iScience, Volume *24*

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+/- 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/ 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.

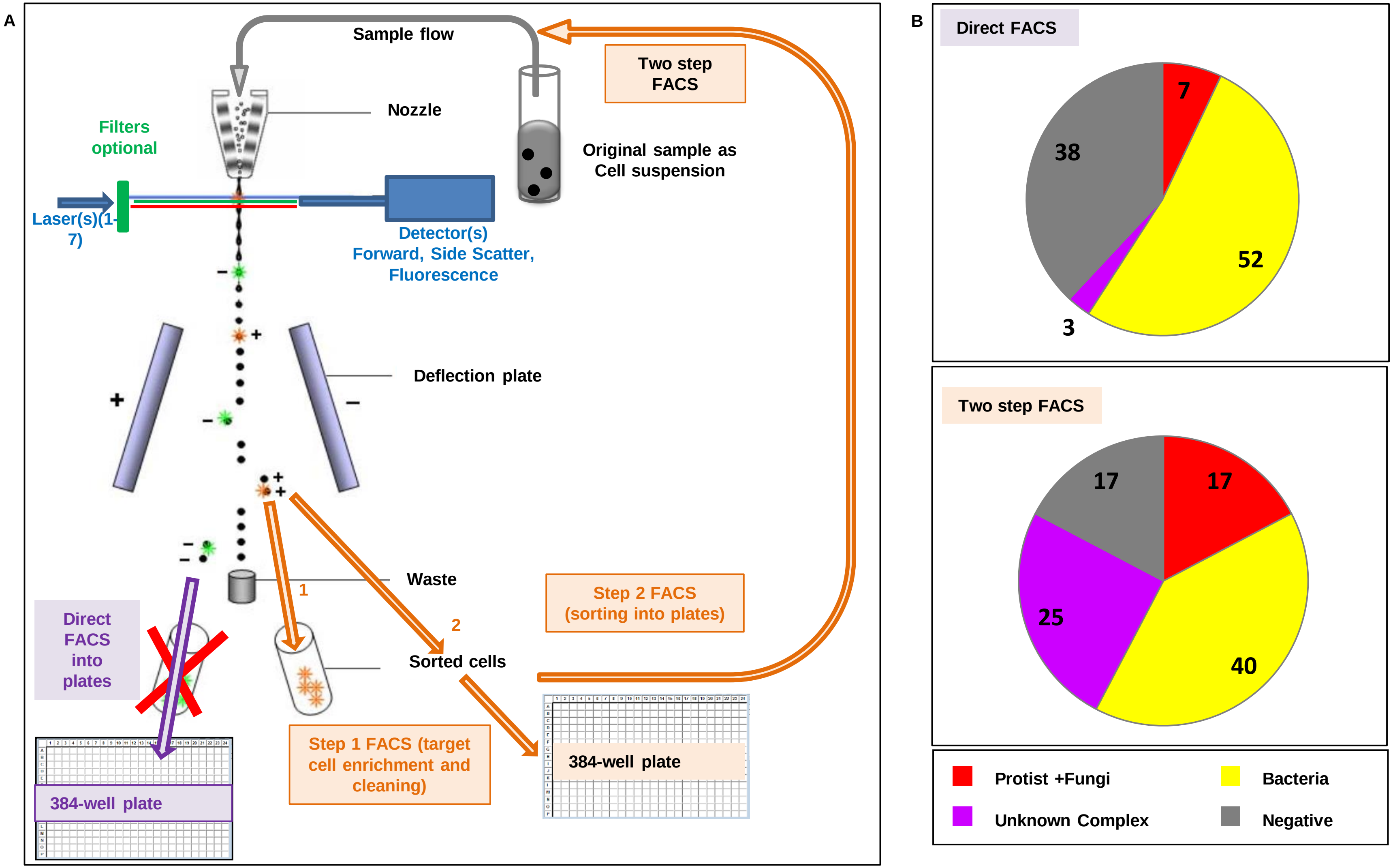


Figure S1. Target EME Single-Cell Isolation from Environmental Samples. Related to Figure 1, step 2. A. Schematics showing FACS target enrichment procedure (two-step FACS). FACS schematics was adapted from BD Influx™ Cell Sorter User's Guide. **B.** Target (here Protist +Fungi) recovery efficiency (numbers 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.

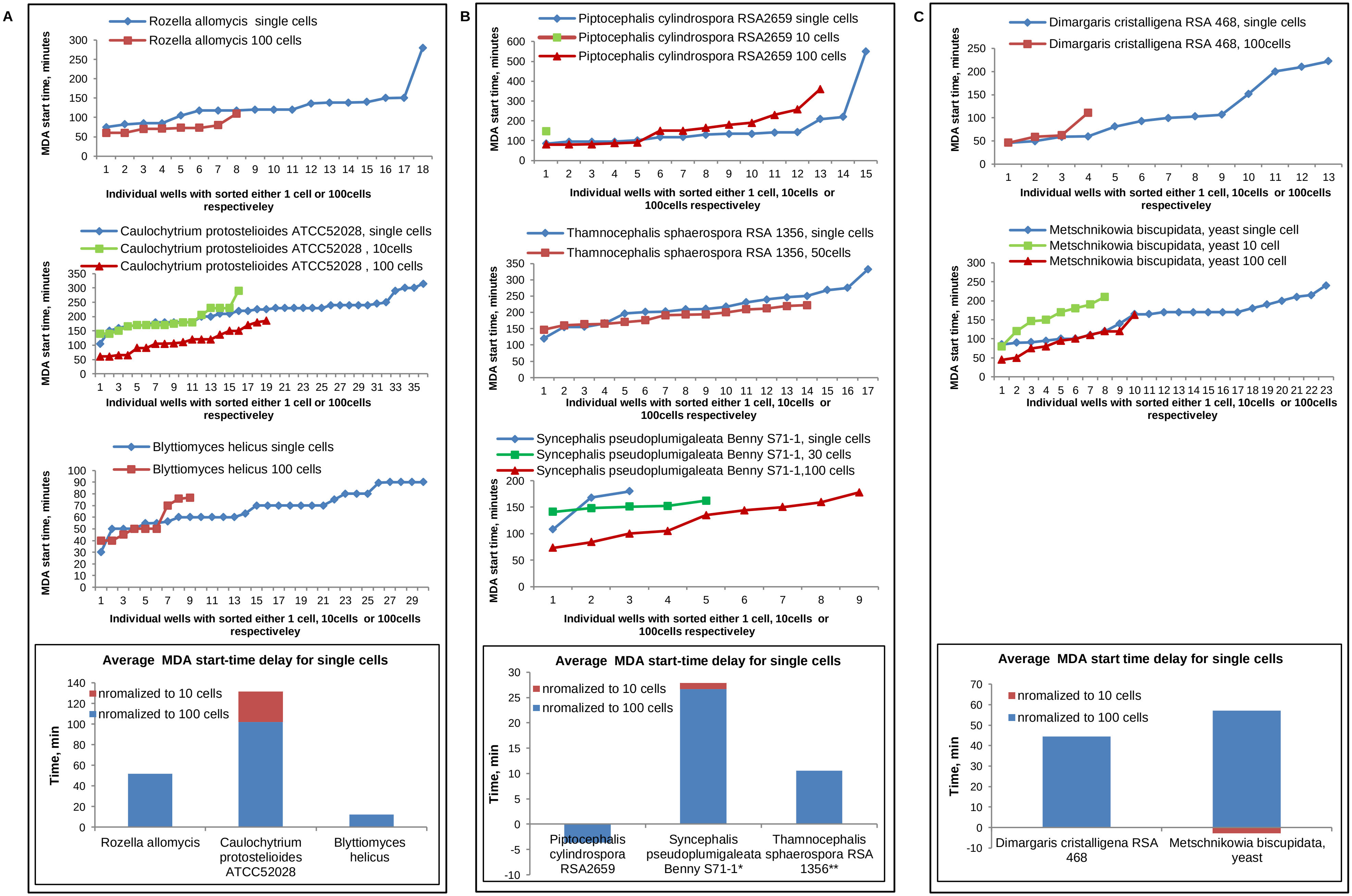
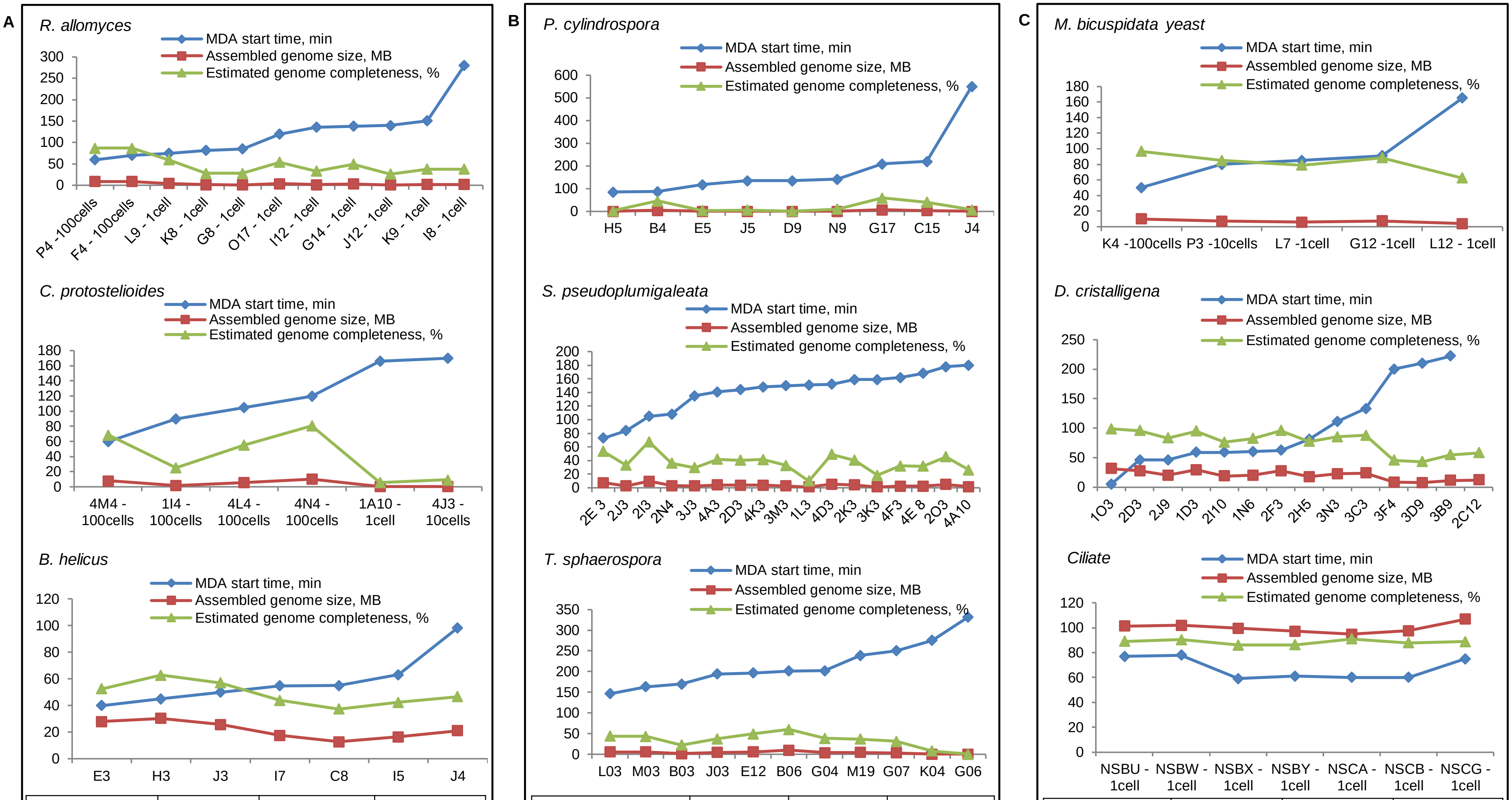


Figure S2. MDA start time comparison for confirmed target species. Related to Figure 1, step 3. and Data S4. Top to bottom – first three plots show raw amplification start times for individual wells sorted with respective number of cells. Bar-graph shows normalized single-cell to 100 cells and 10 cells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

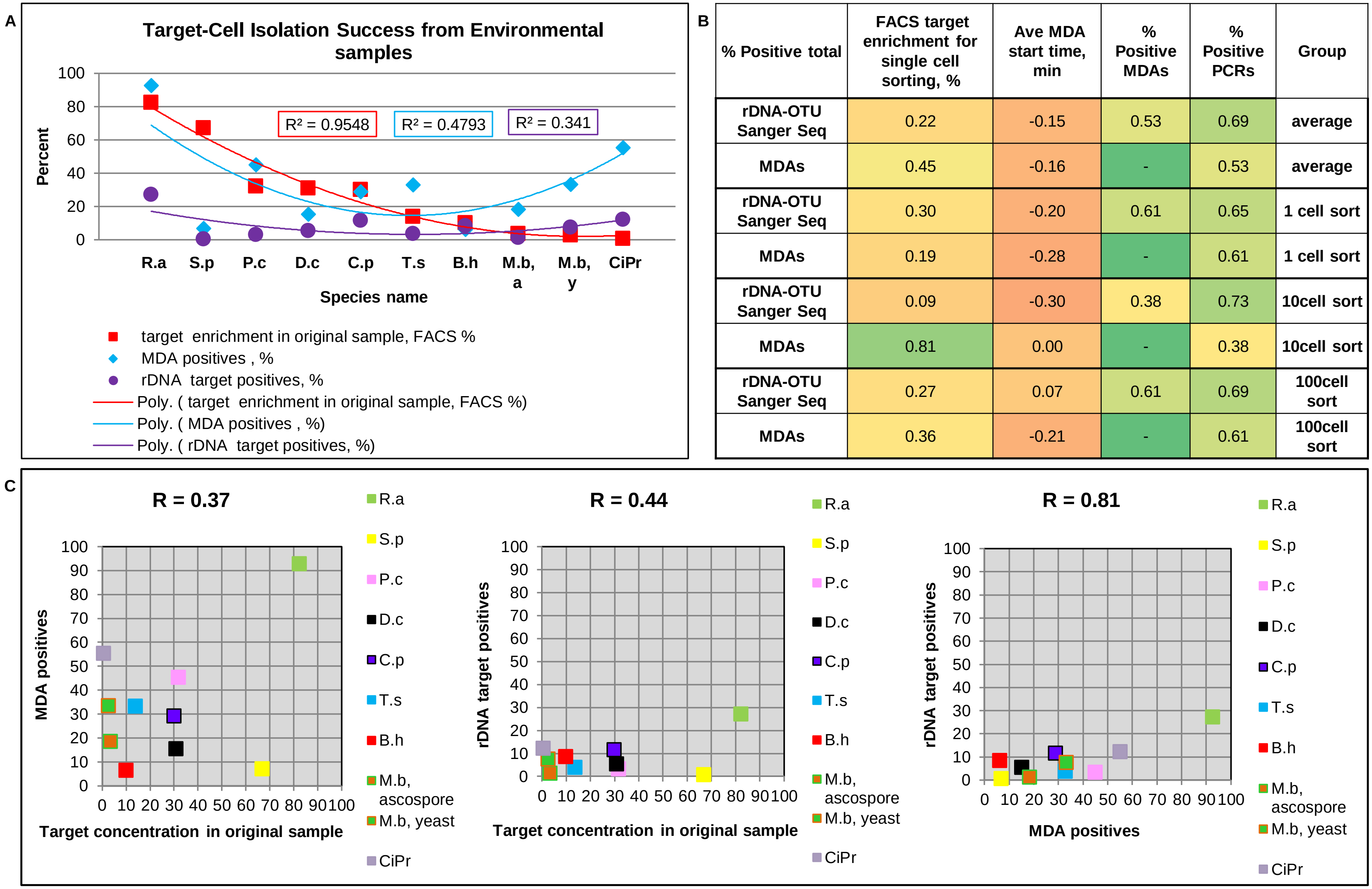


Species	Correlation: MDA start time with genome size	Correlation: genome size with genome completeness	Correlation: MDA start time with genome completeness
R. allomyces	**-0.45**	**0.99**	**-0.43**
C. protostelioides	**-0.58**	**0.99**	**-0.67**
B. helicus	**-0.36**	**0.96**	**-0.37**

Species	Correlation: MDA start time with genome size	Correlation: genome size with genome completeness	Correlation: MDA start time with genome completeness
P. cylindrospora	**-0.05**	**0.99**	**-0.04**
S. pseudoplumigaleata	**-0.49**	**0.95**	**-0.41**
T. sphaerospora	**-0.62**	**0.96**	**-0.72**

Species	Correlation: MDA start time with genome size	Correlation: genome size with genome completeness	Correlation: MDA start time with genome completeness
M. biscupidata	**-0.90**	**0.97**	**-0.95**
D. cristalligena	**-0.84**	**0.98**	**-0.88**
Ciliate	**0.77**	**0.09**	**0.49**

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. cristalligena*) fungi.



% Positive total	FACS target enrichment for single cell sorting, %	Ave MDA start time, min	% Positive MDAs	% Positive PCRs	Group
rDNA-OTU Sanger Seq	0.22	-0.15	0.53	0.69	average
MDAs	0.45	-0.16	-	0.53	average
rDNA-OTU Sanger Seq	0.30	-0.20	0.61	0.65	1 cell sort
MDAs	0.19	-0.28	-	0.61	1 cell sort
rDNA-OTU Sanger Seq	0.09	-0.30	0.38	0.73	10cell sort
MDAs	0.81	0.00	-	0.38	10cell sort
rDNA-OTU Sanger Seq	0.27	0.07	0.61	0.69	100cell sort
MDAs	0.36	-0.21	-	0.61	100cell sort

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.

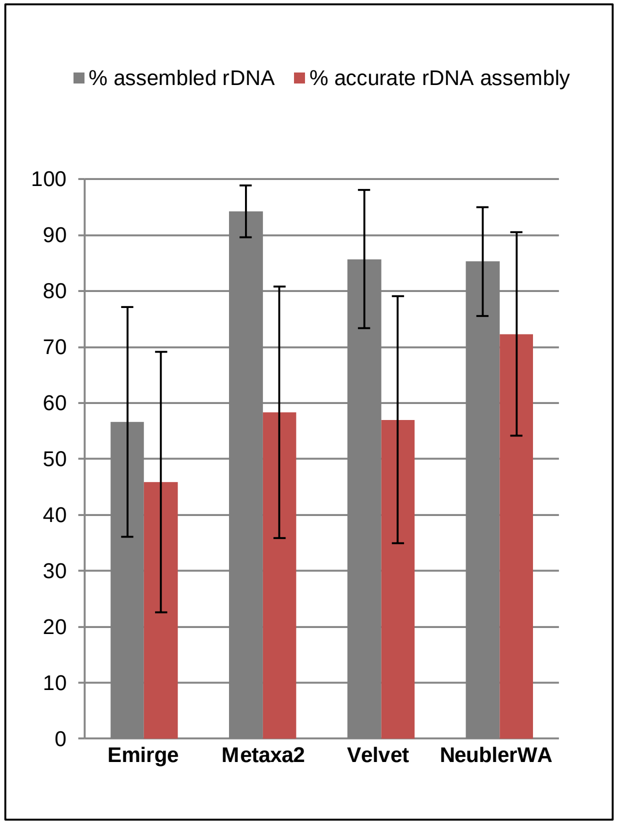


Figure S5. rDNA assembly and OTU identification tools evaluation. Related to Figure 1, step 4. Shown results are average for 8 fungal species (over 80 libraries) with standard deviation between species.

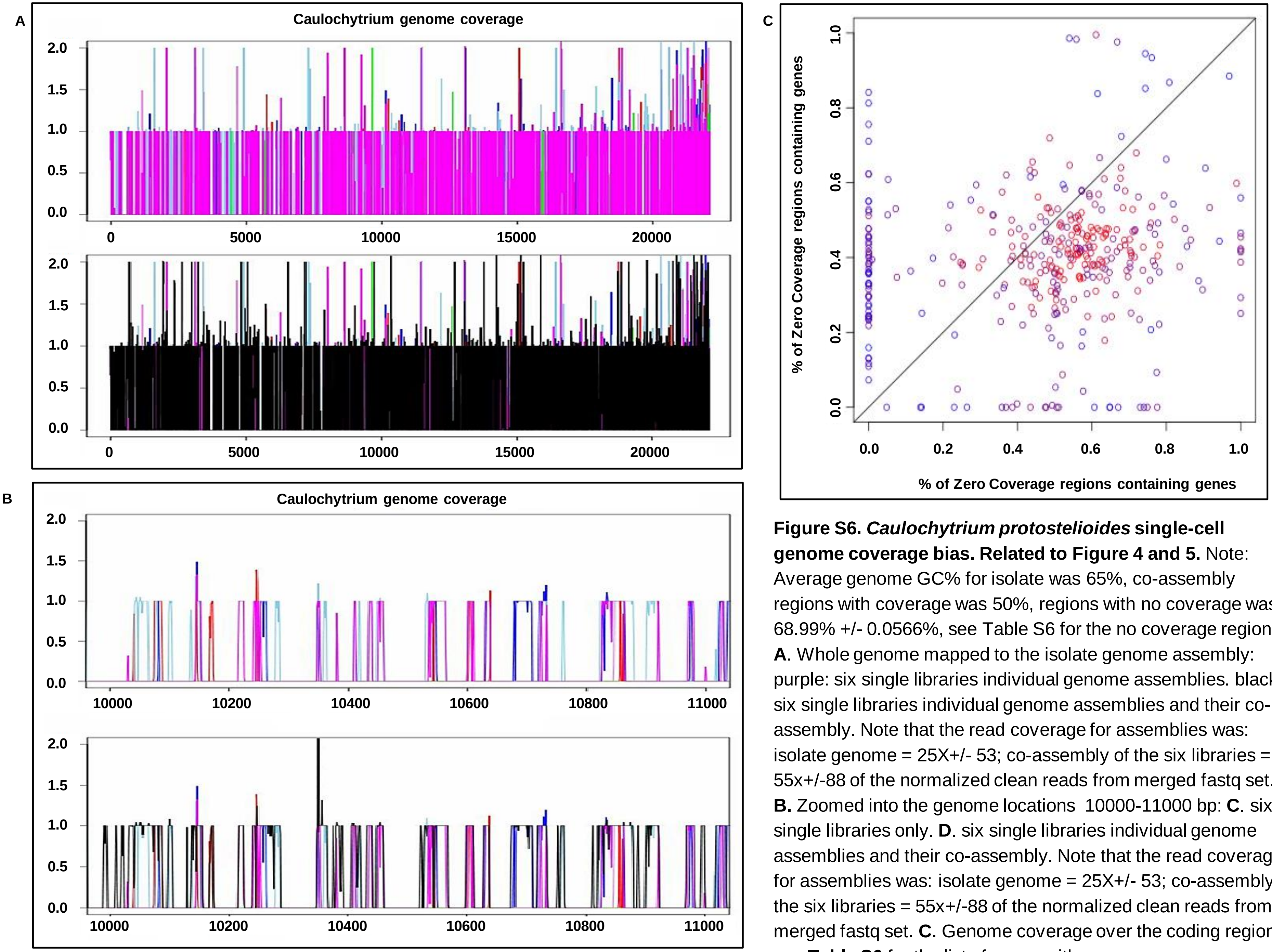
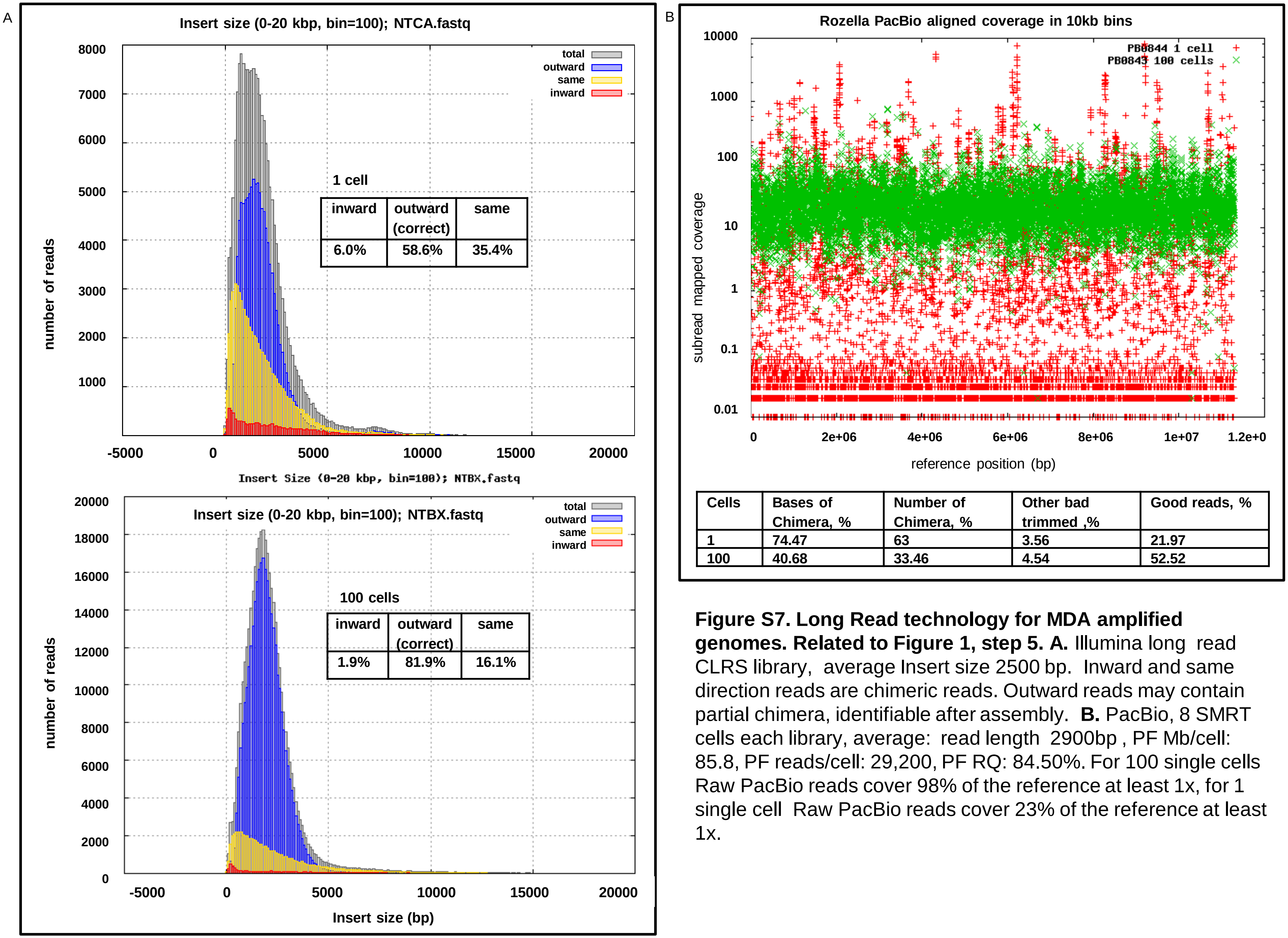


Figure S6. ***Caulochytrium protostelioides*** **single-cell genome coverage bias. Related to Figure 4 and 5.** Note: Average genome GC% for isolate was 65%, co-assembly regions with coverage was 50%, regions with no coverage was 68.99% +/- 0.0566%, see Table S6 for the no coverage regions. **A.** Whole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



1 cell

inward	outward (correct)	same
6.0%	58.6%	35.4%

100 cells

inward	outward (correct)	same
1.9%	81.9%	16.1%

Cells	Bases of Chimera, %	Number of Chimera, %	Other bad trimmed ,%	Good reads, %
1	74.47	63	3.56	21.97
100	40.68	33.46	4.54	52.52

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.

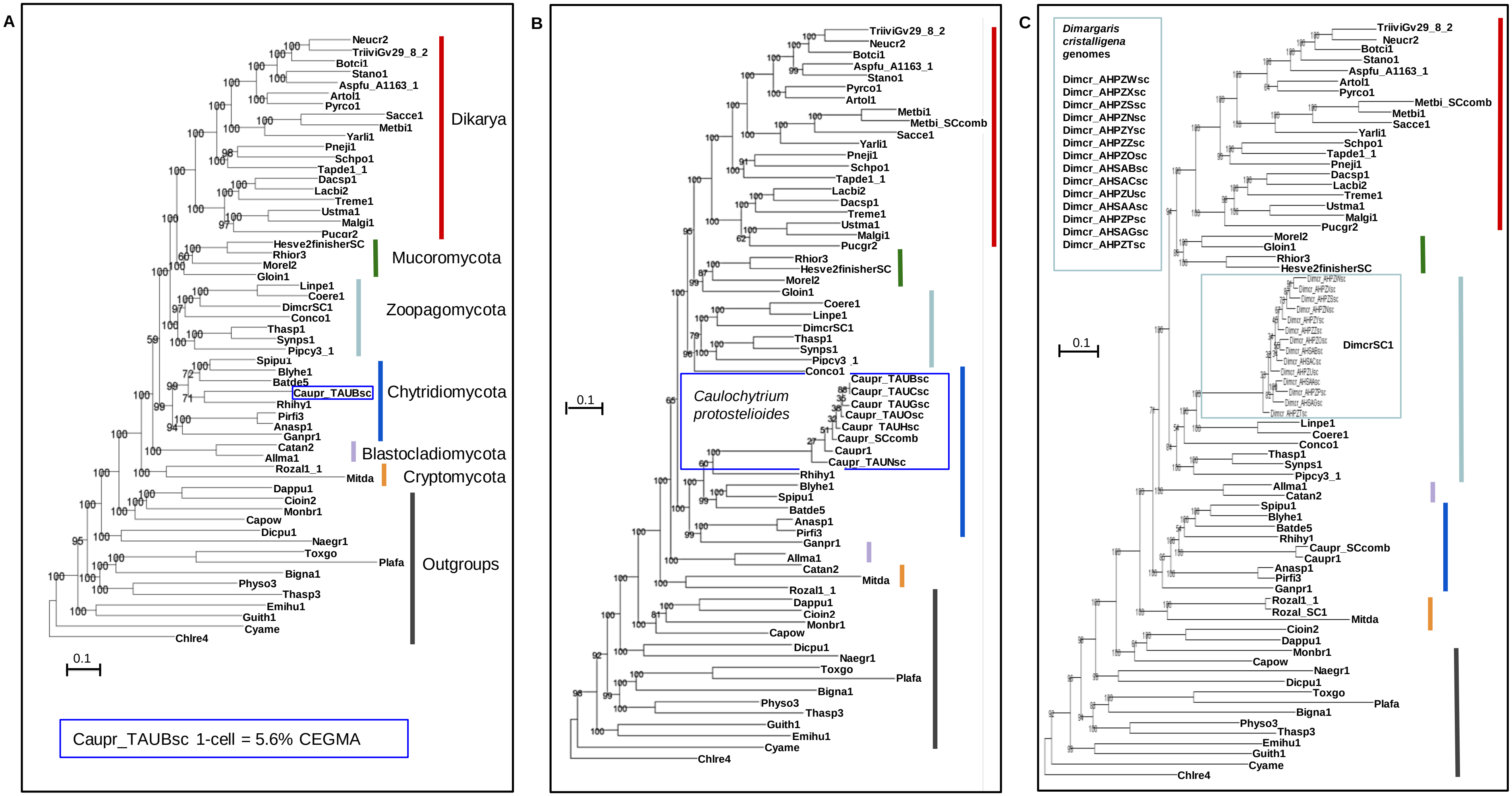


Figure S8. Phylogenomic placement of partial genomes. **Related to Figure 4.** RaxML trees with bootstrap values. Phyla names are on the right side of the color-coded vertical bars. **A.** *C. protostelioides* single-cell with lowest completeness (marked by sc) alone. **B.** *C. protostelioides* single- 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/ 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[1]	ITS4rev	TCCTCCGCTTATTGATATGC
Eukarya, Fungi	ITS	ITS[1]	ITS5for	GGAAGTAAAAGTCGTAACAAGG
Cryptomycota	18S	18SCRYPTO[2]	M13CRYPTO2-2F	GTTTTCCCAGTCACGACCACAGGGAGGTAGTGACAG
Cryptomycota	18S	18SCRYPTO[2]	M13AU4v2	CAGGAAACAGCTATGACGCCTCACTAAGCCATTC
Protists, Eukarya	18S	18SDPD[3]	M13DPD360FE	GTTTTCCCAGTCACGACCGGAGARGGMGCMTGAGA
Protists, Eukarya	18S	18SDPD[3]	M13DPD1492RE	CAGGAAACAGCTATGACACCTTGTTACGRCTT
Eukarya, Fungi	18S	18S_SR[1]	M13SR1RFor	GTTTTCCCAGTCACGACTACCTGGTTGATYCTGCCAGT
Eukarya, Fungi	18S	18S_SR[1]	M13NS4Rev	CAGGAAACAGCTATGACCTTCCGTCAATTCCTTTAAG

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 *Blyttiomyces 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 2.5x10^6 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. *Blyttiomyces 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-µm 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 diH2O; 3 g of CaCO3; 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 x 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 0.2uM 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_2O 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 MgCl2, Ambion, AM9530; 50 mM (NH4)2SO4, 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_2O, 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. subtillis* 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. allomycis* 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. allomycis* 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. allomycis* (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^{TM} (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 Exo-Sap 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 (Neubler) had higher accuracy relative to other tested bioinformatics tools. We further improved this algorithm and named it NeublerWA (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 bellow. 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/1RxgLpIaEzy0QJTnJO_nIbPWCuxGqu-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 bhist= 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 bbtoolsl; bbduk.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 samplereadstarget=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 (Figure5). 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., Prjibelski, 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 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., Otillar, 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.