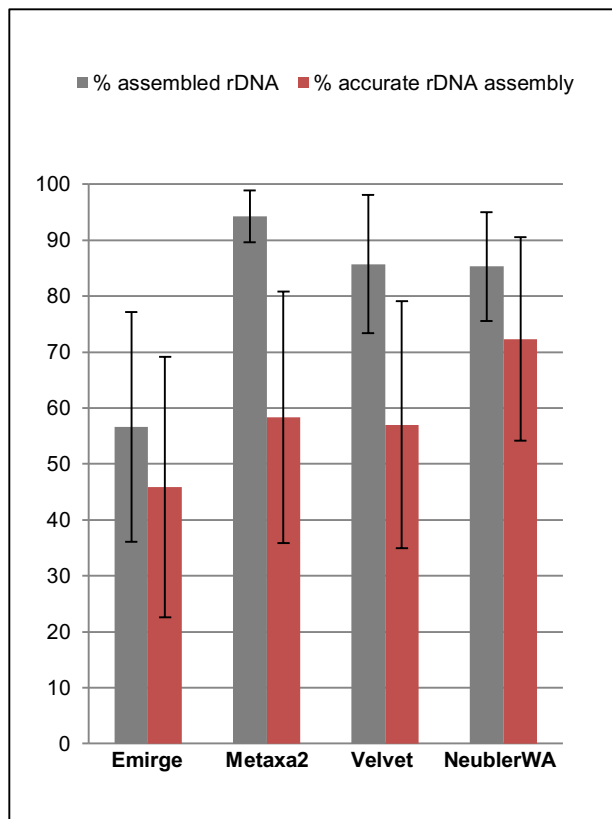


**Figure S3. Correlation between MDA start time, genome size and genome completeness.** Related to Figure 1, 3, 7 and Data S3. Top three plots: Graphic representation of correlation between MDA start time, assembled genome size and CEGMA estimated genome completeness, plotted for individual sorted wells. Well ID shown on the x-axis, for species where single-cell sorts were not enough for meaningful statistics, multiple cell sorts were included and are shown next to the plate well code; for the rest of the species only single-cell sorts are shown. Bottom table shows numerical correlation for these criteria. **A.** Chytrid fungi, **B.** Zoopagomycotina fungi, **C.** Ascomycota (*M.bicuspidata*) and Kickxellomycotina (*D. crystalligena*) fungi.

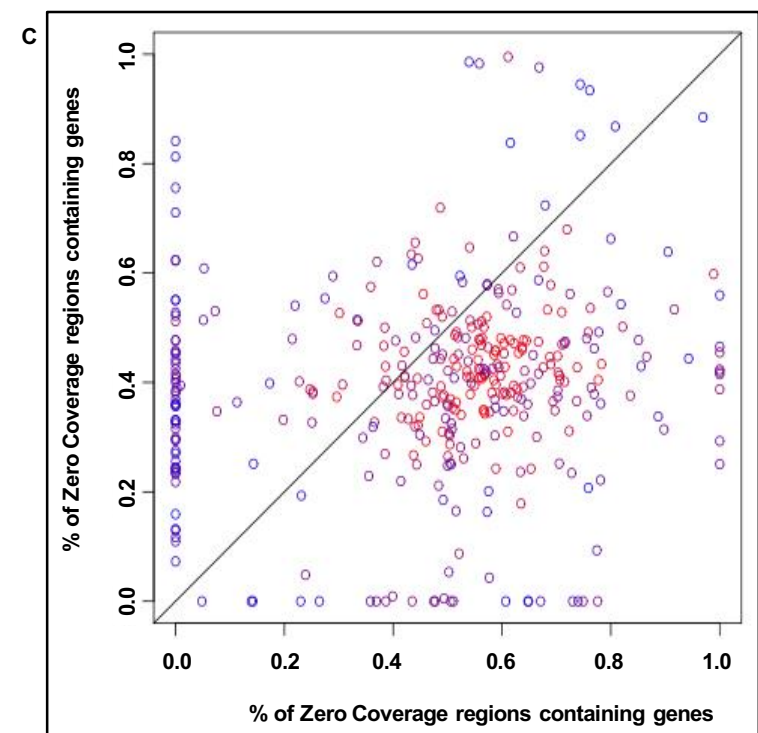
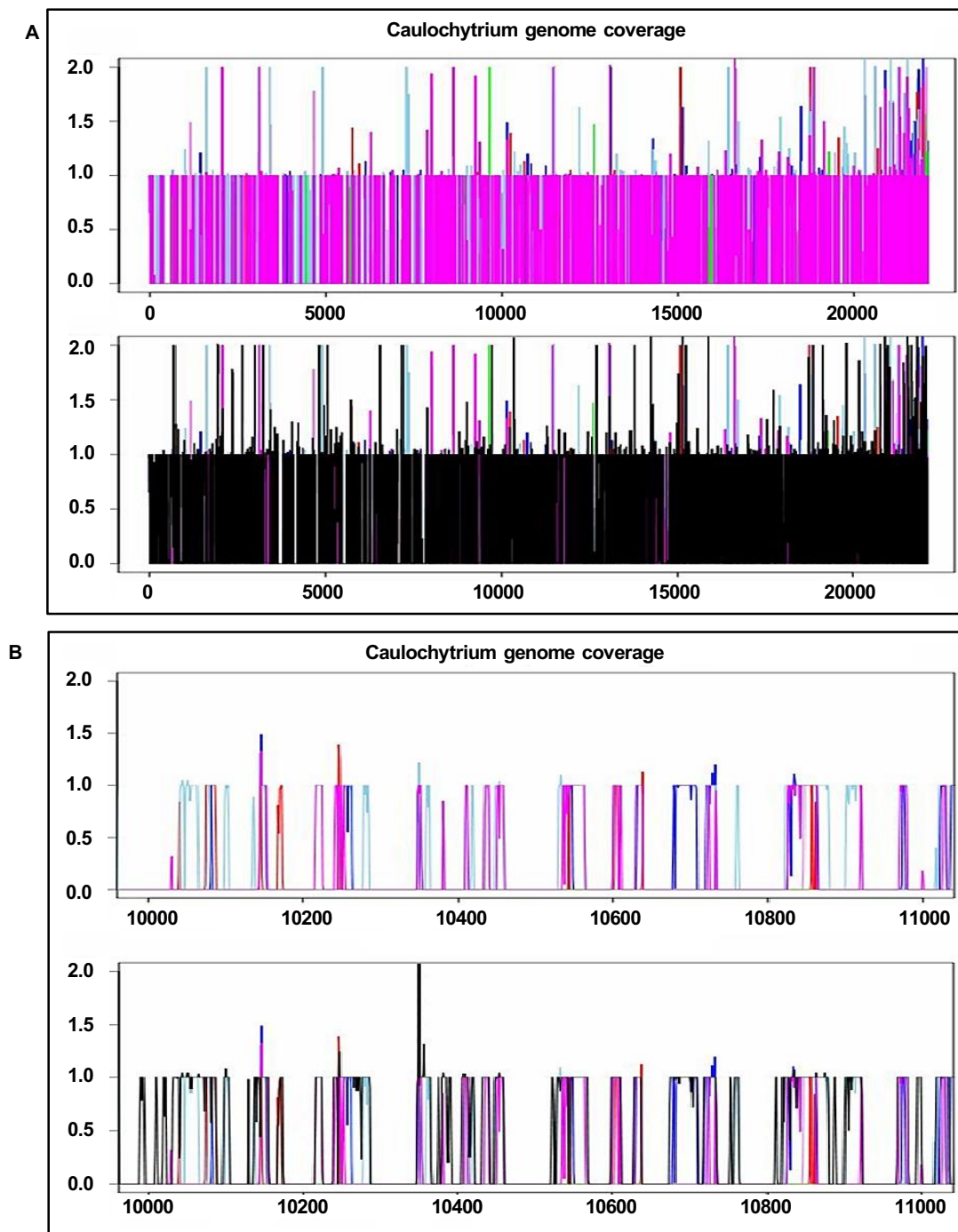




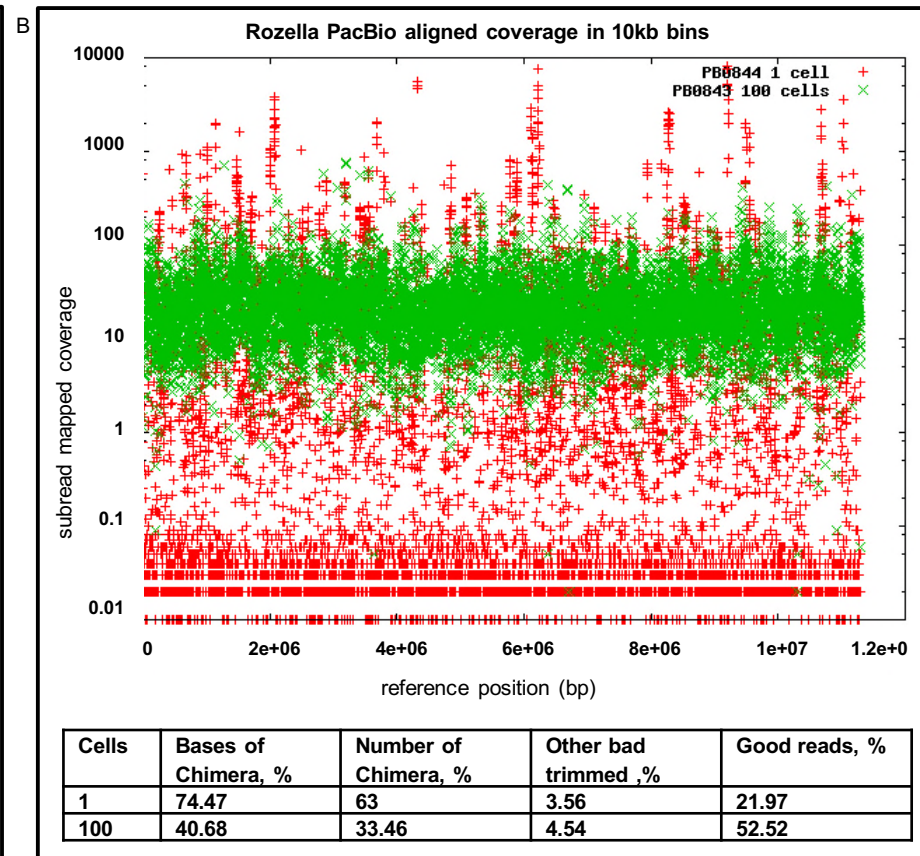
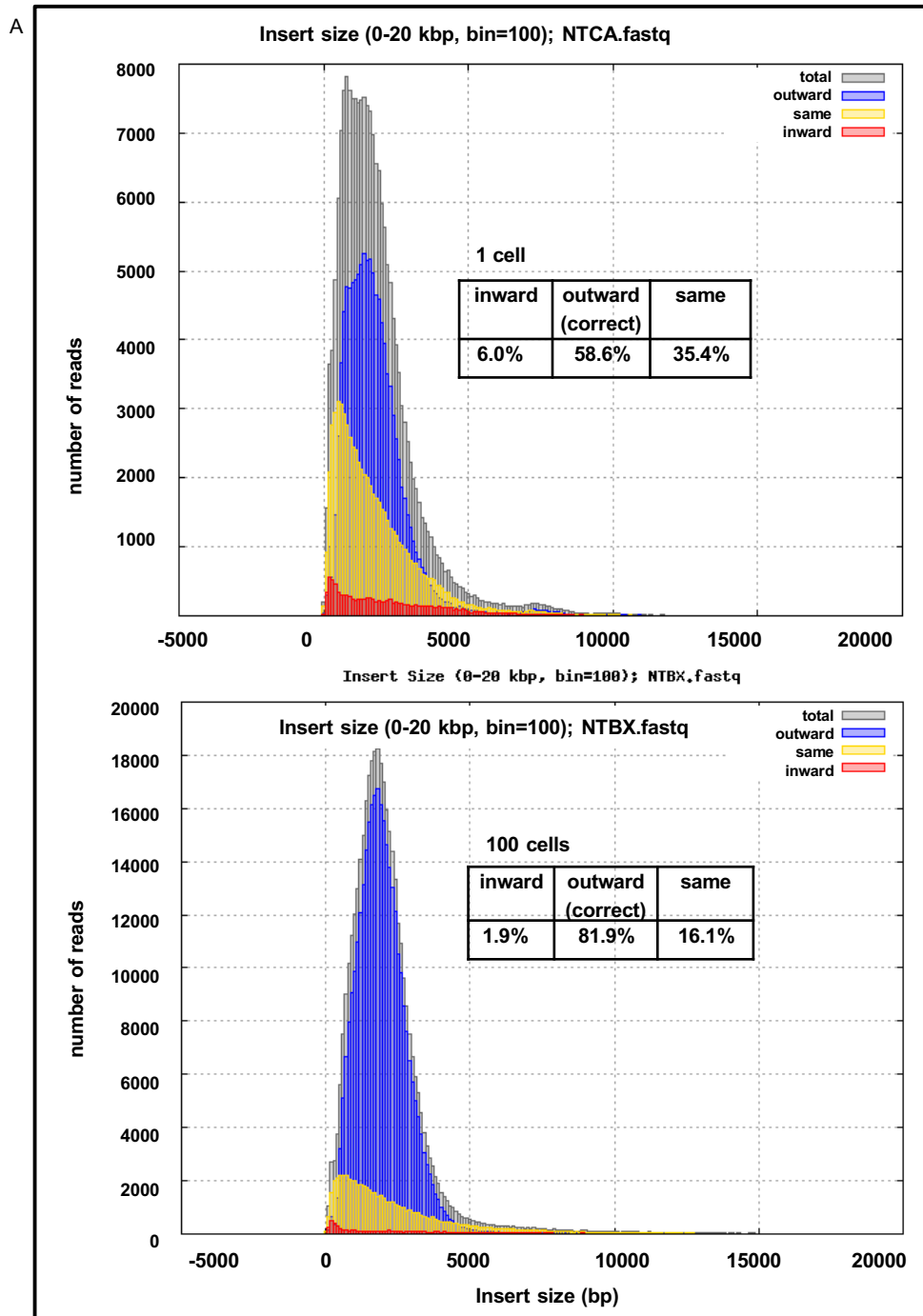
**Figure S4. Target Single-Cell Isolation Success from Environmental Samples. Related to Figure 7.A.** Relationship between FACS estimated target concentration in original sample (red) and total amplified single-cells (blue) and rDNA-PCR-sequencing confirmed target single-cells (purple). Samples on the plot are arranged from high to low target concentration in original sample based on FACS estimation. Polynomial trend curve is the best fitting trend. **B.** Pearson correlation ( $R$ ) between FACS estimated target concentration in original sample and total MDA amplified single-cells, confirmed target single cells identified using rDNA-PCR-sequencing, as well as total amplified genomes and rDNA-PCR confirmed target OTU. Heat map: negative-red, no correlation –yellow, positive correlation – green. **C.** Percent amplified target genomes relationship with other metrics. % positive MDAs - % positive multiple displacement reactions; % positive PCRs - % positive PCR reactions for 16S, 18S, ITS rRNA regions; %positive Sanger - % PCR amplified, Sanger sequenced and BLAST confirmed rRNA for target species.



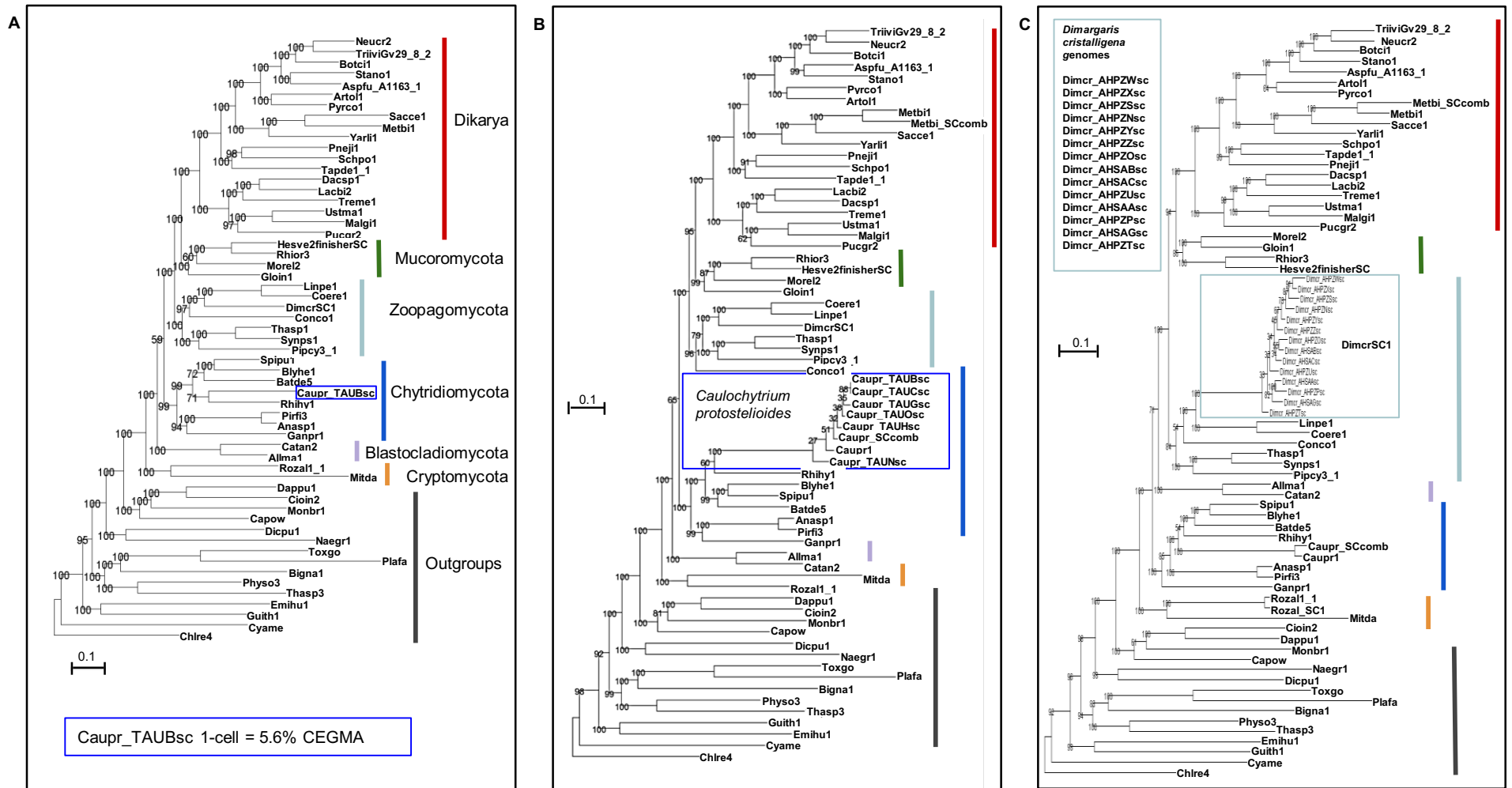
**Figure S5. rDNA assembly and OTU identification tools evaluation. Related to Figure 1, step 4.** Shown results are average for 8 fungal species (over 80 libraries) with standard deviation between species.



**Figure S6. *Caulochytrium protostelioides* single-cell genome coverage bias. Related to Figure 4 and 5.** Note: Average genome GC% for isolate was 65%, co-assembly regions with coverage was 50%, regions with no coverage was 68.99% +/- 0.0566%, see Table S6 for the no coverage regions. **A.** Whole genome mapped to the isolate genome assembly: purple: six single libraries individual genome assemblies. black: six single libraries individual genome assemblies and their co-assembly. Note that the read coverage for assemblies was: isolate genome = 25X +/- 53; co-assembly of the six libraries = 55x +/- 88 of the normalized clean reads from merged fastq set. **B.** Zoomed into the genome locations 10000-11000 bp: **C.** six single libraries only. **D.** six single libraries individual genome assemblies and their co-assembly. Note that the read coverage for assemblies was: isolate genome = 25X +/- 53; co-assembly of the six libraries = 55x +/- 88 of the normalized clean reads from merged fastq set. **C.** Genome coverage over the coding regions, see Table S6 for the list of genes with zero coverage.



**Figure S7. Long Read technology for MDA amplified genomes. Related to Figure 1, step 5. A.** Illumina long read CLRS library, average Insert size 2500 bp. Inward and same direction reads are chimeric reads. Outward reads may contain partial chimera, identifiable after assembly. **B.** PacBio, 8 SMRT cells each library, average: read length 2900bp, PF Mb/cell: 85.8, PF reads/cell: 29,200, PF RQ: 84.50%. For 100 single cells Raw PacBio reads cover 98% of the reference at least 1x, for 1 single cell Raw PacBio reads cover 23% of the reference at least 1x.



**Figure S8. Phylogenomic placement of partial genomes. Related to Figure 4.** RaxML trees with bootstrap values. Phyla names are on the right side of the color-coded vertical bars. **A.** *C. protostelioides* single-cell with lowest completeness (marked by sc) alone. **B.** *C. protostelioides* single- and multiple-cell amplified genomes assemblies with various degree of completeness (marked by sc). Co-assembly is marked by SC\_comb. Isolate unamplified genome is marked by 1. **C.** *D. cristalligena* single-cell or multiple-cell amplified genome assemblies with various degree of completeness (marked by sc).

**Table S1. rDNA qPCR primers used for OTU identification. Related to Figure 1, step 3 and Figure S5.** Pairs are designated by the same color. Superscript refer to the original source:1 [https://sites.duke.edu/vilgalyslab/rdna\\_primers\\_for\\_fungi/](https://sites.duke.edu/vilgalyslab/rdna_primers_for_fungi/) 2- Lazarus, et al., 2017, 3 - Dawson and Pace, 2002. These rDNA qPCR primers were selected and established and most reliable for a wide range of eukaryotes after testing the full list from source 1.

Phylogenetic group	rDNA region	Code name	Primer name	Sequence 5'to 3'
universal	16S	16SV6	926wF-M13pyro	GTTTTCCCAGTCACGACGTTGTAGAAACTYAAAKGAATTGRCGG
universal	16S	16SV6	1392R-M13pyro	AGGAAACAGCTATGACCATACGGGCGGTGTGTRC
Eukarya, Fungi	ITS	ITS <sup>1</sup>	ITS4rev	TCCTCCGCTTATTGATATGC
Eukarya, Fungi	ITS	ITS <sup>1</sup>	ITS5for	GGAAGTAAAAGTCGTAACAAGG
Cryptomycota	18S	18SCRYPTO <sup>2</sup>	M13CRYPTO2-2F	GTTTTCCCAGTCACGACCACAGGGAGGTAGTGACAG
Cryptomycota	18S	18SCRYPTO <sup>2</sup>	M13AU4v2	CAGGAAACAGCTATGACGCCTCACTAAGCCATTC
Protists, Eukarya	18S	18SDPD <sup>3</sup>	M13DPD360FE	GTTTTCCCAGTCACGACCGGAGARGGMGCMTGAGA
Protists, Eukarya	18S	18SDPD <sup>3</sup>	M13DPD1492RE	CAGGAAACAGCTATGACACCTTGTACGRCTT
Eukarya, Fungi	18S	18S_SR <sup>1</sup>	M13SR1RFor	GTTTTCCCAGTCACGACTACCTGGTTGATYCTGCCAGT
Eukarya, Fungi	18S	18S_SR <sup>1</sup>	M13NS4Rev	CAGGAAACAGCTATGACCTCCGTCAATTCCTTTAAG

**Table S2. Four assemblers performance comparison for single-cell microbial eukaryotes with large genomes. Related to Figure 1, step 5.** Shown are top five assembly quality metrics that reflect the degree of fragmentation and completeness relative estimated genome size. For the test where used 51mln 2x150 bp Illumina raw normalized reads from three MiSeq ciliate protist libraries. Sag pipeline is the standardized production pipeline for prokaryote single-cell amplified genomes and consists of IDBA plus Allpaths, metagenome pipeline is SOAP.

assembler	number of contigs	contig N50	Longest contig	assembled genome size	estimated genome size
IDBA-UD	412,972	381 BP	29,832 KB	157.1 MB	n/a
sag pipeline	8,933	2.2 KB	27,532 KB	18.4 MB	150 MB
metagenome pipeline	96,312	3.1 KB	72,415 KB	115.3 MB	n/a
spades 2.4	94,876	635 KB	6,323 KB	50.8 MB	na

**Table S3. Individual single-cell genome library assembly statistics for the metagenome pipeline. Related to Figure 1, step 5.**  
 Assembly metrics for HiSeq 27-30x read coverage for 7 libraries, after normalization, based on 100 MB genome size.

Library name	% reads remaining after normalization	Number of contigs	contig N50	Longest contig	Assembled genome size	Estimated genome size	Estimated genome completeness, CEGMA %	% 20mer uniqueness	Average GC %
NSBU	57.2	32,983	3923 KB	147.871 KB	101.3 MB	113.7 MB	89.1	97	37.98
NSBW	55.9	31,715	3583 KB	138.592 KB	102 MB	112.8 MB	90.4	60	38.04
NSBX	57.1	32,455	4109 KB	189.243 KB	99.6 MB	115.8 MB	86	98	37.61
NSBY	63.5	32,865	4093 KB	211.538 KB	97.1 MB	112.6 MB	86.2	97	37.82
NSCA	61.5	33,566	4369 KB	106.682 KB	94.9 MB	104.3 MB	91	70	38.09
NSCB	57.4	33,654	4296 KB	148.676 KB	97.6 MB	111.2 MB	87.8	98	38.16
NSCG	63.5	35,603	4489 KB	76.398 KB	107 MB	120.4 MB	88.9	98	37.73



## Transparent Methods:

### Step by step Method testing and Optimization of the single cell pipeline

We explored all factors across a set of diverse samples, that can influence the recovery of EME complete genomes. We tested what QC criteria could be used to predict the efficiency and quality of EME single-cell genome recovery. Following the general idea of using shallow sequencing as a prediction tool at an earlier step (Daley et al., 2014), we made a number of simple but highly effective changes to the amplification and screening process of the single-cell amplified genomes prior to the deep sequencing step, which allowed us to reduce costs and significantly improve genome quality of the EME.

### Step 1. Environmental sample collection and target identification

Eleven different samples with various degrees of complexity were used for this study (see Data S1, Table 2 and Figure 2). Our target species were: eight fungal obligate symbionts: six mycoparasites: *Caulochytrium protostelioides* [Chytridiomycota], *Rozella allomycis* [Cryptomycota], *Syncephalis pseudoplumigaleata* [Zoopagomycotina], *Thamnocephalis sphaerospora* [Zoopagomycotina], *Piptocephalis cylindrospora* [Zoopagomycotina], *Dimargaris cristalligena* [Kickxellomycotina], one crustacean parasite *Metschnikowia bicuspidata* [Ascomycota], one saprobe symbiont of pollen *Blyttomyces helicus* [Chytridiomycota], a free living protist from ciliate group plus any number of uncharacterized species from Cryptomycota and Chytridiomycota phyla.

Sample complexity level was estimated based on the combination of such factors: target cell abundance (concentration and total amount in the provided volume), phylogenetic and biometric diversity of organisms, presence of 'competing' cells for sorting process (e.g. cells with similar biometric characteristics), shape of the target cell, target cell wall complexity, target cell fragility (see Table 2). Samples were collected in their natural environment in different locations and shipped to the Joint Genome Institute (JGI), where all subsequent work was performed. Seven of our samples were obtained from dual non-axenic cultures re-creating host-parasite environment in laboratory conditions. Other four of our samples were collected directly from the environment. One of them had media and nutrients added to enrich for the target species.

Specifically: A compost sample enriched with microcrystalline cellulose was prepared as described in Eichorst et al., 2013. The sample was received at JGI two weeks later. The ciliate protist lifestyle was observed and documented for 2 months. The sample was continuously stored at room temperature in either a wet or dry state containing microcrystalline cellulose particles. The *Rozella allomycis* CSF55 sample was prepared as described in James et al., 2013 and shipped to JGI on ice in 10% glycerol, after which it was stored at -80°C. The sample was thawed on ice and stored at room temperature after the zoospores regained motility. A dual culture of *Caulochytrium protostelioides* ATCC52028 with its host *Sordaria* was used to isolate parasitic zoospores at  $2.5 \times 10^6$  per ml. The zoospore suspension was preserved in 10% DMSO with 10% fetal bovine serum, shipped on dry ice, and stored at -80°C. The DNA isolated from this sample was prepared via multiple cleaning steps of the zoospores of the dual culture at the Timothy Y James laboratory at the University of Michigan, USA. *Blyttomyces helicus* was grown through enrichment methods using spruce pollen in bog water. The sample was obtained from Perch Pond Fen near Old Town, Penobscot County, Maine, in June 2014. This enrichment culture was filtered through 40- $\mu$ m mesh (removing pollen and sporangia) and concentrated by centrifugation. *Metschnikowia bicuspidata* standard was isolated from an infected population of the water flea *Daphnia dentifera* grown under laboratory conditions in the Meghan Duffy laboratory at the University of Michigan, USA. *Daphnia* were dissected under a stereoscope. First, *Daphnia* were rinsed repeatedly with deionized water. Then, insect pins were used to puncture the *Daphnia* carapace, and a micropipette was used to collect *Daphnia* hemolymph, which contained a mixture of yeast cells and ascospores of *M. bicuspidata*. Cells were preserved in 10% glycerol at a concentration of  $10^5$  spores per ml and stored at -80°C. *Dimargaris cristalligena* RSA 468 was grown on V8 juice agar [1 small can of original V8 juice [5.5 oz, 163 ml], diluted to 1 L with diH<sub>2</sub>O; 3 g of CaCO<sub>3</sub>; 20 g of agar] and cultured with *Cokeromyces recurvatus*. Spores were shipped in 10% sterile glycerol. *Syncephalis pseudoplumigaleata* Benny S71-1 was grown on *Mucor moelleri* on 10% wheat germ agar [Wg10, Benny et al., 2016]. Parasite hyphae and spores were shipped to JGI by Jerry Benny in 50% glycerol. *Thamnocephalis sphaerospora* RSA 1356 was grown in dual culture with the fungal host *Microascus* and harvested from Petri plates. The sample was stored in 50% glycerol at -80°C. *Piptocephalis cylindrospora* RSA 2659 was cultivated on potato dextrose agar with its fungal host *Cokeromyces*. The culture was grown on many Petri dishes, and the spores of both the fungus and the host *Cokeromyces* were removed from the culture by washing the plates with 0.2% Triton X-100. An estimated  $2.5 \times 10^7$  spores/ml of parasite with host were obtained and preserved in 10% glycerol at -80°C.

Ribosomal DNA screening or microscopic examination was used to confirm target species or phyla presence as well as overall taxonomic diversity of the sample. For two of the environmental samples which did not have visually identifiable taxa we used rDNA screening only (Data S1t,u). For another environmental sample rDNA screening failed to identify target EME in the original environment and was initially identified using microscopy only (Data S1r,s). In the rest of the samples target species were confirmed both by microscopy and rDNA-PCR (Data S1a-p). Two of the co-cultures were used for obtaining unamplified genomes from bulk DNA isolates for benchmarking single- and multiple-cell amplified genomes (Data S1a-d).

**Optimization: Enrichment of the target via filtration steps can improve recovery rate and reduce time costs in the next steps.**

Ultra-low abundant target organisms that are identifiable via microscopy and could have their size determined, but their OTU fails to be identified by rDNA PCR, can be enriched via layered filtration steps (See Data S1i) and resuspension in the original 0.2µM filtered environment. We do not recommend gradient centrifugation for size separation due to drastic change of the living environment. Preservation of samples in their original media and conditions is more favorable than freezing- thawing in media with cell-stabilizers. Shipping on wet-ice for samples that tolerate cold in their native environment is better, than freezing/shipping on dry ice. However, if the samples have been frozen, they should be kept as such (e.g. shipped on dry ice) until FACS isolation during next step.

## **Step 2. Single-cell FACS isolation**

The single-cell isolation process is shown in Figure S1. Single-cells were isolated via FACS as shown in Data S1 and Figure S1. FACS was performed using a BD Influx™ Cell Sorter according to the manufacturer's instructions ([BD Influx™ Cell Sorter User's Guide](#)). The instrument fluidics lines were sterilized prior to each use with 10% bleach solution, followed by extensive rinsing with deionized and filter-sterilized Milli-Q water and sterilized sheath fluid, prior to each use. For the sheath fluid, a sterile 0.01-µm- or 0.02-µm-filtered 1x or 0.5x PBS solution was used. The instrument was calibrated for each light source used, and for fluid stability, each time prior to sorting using 2-µm green fluorescent beads from BD. For fluorescent cell labeling, SYBR Green, SYTO 9, Tubulin Tracker and wheat germ agglutinin (WGA) with various fluorophores were used (see details for which dye and organism in Data S1). We tested these commonly used, non-specific labeling techniques that are capable to stain live (non-fixed) organisms and differentiate clearly different populations, e.g. - the target organisms from the undesired ones. We started with SYBR Green as the most common used DNA labeling method that allows removal of abiotic impurities in addition to size gating. However, we discovered that most of the fungal targets and a wide range of bacterial contaminants do not take this label well. The same organisms will take better Syto9 and will not bleach or excrete it as fast as SYBR Green. We then tested both SYBR Green and Syto9 on more complex samples that had non-target species of the same size as the target species and in this case the DNA labeling would not allow to differentiate them by size gating. For these samples using Tubulin tracker allowed to differentiate the flagellated (in some cases target and, in some samples, undesired) species from the rest. WGA allowed to differentiate between fungi and algae in some samples where both were flagellated and similar size.

Additionally, the cell sorting accuracy was verified using a Zeiss Axio Observer D1 microscope and published species morphological descriptions, when available.

**Optimization: Enrichment and quantifying samples prior to sorting increase recovery of target cells.**

Repeated microscopic evaluation prior and after FACS in step2 proved to be valuable tool for validating target organism concentration post sample collection in step1. For example: In one sample harboring Chromista (SAR) species, rDNA-OTU screening of the bulk sample failed to detect the target due to its ultra-low concentration, however large size and distinct morphology allowed for target detection via microscopy. In this case enrichment via layered size filtration allowed for target enrichment sufficient for the two-step FACS, which improved target genome isolation (Figure S1). In two environmental samples (Supplemental Figure 1u, v), where target phyla were at ultra-low concentration and had small size similar with the majority of the organisms in the sample, microscopy was not very useful. For these two samples rDNA profiling of the multiple-cell sorts of various tight FACS populations was the only tool capable to evaluate target phyla presence and abundance. rDNA profiling of these samples displayed an ultra-low abundance (0-0.1%, see Supplemental Figure 1v). Target recovery rate for these samples was 1 genome in 500 cells and were classified as the lowest threshold for the current pipeline.

Most importantly, a two-step FACS enrichment prior single-cell sorting into 384-well plates increased recovery of target cells more than double (Figure S1). For two samples (target species *B.*

*helicus* and *M. bicuspidata*) a limited amount of target (2.6% and 3.5% in 2ml and 5% in 5ml respectively, Table 1) in starting material did not allow for a two-step FACS enrichment (Data S1e,f,o,p and S4). The number of clean target single-cell genomes in these samples was smaller than for the other fungal species, where two-step FACS was used (Data S2f) and the few *M. bicuspidata* isolated ascospores did not result in a genome assembly due to high contamination rate. Nevertheless, drastic differences in cell size and shape between target and non-target cells in these samples allowed us to recover enough single-cells that were co-assembled into high quality genomes (Figure 5 and 6).

In summary, we found that combining both size filtration and FACS enrichment, when non-target organisms and matter are at a prohibitively high rate (>90%) in the sample (Data S1k, l, m, n, r, s) significantly reduced carry-on 'contaminants' and increased the number of clean single-cells which ultimately led to higher quality single-cell and species co-assembled genomes (Figures 4-6 and S4).

### Step3: Cell Lysis and genome amplification

Further conditions are required to ensure a full single-cell genome recovery: 1. Single-cell lysis and genome amplification should happen in one-tube reaction to avoid loss of minute genomic material. 2. Uniform amplification and complete genome recovery are facilitated by easy and equal access of the MDA reagents to the cell's DNA. Consequently, the assumption that efficient lysis may result in an earlier Start of the Genome Amplification reaction (SGA) has been proposed for investigation as a possible QC step. We used purified DNA and incremental cell sorts (1, 10, 30, 50, and 100 cells per reaction) to test a range of lysis-MDA conditions (described in Data S3). SGA criterion was used for evaluation of lysis-MDA efficiency in these tests. Cell lysis solutions (described in Data S3) were prepared using the following reagents: KOH dry pellets reconstituted to 500 mM with nuclease-free water (H<sub>2</sub>O sc), 1 M DTT, and HCl (stop buffer) obtained from the REPLI-g® Single Cell Kit from Qiagen (part # 150345) and following the kit protocol; Tween 20 (SIGMA, P9416-100 ml), 0.5 M EDTA (Ambion, AM92606), Proteinase K (NEB, P8107S), PMSF (SIGMA, 532789-5 g), and EGTA (SIGMA, E3889-10 g) were purchased and sterilized separately.

For MDA, one of the following was used: REPLI-g Single Cell kit (Qiagen part # 150345), RepliPHI kit, (Epicenter catalog #RH040110) or separate reagents (10 mM dNTP, NEB part #N0447L; 500 mM hexamers IDT order #37617009; phi29 polymerase 10,000 U/ml, NEB #M029L; DMSO 99.9% pure, SIGMA D8418-50 ml) and JGI homemade 10X buffer (400 mM Tris-HCl, pH 7.5, Ambion, AM9855; 500 mM KCl, Ambion, AM9640G; 100 mM MgCl<sub>2</sub>, Ambion, AM9530; 50 mM (NH<sub>4</sub>)<sub>2</sub>SO<sub>4</sub>, Sigma, AA4418-100G; 20 mM DTT, Invitrogen, P2325) supplemented with SYTO-13 (Invitrogen, part # S7575) diluted 1.27E+05 times for real-time tracking. For either of the chemistries, the reactions were carried out at 29°C-30°C until a desirable amplification level was achieved, from 2 h to 14 h. For either of the chemistries, all plasticware was UV-sterilized for 1 h prior to solution preparation in a *Stratagene UV Stratalinker 2400*. H<sub>2</sub>O, lysis buffers, HCl and 10X reaction buffers were UV-sterilized for an additional 1 h prior to final solution preparation. For the RepliPHI kit and NEB-phi29 homemade MDA kit, the final reaction was UV-sterilized for an additional 1 h after each of the reagents (except the enzyme) were UV-sterilized for 3 h in UV-sterilized plasticware. All the work was conducted in a sterile hood without airflow. Hood sterilization was performed as follows: 70% ethanol, followed by 10-50% bleach, 70% sterile isopropanol (TexWipe #TX3270), and 1 h UV sterilization. Personnel were gowned with sterile single-use gloves and a coat for each reaction setup. All reactions were performed on pre-sterilized (for 10-15 minutes in a *Stratagene UV Stratalinker 2400*) Bio-Rad 384-well plates (#HSP3805).

Our tests show that some lysis-MDA conditions are suitable for some species more than other species (Figures S4), based on start of genome amplification (SGA). As a result of these tests, we chose a single protocol for lysis-MDA that had acceptable efficiency for broad phylogenetic sampling despite being suboptimal for some of the samples (Data S3). For the chosen protocol, single-cell SGA happened 30 min after positive control (10pg purified DNA or 10-100 cells) varying from 5 min to 1hr 50min between different species (Figure S2). The average success rate of MDA was 33% in 288-576 sorted cells per sample, ranging from 6.9% to 93% for individual samples (Figure S4). These numbers indicate that for some samples a large number of sorted cells neither lyse nor amplify. These trends differ between single-cell and multiple-cell sorts within the same species, as well as between species, indicating that SGA alone cannot be used for prediction of cell lysis-MDA efficiency in environmental eukaryotes, which was confirmed by our PCA analysis (Data S2).

Correlation between MDA start time and genome quality is shown in Figure S6: For four fungal species we observed a negative correlation between start of the genome amplification and genome completeness, in the other four and the ciliate species there is no correlation. Correlation between % positive WGA-MDA reactions and % positive rDNA-qPCR reactions are shown in Figure S4. Percent positive rDNA-qPCR reactions of the target species was based on the BLAST of the Sanger Sequencing

of the qPCR product for 1 cell sorts, 10 cells (20, 30, 50 in 3 cases) and 100 cells sorts (50 in 2 cases). We observed a positive linear correlation between % MDA positives and % PCR positives for all species. The number of confirmed target species by BLAST rDNA-PCR was significantly smaller than the number of total qPCR positives in most species, indicating to the fact that a lot of cells from the target population either contain a high number of prokaryote symbionts (in case of the two-step FACS, see Figure S1) or contaminants (for direct FACS, see Figure S1).

**Earlier start times for MDA do not always predict library quality.** Although four of the species had MDA start time inversely correlated with genome CEGMA, the support is much weaker (Data S2c and S4) than expected based on prokaryote single cell data (Clingenpeel et al., 2015). The correlation between MDA start time and assembled genome size overall is weak as well, see Data S2c and S4. Overall, start of the amplification time was concluded to be a poor QC criterion, instead the number of positive MDA reactions was a better predictor of the number of recovered target cells, which for most species correlated with a better co-assembled species genome. We found that fold amplification of the genome inversely correlates with genome quality (Figure 3 and S3) and can be used as a criterion, when genome size can be approximated. However, reducing the MDA total time to a minimum will aid to the quality of the genome due to reduction of the amplification bias.

**Optimization:** Lysis-MDA efficiency was evaluated using start of the genome amplification. Due to variation in the kinetics of the MDA reaction between each run, we used purified DNA, 100 cells, 50 cells and 10 cells as controls to normalize the single-cell MDA start (Data S3). For lysis-MDA reaction mix compatibility test, we first used purified genomic DNA from *E.coli* in the amount that equals to one *E.coli* cell (5-7 fg) (Supplemental Figure 4a). From the top panel, we observed, that detergent alone had a catalyst like effect on MDA kinetics, facilitating an early and congruent start of amplification comparing to either standard Alkaline1 lysis or no Lysis solution added. Such effect has been reported for other DNA polymerases (Zhulin et al., 2006), but not for phi29 polymerase. Using standard Alkaline1 lysis on purified DNA resulted in delayed start of the MDA, most likely due to DNA damage. To check this supposition, we reduced to 0.2x alkaline concentration, which resulted in an earlier start of amplification than higher concentration alkaline and similar to detergent alone as seen in the third panel of the figure. From the fourth panel is visible that 1mM EDTA in the composition of Lysis buffer does not inhibit the MDA reaction and that the combination of 0.3% Tween, 1mM EDTA and alkaline are fully compatible and enhance MDA to the same extent as detergent alone.

To test lysis efficiency for single-cells we first used axenic *E.coli* and *B. subtilis* cultures and environmental soil-dwelling single-cells and found most efficient lysis formulas (Data S3b and c). For the soil dwellers which are most difficult to lyse cells we improved cell lysis by adding 1mM EDTA to the Alkaline1 with 0.3% Tween lysis buffer (Data S3b). For fungal single-cell samples we chose three of the most promising approaches (Data S3d). Our results show that the lysis of *R. allomyces* single-cells in the lysis buffer containing detergent prior to the addition of KOH resulted in an earlier start of MDA, consistent with the DNA based tests in Data S3a. However, we found that a similar result on the amplification had replacing NEB MDA chemistry with the Qiagen REPLI-g single-cell WGA chemistry using just the standard alkaline lysis buffer (Data S3d). The latter chemistry allowed for a shorter hands-on and amplification time and was used for subsequent tests on another fungus, *C. protostelioides* (Data S3e): In this sample, the 100-cell control did start to amplify as early as in the *R. allomyces* sample with similar conditions. However, the single-cell MDA start in the *C. protostelioides* sample had a wider distribution. The use of detergent prior to the addition of alkaline delayed genome amplification in this fungus. Proteinase K addition to the cells prior the alkaline buffer lead to 100-cells and 10-cells amplification shift to an earlier point compared to the alkaline alone lysis buffer, however single-cell genome amplification was delayed. We postulated that due to efficient Proteinase K lysis, subsequent alkaline treatment caused some DNA damage reducing DNA amount and delaying the start of amplification. We verified this by diluting the alkaline used after Proteinase K treatment, and improved single-cell amplification significantly. Nevertheless, these results were very similar to the standard alkaline lysis results (Data S3e). We further explored the effect of alkaline concentration on *C. protostelioides* single-cell lysis (Data S3f). We found that the final concentration of 25mM alkaline as opposed to 10mM recommended in the standard lysis protocol for this chemistry had most beneficial effect on the start of genome amplification. This combination of the Lysis and MDA chemistry showed similar results for *R. allomyces* (Data S3f). This protocol (alkaline lysis at 0.25mM with DTT at 0.088mM final concentration in MDA, without other additives and incubate the cells at room temperature for 3-5minutes, prior the addition of the neutralizing buffer and MDA reaction) was the most succinct, eliminating additional steps and reagents that can increase the level of contaminating DNA and we used it for the other species in this study.

#### Step 4. SAG OTU(s) identification via rDNA

To identify both target OTU and possible contaminants carried over during FACS step or introduced via Lysis-MDA process we used universal and specific primers and Kappa SYBR Fast qPCR 2x mix (KK4611). The cycling conditions for the primers (Table S1) were as follows. For 16S (universal), the program was 95°C for 3 min, followed by 25 cycles (95°C for 10 sec, 56.8°C for 30 sec, 72°C for 45 sec). For ITS 18SCRYPTO, the program was 95°C for 3 min, followed by 30 cycles (95°C for 10 sec, 58.6°C for 30 sec, 72°C for 45 sec). For ITS 18SDPD, 18S\_SR, the program was 95°C for 3 min, followed by 28 cycles (95°C for 30 sec, 57.5°C for 30 sec, 72°C for 45 sec). All PCRs ended with a melting curve (65°C for 5 sec and 95°C for 30 sec) and cool down. A Bio-Rad CFX384 Real-time thermocycler was used for all qPCR reactions.

Sequenced fragments were treated with ExoSap-IT™ (Thermo Fisher Scientific, 78201.1. ML treated (37°C for 30 min, 80°C for 15 min), and either forward or reverse primer was added to the ExoSap treated mix, which was then submitted for sequencing at the UCB DNA Sequencing Core facility. Reaction volumes followed UC Berkeley DNA Sequencing core facility recommendations. The obtained sequences were analyzed by BLAST against the NCBI nucleotide or AFTOL databases.

qPCR of the rDNA followed by Sanger sequencing approach revealed on average 34% of the target OTU, ranging from 5.3% to 74% between samples (Figure S4). The limitation of this method was the inability to resolve multiple DNA sequences which occurred from either symbiotic or contaminating organisms or highly diverged copies of rDNA of the same species. For example, despite the high number of MDA and PCR positives for *D.cristalligena*, initially we found an extremely low rate of target OTU. When we examined all recovered rDNA sequences from this sample we observed a high rDNA divergence rate opposing high whole-genome similarity of this species (Figure 6a).

The Newbler assembler was used with in-house modifications to assemble a set of 18S sequences from the reads obtained from Illumina shallow sequencing. Briefly, an 18S HMM model was used for 18S rRNA assembly. The HMM-based tool uses *hmmsearch* against the model to pull reads for 18S rRNA assembly. *Hmmsearch* is sensitive when a sequence is not similar to anything in the database, and Newbler was found to produce few chimeras.

**Optimization:** We tested all the primers ranging from universal to taxa specific that target different rDNA regions from the original sources listed in Table S1. We selected the most reliable and broad range primers, shown in Table S1. Due to limitations of the rDNA qPCR followed by Sanger Sequencing (see above), we tested a different approach for screening: A shallow sequencing step (illustrated in the next step5 of the pipeline) originally introduced to screen out biased genomes and low-quality libraries was tested as an alternative approach for OTU identification, via rDNA assembly. We tested a number of rDNA assembly methods combined with different library creation methods and different Illumina sequencing platforms and uncovered a wide discrepancy between approaches. We benchmarked this approach against rDNA qPCR-Sanger results and rDNA from whole genome assembly (Figure S5). We found that several bioinformatics tools failed to assemble correct rDNA from the NGS reads of the MDA amplified genome. Some rDNA assembly methods performed better or worse depending sequencing quality and sequencing platform. None of the tools had same accuracy as the rDNA-PCR followed by Sanger sequencing. One of the tools (Newbler) had higher accuracy relative to other tested bioinformatics tools. We further improved this algorithm and named it NewblerWA (after William Andreopoulos) who modified existing tool (see above) and increased the accuracy and taxonomic resolution level (from phylum to species or genus).

Thus, OTU identification prior genome sequencing allowed further screening out undesirable for sequencing single-cell genomes and thus reduced the pipeline costs. Steps 4 and 5 can be combined into a single step to further decrease costs and increase the recovery of genomes that have endosymbionts and to minimize false negatives caused by poor PCR amplification. We found that correct combination of the sequencing and rDNA assembly method were able to detect both target and symbiont rDNA OTU and evaluate genome amplification bias in one shallow sequencing step. This approach can be further modified for assembly of other marker DNA regions in addition to rDNA or instead of rDNA, when dealing with current poor representation of early diverging eukaryotic species in rDNA databases.

#### Step 5. NGS library and SAG genome quality screening

In step 5 we implemented shallow sequencing of the NGS libraries for SAG quality screening. This step was automated through a JGI pipeline accessible to JGI users at <https://rqc.jgi-psf.org/> and is described in detail below. Two essential steps were adjusted for the EME single-cell genomics pipeline: 1. For each read in the sequence data, a set of 20 bases (20-mer) was selected from a random starting position in the read and stored in a hash function. If the 20-mer already existed in the hash function, a

counter was incremented to indicate the number of times in which the 20-mer was seen. Every 25,000 reads, the uniqueness was calculated by dividing the number of unique 20-mers seen by the total number of reads sampled. 2. The contamination used BB'Tools' seal program: <https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/seal-guide/>

For the NGS library and SAG quality screening after shallow sequencing we used JGI Read QC (RQC) pipeline for a quick and inexpensive way to estimate the quality of the genomes in hand. Pipeline details are here <https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/data-preprocessing/> ) and described here:

### Read QC pipeline metrics criteria:

#### Illumina Read Quality metrics

#### Read size distribution

#### Read GC%

#### Read random twentymer uniqueness

#### Contaminant %

#### Table of organisms reads map to with percentage

#### Mitochondria and Ribosomal %

#### Read QC Pipeline

The Read QC pipeline performs QC for Illumina sequencing

Command: `module load jgi-rqc; readqc.py --fastq FASTQ_FILE --output-path OUTPUT_PATH [--skip-cleanup --skip-subsample --skip-blast --skip-localization]`

Parameter	Meaning
FASTQ_FILE	Gzip'd or raw fastq
OUTPUT_PATH	File system location to run analysis and store results

#### Toggle Options

- "--cut": set read cut length (bp) for read contamination detection (default: 50bp)
- "--skip-cleanup": skip cleaning temporary files
- "--skip-subsample": skip subsampling of the input fastq
- "--skip-blast-nt": skip BLAST search against nt
- "--skip-blast-refseq": skip BLAST search against refseq.archaea and refseq.bacteria
- "--skip-localization": skip localization of BLAST reference database files

**i** *in RQC framework, the raw fastq file is used as the input.*

readqc.log is the main log file that shows the log time and pipeline step.

Qsub options: `-b yes -j yes -m n -w e -terse -`

`l ram.c=5.25g,h_vmem=5.25g,disk.c=20G,h_rt=43199,s_rt=43194 -pe pe_slots 8`

Database description: [https://docs.google.com/document/d/1RxgLplaEzy0QJTnJO\\_nlbPWCuxGqu-YuOVpqy\\_FQh0w](https://docs.google.com/document/d/1RxgLplaEzy0QJTnJO_nlbPWCuxGqu-YuOVpqy_FQh0w)

#### Read QC Process

1. Read subsampling

- module load bbtools; reformat.sh in=IN out=OUT samplerate=0.01 qin=33 qout=33 ow=t gcplot=t bhis t= qhist= gchist= gcbins=auto bqhist= bqhist=
  - 2. Unique 20-mer/25-mer analysis
    - module load bbtools; bbcountunique.sh k=[20 or 25] interval=25000 in=IN out=OUT percent=t count=t cumulative=f int=f ow=t
  - 3. GC analysis
    - Generates GC statistics and histogram plots
  - 4. Read quality checking
    - Generates read quality plots
  - 5. Base quality checking
    - Generates base quality statistics
  - 6. Quality score analysis
    - Generates quality score statistics and plots
  - 7. 21-mer analysis
    - Skipped
  - 8. Common motifs checking
    - patterN\_fastq.pl -analog -PCT 0.1 -in IN > OUT
  - 9. Duplicates removing
    - Accepts one or more files containing sets of sequences (reads or scaffolds). Removes duplicate sequences, which may be specified to be exact matches, subsequences, or sequences within some percent identity.
- module load bbtools; dedupe.sh in=IN out=null qin=33 ow=t s=0 ftr=49 ac=f int=f> OUT 2>&1
- 10. Tag dust
    - Skipped
  - 11. Contamination detection
    - module load bbtools; seal.sh in=IN out=null ref=[reference file] k=22 minskip=7 hdist=0 stats=OUT k=22 hdist=0 ow=t

Reference file location:

Reference file	File location
ARTIFACT (no spikein)	/global/dna/shared/rqc/ref_databases/qaqc/databases/illumina.artifacts /Illumina.artifacts.2012.10.no_DNA_RNA_spikeins.fa
ARTIFACT (first 50bp)	/global/dna/shared/rqc/ref_databases/qaqc/databases/illumina.artifacts /Illumina.artifacts.2012.10.no_DNA_RNA_spikeins.fa
ARTIFACT (DNA spikein)	/global/dna/shared/rqc/ref_databases/qaqc/databases/illumina.artifacts /DNA_spikeins.artifacts.2012.10.fa.bak
ARTIFACT (RNA spikein)	/global/dna/shared/rqc/ref_databases/qaqc/databases/illumina.artifacts /RNA_spikeins.artifacts.2012.10.NoPolyA.fa

CONTAMINANTS	/global/dna/shared/rqc/ref_databases/qaqc/databases/JGIContaminants.fa
FOSMID	/global/dna/shared/rqc/ref_databases/qaqc/databases/pCC1Fos.ref.fa
MITOCHONDRION	/global/dna/shared/rqc/ref_databases/qaqc/databases/ncbi.refseq/refseq.mitochondrion.fa
PHIX	/global/dna/shared/rqc/ref_databases/qaqc/databases/phix174_ill.ref.fa
PLASTID	/global/dna/shared/rqc/ref_databases/qaqc/databases/ncbi.refseq/refseq.plastid.fa
RRNA	/global/dna/shared/rqc/ref_databases/qaqc/databases/rRNA.fa
NON-SYNTHETIC	/global/projectb/sandbox/gaag/bbtools/commonMicrobes/fusedERPBBmasked.fa.gz
SYNTHETIC	/global/projectb/sandbox/gaag/bbtools/data/Illumina.artifacts.2013.12.no_DNA_RNA_spikeins.fa.gz
ADAPTERS	/global/projectb/sandbox/gaag/bbtools/data/adapters.fa

Additional information: [Microbe Read Filtering: SOP 1077](#)

#### 12. Sciclone analysis

- module load bbtools; bbdduk.sh in=IN ref= out=null fbm=t k=31 mbk=0 stats=OUT statscolumns=3

#### 13. Subsampling for Blast search

- module load bbtools; reformat.sh in=IN out=OUT samplerate=RATE qin=33 qout=33 ow=t or  
module load bbtools; reformat.sh in=IN out=OUT samplereadtarget=25000 qin=33 qout=33 ow=t

#### 14. Blast search vs. refseq.archaea

- Default Blast options: -evalue 1e-30 -perc\_identity 90 -word\_size 45 -task megablast -show\_gis -dust yes -soft\_masking true -num\_alignments 100 -outfmt '6 qseqid sseqid bitscore evalue length pident qstart qend qlen sstart send slen staxids salltitles'

```
module load jgi-rqc; run_blastplus.py -d refseq.archaea -o OUTDIR -q QUERY -s > blast.log 2>&1
```

#### 15. Blast search vs. refseq.bacteria

- module load jgi-rqc; run\_blastplus.py -d refseq.bacteria -o OUTDIR -q QUERY -s > blast.log 2>&1

#### 16. Blast search vs. nt

- module load jgi-rqc; run\_blastplus.py -d nt -o OUTDIR -q QUERY -s > blast.log 2>&1

#### 17. Multiplex analysis

#### 18. Adapter checking

- kmercount\_pos.py --plot PLOT /scratch/rqc/Artifacts.adapters\_primers\_only.fa IN > OUT

#### 19. Insert size analysis



- module load bbtools; bbmerge.sh in=IN hist=OUT reads=1000000

## 20. GC divergence analysis

- module load R/3.2.4; module load jgi-fastq-signal-processing/2.x; format\_signal\_data --input IN --output OUT --read both --type composition
- module load R/3.2.4; module load jgi-fastq-signal-processing/2.x; model\_read\_signal --input IN --output OUT

## 21. Post-processing

## 22. Cleanup

### **Optimization: Genome amplification bias early detection and proposed reduction**

Several read quality metrics produced by this pipeline were used to evaluate their predictability for genome completeness (Data S2). Two of these criteria: Random 20-mer uniqueness (RTU) and contaminant percent proved especially useful for predicting genome quality (Figure 3 and S3). RTU was found to be predictable of the amplification bias. Thus, RTU value above 60% correlated with nearly complete genomes and RTU value below 10% guaranteed highly incomplete genomes. A cut-off of the reagent contaminant carry-over below 3% proved to be efficient for subsequent steps.

From all QC criteria listed in Data S2a, random 20-mer uniqueness (RTU) proved to be most useful for amplification bias assessment. For this criterion, we chose 1 million reads input as a quality prediction cut-off and examined the results across 9 species, illustrated in Data S2 and Figure 3. Overall, our data confirmed increased genome amplification bias (GAB) with fold of amplification, e.g. all genomes were amplified for the same amount of time and end quantity, in which case smaller genomes were exposed to higher fold of amplification than larger genomes. Thus, smaller genomes showed a higher amplification bias (GAB) than larger genomes (Figure 4,5 and 7). *C. protostelioides* single-cells had lowest RTU and highest amplification bias, however *C. protostelioides* genome size is similar to two other species, which showed better RTU and less genome amplification bias (Figure 5). Worst *C. protostelioides* amplification bias occurred in the higher than 65% GC regions (Figure S6). Our attempt to correct the situation, using high GC% hexamers during amplification, resulted in very poor read quality (not shown here). Because MDA chemistry should be GC-bias free and Illumina sequencing was reported to be biased against high GC% regions we compared the amplified DNA with isolate DNA results for coverage level and found that the isolate DNA had a mean of 25.46-fold coverage with StDev of 53.57 and the co-assembly had a mean of 55-fold coverage with StDev of 88.5. We considered bias due to specific DNA structures and looked at the structure of these regions. We did not find any long homopolymeric stretches in the biased areas. We found mostly coding regions for a number of proteins (Figure S6 and Table S6). We excluded poor lysis because single-cell lysis efficiency was high (and high % of target OTU) with early start of amplification; we excluded amplification bias during Illumina sequencing because the isolate unamplified DNA underwent 20 cycles of amplification after library construction to meet sequencer loading needs, while the MDA amplified genomes had unamplified libraries; we excluded coverage bias because MDA amplified genomes had twice higher read coverage than the unamplified genome. Therefore, we conclude that the missing regions in the co-assembly are not due to the low coverage of the amplified DNA, but rather due to high GC% regions - MDA amplification bias.

In summary, shallow sequencing of libraries in Step5 was found to be essential for weeding-out low-quality genomes and was necessary for significant cost-saving when working with genomes larger than 8-10 Mb.

### **Step 6. Single-Cell Genome assembly and coassembly**

Two of the target species were used to benchmark various genome assemblers of the Illumina reads for amplified single-cell genomes. Genome assembly quality was judged using a set of criteria from the QUAST software (Tables S2-4 and Data S2) and their correlation with CEGMA (Parra et al., 2007) (Data S2) as a measure of genome completeness. We tested the use of long read: PacBio platform and LMP-Illumina libraries and short-read Illumina sequencing platforms: Due to the formation of the long chimeric regions during MDA, long read sequencing technology was not suitable for the MDA-amplified single cell genomes, where the short reads (150-600bp) performed the best (Figure S7, Tables S2-4). Illumina LMP library did not provide a significant improvement of the short-read assembly made from the same single cell MDA genome (e.g. 0.52% reads aligned to long edges). For the protist genome (Table

S2,S3), in our tests, the standard JGI prokaryote single-cell amplified genome (sag) pipeline (IDBA+Allpaths) (Peng et al., 2012, Butler et al., 2008) is estimating a large assembly size but only assembling a small fraction of that; IDBA-UD (Peng et al., 2012) produces more fragmented assemblies but has a reasonable assembled genome size; the metagenome pipeline produces a reasonable sized assembly with the largest pieces; SPAdes 2.4 (Bankevich et al., 2012) also produces a smaller than expected genome size. As a result of these tests a combination of normalization of the read coverage with the sag pipeline and subsequent assembly with the metagenome pipeline produced longest contigs and assembled a reasonable size genome (Tables S2-4). For our test fungal genome, IDBA-UD and IDBA+Allpaths failed to run before finishing and could not be used. For the fungal genome, SPAdes Single Cell v2.4 (subsequently replaced by SPAdes Single Cell v3.6 and higher) performed the best in terms of time, number of contigs and assembled genome size (Table S4). This assembler was used for the rest of the fungal amplified single or multiple cell genome libraries.

**Optimization:** Our results show that for medium size genomes (12-30Mb) Single Cell SPAdes assembler v2.4 and higher (Bankevich et al., 2012) performed the best, while for large genomes (>100 Mb) SOAP (Luo et al., 2012) performed the best. We examined 16 criteria to assess genome assembly quality and as predictors of high genome completeness (Data S2). Many of them did correlate with assembly CEGMA value and genome size and the number was reduced due to redundancy. The number of scaffolds in the range of 10-25kb correlated directly with assembled genome size, while main genome scaffold\_N50 and the number of scaffolds between 2-10kb directly correlated with predicted genome size, usually larger than assembled. The number of scaffolds in the range of 25-50kb correlated with a higher CEGMA and less with assembled or predicted genome size. Interestingly, assembled genome size and predicted genome size do not correlate as strongly as expected with CEGMA genome completeness, perhaps due to a high non-coding proportion in EME genomes.

Besides single-cell genome assemblies, we tested two strategies for co-assembly of the amplified genomes from the same target OTU from individual libraries: (1) all libraries combined (Table S5) and (2) a selection of fewer individual libraries with the highest CEGMA values (Figure 4-7). Our results showed that the second approach is not only faster, but also can result in co-assemblies with larger genome and/or CEGMA values (Figure 4-7). To produce the co-assembly: Data from multiple single-cell runs were combined in a single fastq file to produce a co-assembly for a species. The fastq files were normalized with bbnorm to bring the coverage to a uniform level; this step reduced the co-assembly runtime drastically since MDA coverage bias caused some areas to have very high coverage. SPAdes 3.6 without the error correction step was used on the combined normalized fastq file. From the co-assembly, only the scaffolds of length 2 KB or longer were kept, in order to remove contaminants. In our experience, more than 50% of the contigs of 2 kb and smaller tend to be phylogenetically ambiguous and thus require manual curation; therefore, their use is strongly advised against in an automated pipeline. These contigs were saved as a separate pool for optional manual organelle assembly and/or symbiont/contaminant assembly.

Co-assembly of individual amplified single- or multiple- cell improved genome quality for several species (Figure 7a), for a few species the multiple-cell genome assembly produced as high CEGMA as the co-assembly (Figure 7d). **The improved quality of the co-assembly is due to random amplification bias**

#### **Steps 4, 5, and 6. Phylogenetic and phylogenomic calculations**

18S rDNA trees were constructed using 18S sequences obtained from the amplified single-cell genomes OTU-Sanger screening step. All sequences were verified against the assembled genome, and ambiguous (N) Sanger sequencing reads were corrected with Illumina reads, if necessary. All sequences were trimmed to the same region (v6-v9) or used as full length (for the ciliate protist). For the outgroups or related species and genera, the 18S sequences were originally obtained from NCBI and manually curated. Each sequence set was aligned using MAFFT (Katoh et al., 2005; Yamada et al., 2016) using TREX server (Yamada et al., 2016) and manually corrected to reposition gaps if necessary. Phylogenetic trees were calculated and constructed using PhyML on TREX and ATGC (Guindon et al., 2010; Lefort et al., 2017). Optimal parameters for each set were selected and used for the version presented here.

Accurate evaluation of the phylogenetic identity of individual libraries is essential before co-assembly. Thus, we compared the use of 18S rDNA and whole-genome distance of each single-cell (Figure 4 and 6). Our choice of the 18S instead of the ITS for fungal species was based on the lower or null availability of the ITS sequence in the public databases for the early diverging fungi and protists. On the contrary 18S rDNA has been used as a phylogenetic tool for a while to assess early diverging fungi and protists diversity (Berbee et al., 2017, Lazarus et al., 2015, Caron et al., 2009). For all but one target, the majority of the single-cell rDNAs constituted one taxonomic unit with short phylogenetic distance (Figure 4 and 6). For one species, *D. cristalligena* the distance between some of the single-cell rDNA was

as big as the interspecific distance with *D. bacillispora* (Figure 6). *D. cristalligena* single-cells were isolated from one sporulation event, implying a low probability of different strains, and excluding the possibility of different species, indicating a very high evolutionary rate of the 18S rRNA in this species (Figure 6).

Because of the possibility of rDNA evolutionary rate being different from the whole genome evolution rate, we tested the usability of the genome-to-genome comparison tools used routinely for prokaryotes (e.g. ANI, GGDC) to evaluate intraspecific genome similarity between single-cells, or low number of cells for low input amplified-DNA genomes. Average nucleotide identity (ANI) analysis of the seven closest ciliate genomes, revealed that about 55-62% of the genome of the used single cells has 98.8% and higher identity (Table S8), while the other half of the genome has lower than 70% identity. For the same organism genome-to-genome distance calculator (GGDC, formula 2 only) was able to calculate entire genome distance (Figure 6c and Table 3) with very high confidence. Similarly, ANI could not be completed on *R. allomycis* genome. Contrary to ANI, GGDC performed very well both for fungi and protists (Figure 4 and 6, and Table 3). Although RaxML can be used when annotated genomes are available, it is more resource intensive and depends on the annotation pipeline which is computationally more expensive than GGDC for medium and large genomes.

**Optimization:** We tested the use of 18S rDNA sequences of each single-cell, ANI ([Han et al., 2016](#)) and GGDC ([Auch et al., 2010](#), [Meier-Kolthoff et al., 2013](#), [Riley et al., 2016](#)) to determine phylogenetic distance of the single-cell genomes and to group closely related genomes under the umbrella of one species prior to the co-assembly step (Figure 4 and 6, Table 3 and S8). Our results show that for most species rDNA phylogeny is a valuable tool, but is hampered by instances of unusual divergence rate of rDNA for some species. A more robust and accurate evaluation of the whole genome distance was achieved by GGDC formula 2, that was found to be useful for incomplete, amplified genomes (Figure 4 and 6, Table 3). This tool was developed and tested on prokaryote genomes and some fungal unamplified small genomes before ([Auch et al., 2010](#), [Meier-Kolthoff et al., 2013](#), [Riley et al., 2016](#)). Our results showed that this tool is of great use for medium and large eukaryotic genomes obtained via MDA amplification. Another genome distance calculator: ANI, used successfully for prokaryotic genomes ([Han et al., 2016](#)) failed the test for eukaryote genomes (Table S8).

### **Step7: Annotation of amplified genomes for functional predictions**

For genome annotation we used an existing pipeline described in [Grigoriev et al., 2014](#). Measuring genome completeness for de-novo assemblies is an imperative requirement for quality evaluation. However, only approximate estimates could be obtained using mathematical algorithms. Two tools developed for eukaryotic genomes looked most promising: CEGMA ([Parra et al., 2007](#)) and BUSCO ([Simão et al., 2015](#)). We used CEGMA for our pipeline evaluation and later tested newly developed BUSCO. Unexpectedly BUSCO did perform worse (detected less genes) than CEGMA for the early diverging fungi. BUSCO inaccurate performance for early diverging fungi could be due to lower availability of a statistically significant number of early diverging fungi of a specific phylum and high diversity within phylum. We decided not to use this engine until a larger database of early diverging fungal annotated genomes is acquired.

## Supplemental References:

- Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Pribelski, A.D., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19, 455–477.
- BD Influx™ Cell Sorter User's Guide. (2011) [bdbiosciences.com](http://bdbiosciences.com) 23-11543-00 Rev. 01 4/2011
- Berbee, M.L., James, T.Y., and Strullu-Derrien, C. (2017). Early diverging fungi: diversity and impact at the dawn of terrestrial life. *Annu. Rev. Microbiol.* 71, 41–60.
- Benny, G.L., Ho, H.M., Lazarus, K.L., and Smith, M.E. (2016). Five new species of the obligate mycoparasite *Syncephalis* (Zoopagales, Zoopagomycotina) from soil. *Mycologia* 108, 1114–1129.
- Caron, D.A., Countway, P.D., Savai P., Gast, R.J., Schnetzer, A., Moorthi, S.D., Dennett, M.R., Moran, D.M., and Jones, A.C. (2009). Defining DNA-Based Operational Taxonomic Units for Microbial-Eukaryote Ecology. *Applied and environmental microbiology*, Sept. Vol. 75, No. 18, 5797–5808.
- Dawson S.C. and Pace N.R. (2002). Novel kingdom-level eukaryotic diversity in anoxic environments. *PNAS.* 99, 8324–8329.
- Grigoriev, I.V., Nikitin, R., Haridas, S., Kuo, A., Ohm, R., Otilar, R., Riley, R., Salamov, A., Zhao, X., Korzeniewski, F., et al. (2014). MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Res.* 42, D699–D704.
- Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307–321.
- James, T.Y., Pelin, A., Bonen, L., Ahrendt, S., Sain, D., Corradi, N., and Stajich, J.E. (2013). Shared signatures of parasitism and phylogenomics unite Cryptomycota and Microsporidia. *Curr. Biol.* 23, 1548–1553.
- Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I.A., Belmonte, M.K., Lander, E.S., Nusbaum, C., and Jaffe, D.B. (2008). ALLPATHS: *de novo* assembly of whole-genome shotgun microreads. *Genome Res.* 18, 810–820.
- Katoh, K., Kuma, K., Toh, H., and Miyata, T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* 33, 511–518.
- Lazarus, K.L., Benny, G.L., Ho, H.M., and Smith, M.E. (2017). Phylogenetic systematics of *Syncephalis* (Zoopagales, Zoopagomycotina), a genus of ubiquitous mycoparasites. *Mycologia* 109, 333–349.
- Lefort, V., Longueville, J.E., and Gascuel, O. (2017). SMS: smart model selection in PhyML. *Mol. Biol. Evol.* 34, 2422–2424.
- Meier-Kolthoff, J.P., Auch, A.F., Klenk, H.P., and Göker, M. (2013). Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* 14, 60.
- Pang, Z., Al-Mahrouki, A., Berezovski, M., Krylov, S.N. (2006) Selection of surfactants for cell lysis in chemical cytometry to study protein-DNA interactions. *Electrophoresis* 27, 1489–1494.
- Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 23, 1061–1067.

- Peng, Y., Leung, H.C., Yiu, S.M., and Chin, F.Y. (2012). IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28, 1420–1428.
- Riley, R., Haridas, S., Wolfe, K.H., Lopes, M.R., Hittinger, C.T., Göker, M., Salamov, A.A., Wisecaver, J.H., Long, T.M., Calvey, C.H., et al. (2016). Comparative genomics of biotechnologically important yeasts. *Proc. Natl. Acad. Sci. U. S. A.* 113, 9882–9887.
- Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212.
- Yamada, K.D., Tomii, K., and Katoh, K. (2016). Application of the MAFFT sequence alignment program to large data-reexamination of the usefulness of chained guide trees. *Bioinformatics (Oxford, England)* 32, 3246–3251.