Single Cell Genomics and Transcriptomic for Unicellular Eukaryotes

Doina Ciobanu[1*], Alicia Clum[1], Vasanth Singan[1], Asaf Salamov[1], James Han[1], Alex Copeland[1], Igor Grigoriev[1], Timothy James[2], Steven Singer[3], Tanja Woyke[1], Rex Malmstrom[1], and Jan-Fang Cheng[1]

[1]DOE Joint Genome Institute, Walnut Creek, California
[2]University of Michigan, Ann Arbor, Michigan
[3]DOE JointBioEnergy Institue, Emeryville, California
*Email Address: dgciobanu@lbl.gov

March 2014

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California.

Single Cell Genomics and Transcriptomics for Unicellular Eukaryotes

Doina Ciobanu[1]*(dgciobanu@lbl.gov), Alicia Clum[1], Vasanth Singan[1], Asaf Salamov[1], **James Han[1]**, Alex Copeland[1], Igor Grigoriev[1], Timothy James[2], Steven Singer[3], Tanja Woyke[1], Rex Malmstrom[1], and Jan-Fang Cheng[1]

[1]DOE Joint Genome Institute, Walnut Creek, California; [2]University of Michigan, AnnArbor, Michigan; [3]DOE Joint BioEnergy Institute, Emeryville, California.

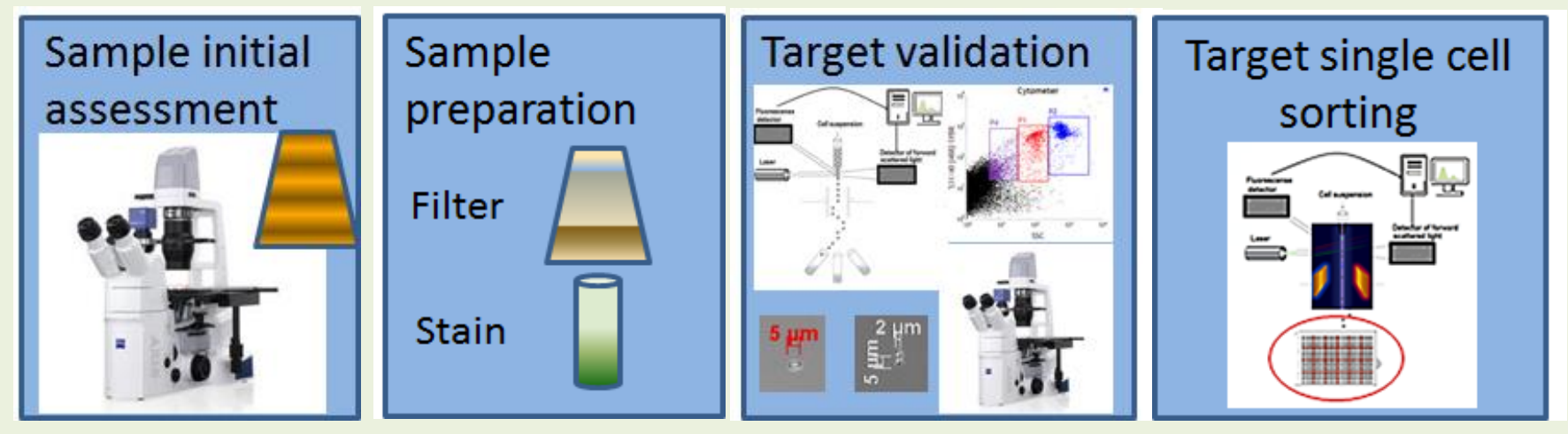
JGI JOINT GENOME INSTITUTE DEPARTMENT OF ENERGY

Introduction

Unicellular eukaryotes have complex genomes with a high degree of plasticity that allow them to adapt quickly to environmental changes. They live with prokaryotes and higher eukaryotes, frequently as symbionts or parasites. The vast majority of eukaryotic microorganisms are uncultured or unculturable, and thus not sequenced so far. To this day their contribution to the dynamics of the environmental communities remains to be understood. Here, we present four components of our approach to isolate, sequence and analyze eukaryotic microorganisms: target isolation and genome/transcriptome recovery for sequencing; sequence analysis for single cell genome and transcriptome, and genome annotation. We have tested some of our tools and some are being still tested, using six species: an uncharacterized protist from cellulose-enriched compost identified as *Platyophrya*, a close relative of *P. vorax*; the fungus *Metschnikowia bicuspidate,* a parasite of water flea *Daphnia*; the mycoparasitic fungi *Piptocephalis cylindrospora,* a parasite of *Cokeromyces* and *Mucor*; *Caulochytrium protosteloides*, a parasite of *Sordaria; Rozella allomycis*, a parasite of the water mold *Allomyces*; and the microalgae *Chlamydomonas reinhardtii*.

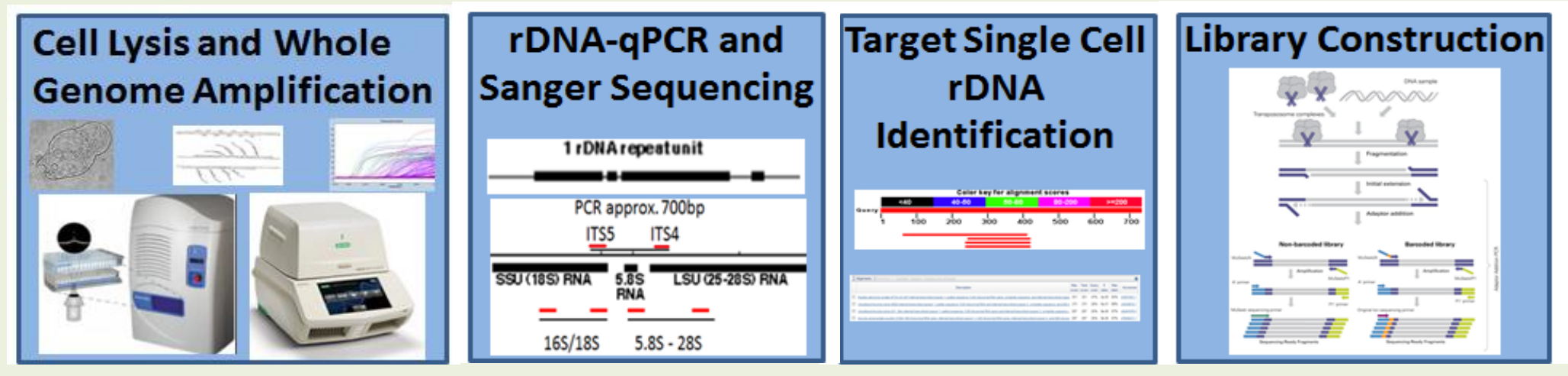
METHODS : LABORATORY PROCESS BEFORE SEQUENCING

Single Cell Isolation Critical Steps



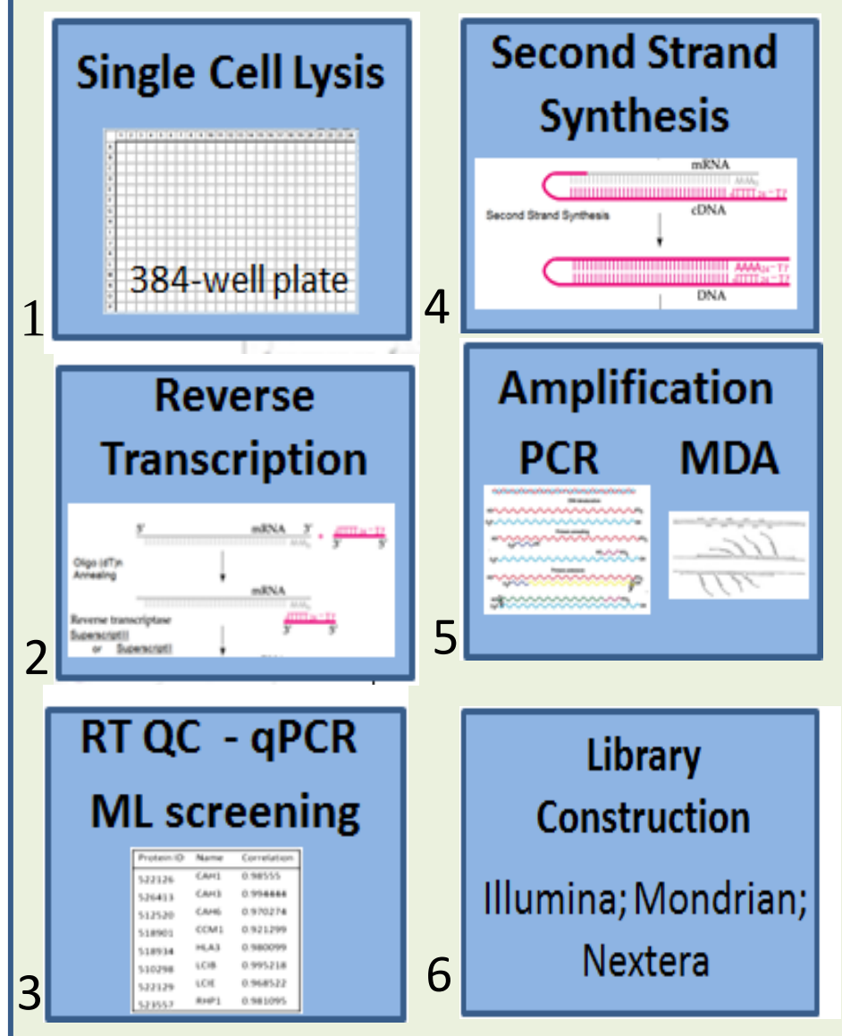
Sample initial assessment: Morphology and standard DNA stains, as well as various specific stains are used for identifying the target. among the heterogeneous content of the environmental samples. **Sample preparation:** Separation of different size populations is done by filtering and/ or pre-sorting, which is followed by **target validation** using the cell sorter and the microscope, to identify the correct population to be used for sorting into 384-well plates.

Single Cell Processing After Sorting for Genomics



Cell Lysis: first critical step for genome recovery of single cells. Several methods have been tested for efficient eukaryote single cells. **Whole genome amplification (WGA) i**s the next critical step . Several parameters are being tracked: **MDA "start" time** – likely to be reflective of cell lysis and DNA denaturation efficiency; possible reflective of the genome coverage; **MDA total time** – directly proportional with degree of amplification bias; **rDNA-qPCR:** We have tested several primer sets for eukaryotic rDNA region, for 18S, ITS and 28S subunits. Currently we are using 18S and ITS regions and NCBI database. **Library constructions:** we tested several different protocols for Illumina method.

Single Cell Transcriptomics Method Development Critical Steps

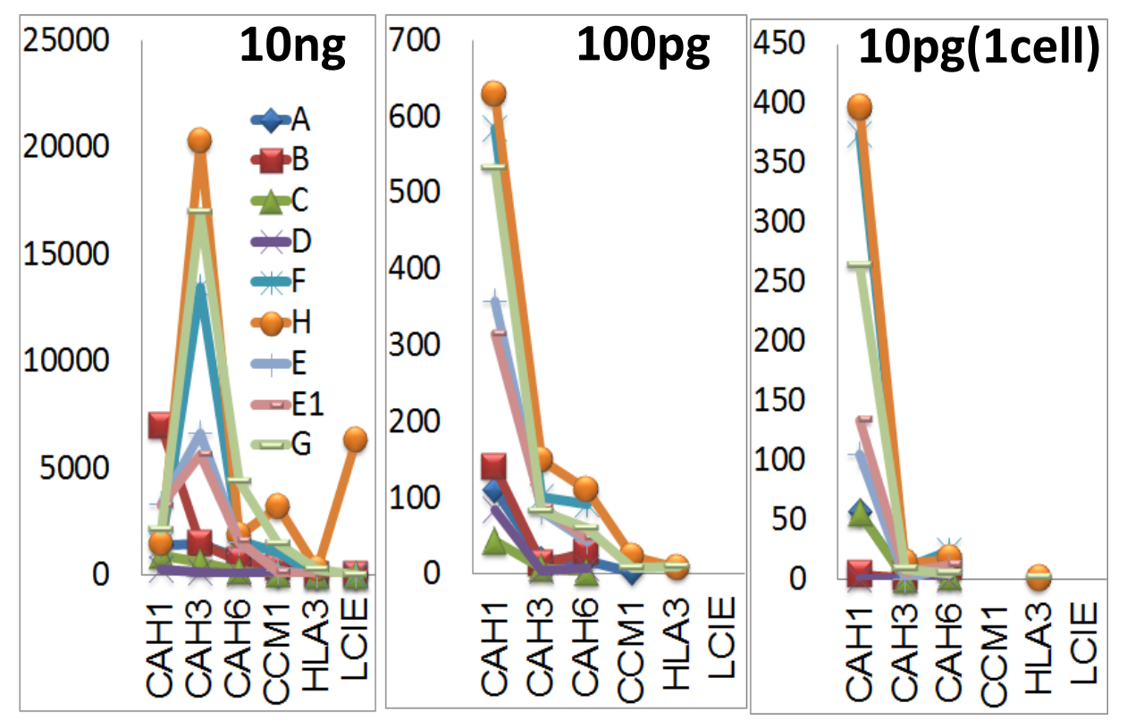


1. Several lysis methods has been tested for single cell transcriptomics, selection criteria were: compatibility with high-throughput format; compatibility with the downstream process and chemistry; transcriptome recovery; time; cost; purity of the reagents. Commercial kits, versus direct lysis and LiCl -based lysis were tested on single cells.
2. Eight Reverse Transcription methods were tested on purified total RNA in amounts equivalent to 1000; 100; 10 and 1 cells for single cell eukaryotes. Methods tested were: Superscript II (A); Superscript III (B); Thermoscript (C); PrimeScript with gDNA eraser (D), for all following manufacturer protocol; Superscript II and Superscript III with essential chemistry modifications (E; E1) and (F) respectively; SmartSeq2 (Nature Methods,Vol10 NO11:1096-1098 (G); SmartSeq2 modified protocol and components (H).
3. Reverse Transcription Quality Check was done using six C.reinhardtii 4a+ genes, shown to have a high correlation with tRNAseq transcriptome analysis (Cell,Vol.24:1876–1893,May 2012).
4. Second Strand Synthesis was performed differently for different RT methods. Efficiency was estimated in preliminary tests, not shown here.
5. Amplification of the cDNA was tested by T7-IVT, PCR or MDA. First method was dropped from further experiments due to much higher costs, however, it did show a higher efficiency than PCR or MDA.
6. For the library construction three methods are being tested: Illumina Fragment 500bp; Mondrian (Ovation SP+ for Ultra low input) and Nextera XP for low input.

RESULTS: TRANSCRIPTOME METHOD DEVELOPMENT

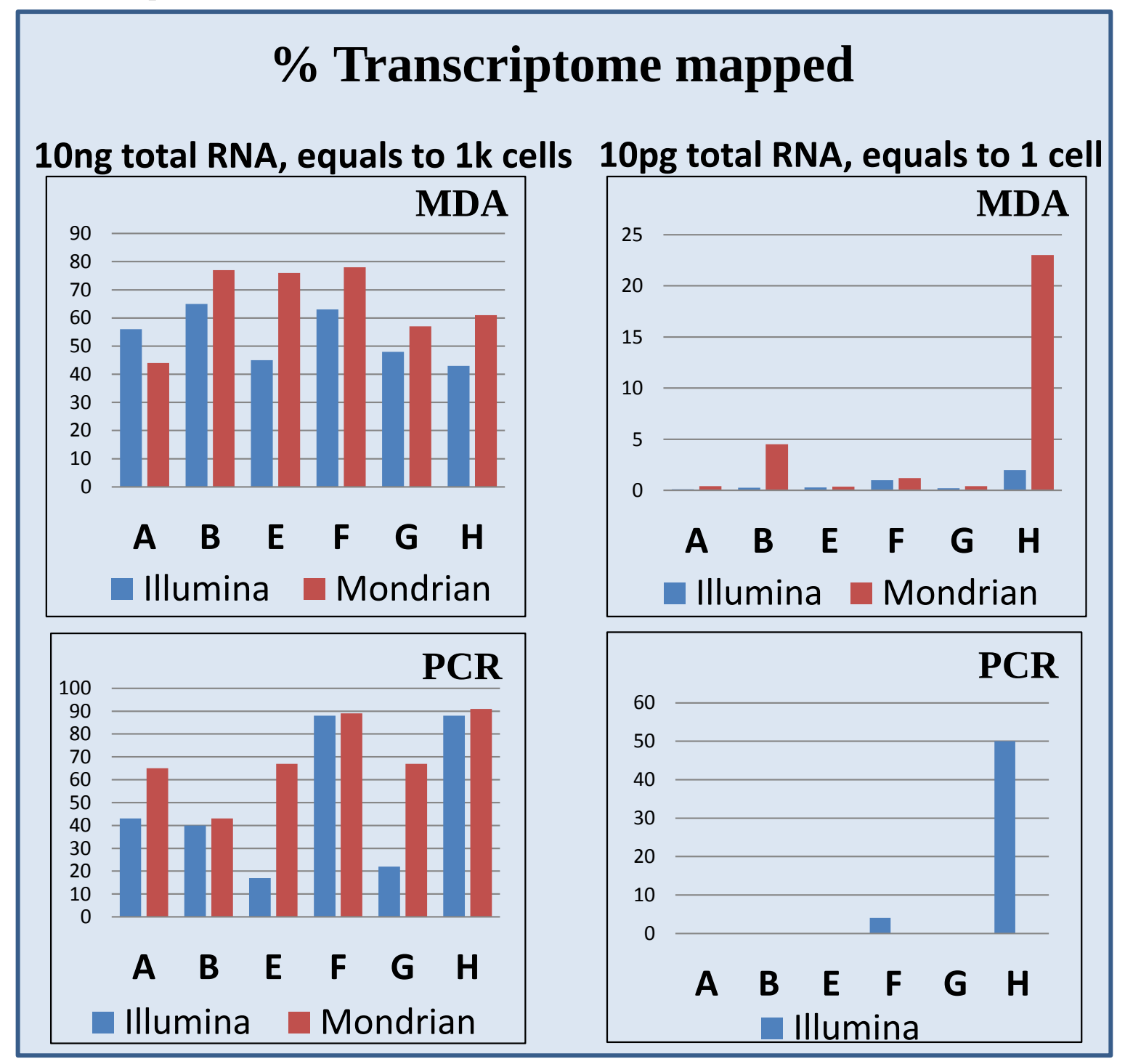
RT Method Selection using multi-locus screening

Input total RNA:

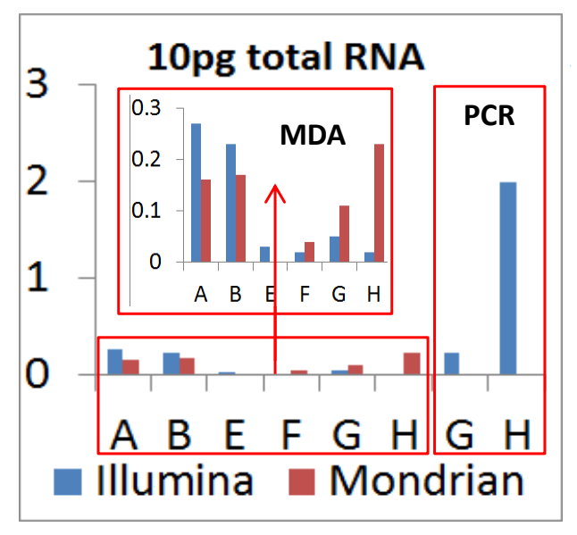


Shown above are relative expression levels for 6 genes for each of the RT methods. As a result of this analysis three methods were selected as most efficient: H,F,G

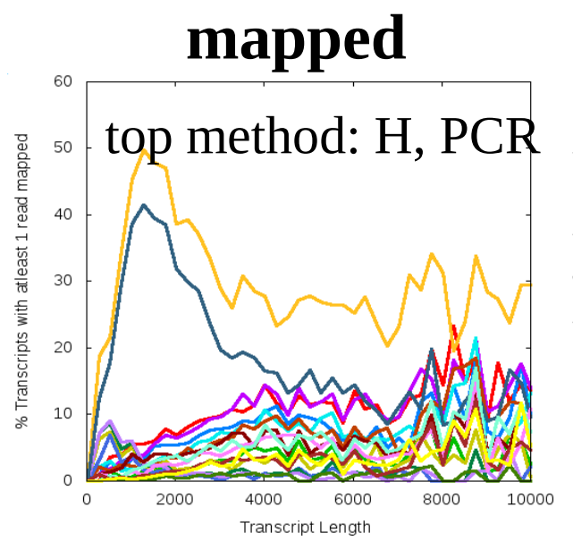
Comparison between methods: A,B,E,F,G,H



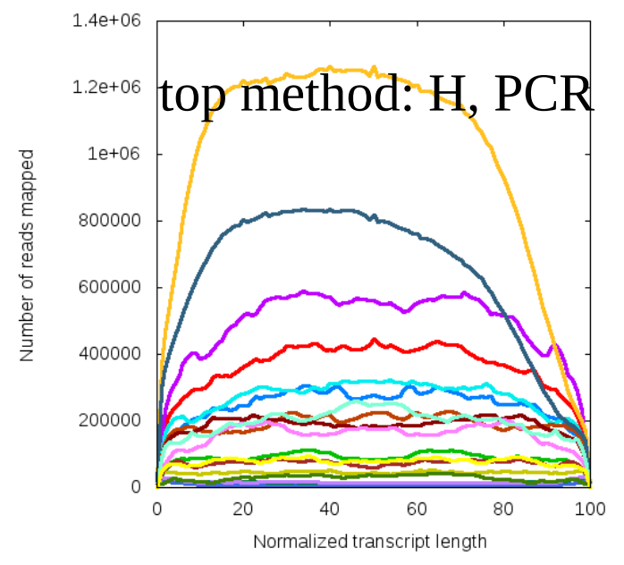
Transcriptome base coverage



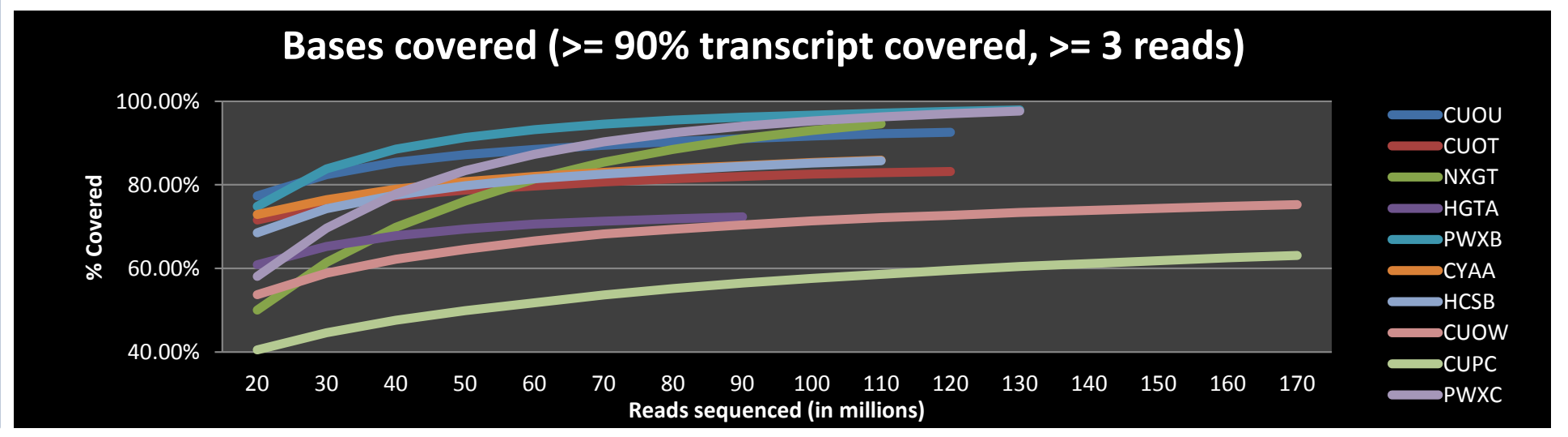
% Transcripts with at least 1 read mapped



Transcript distribution

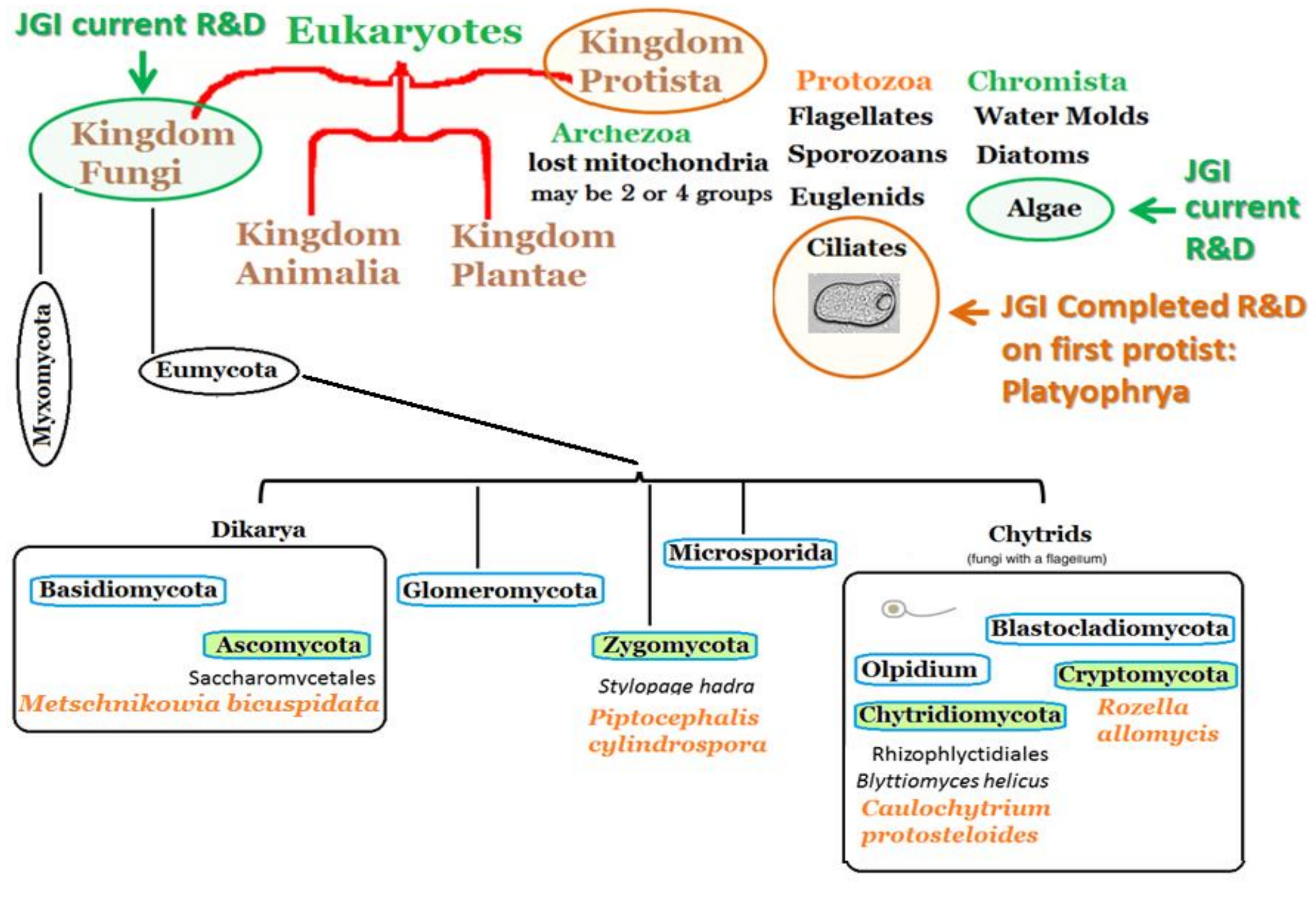


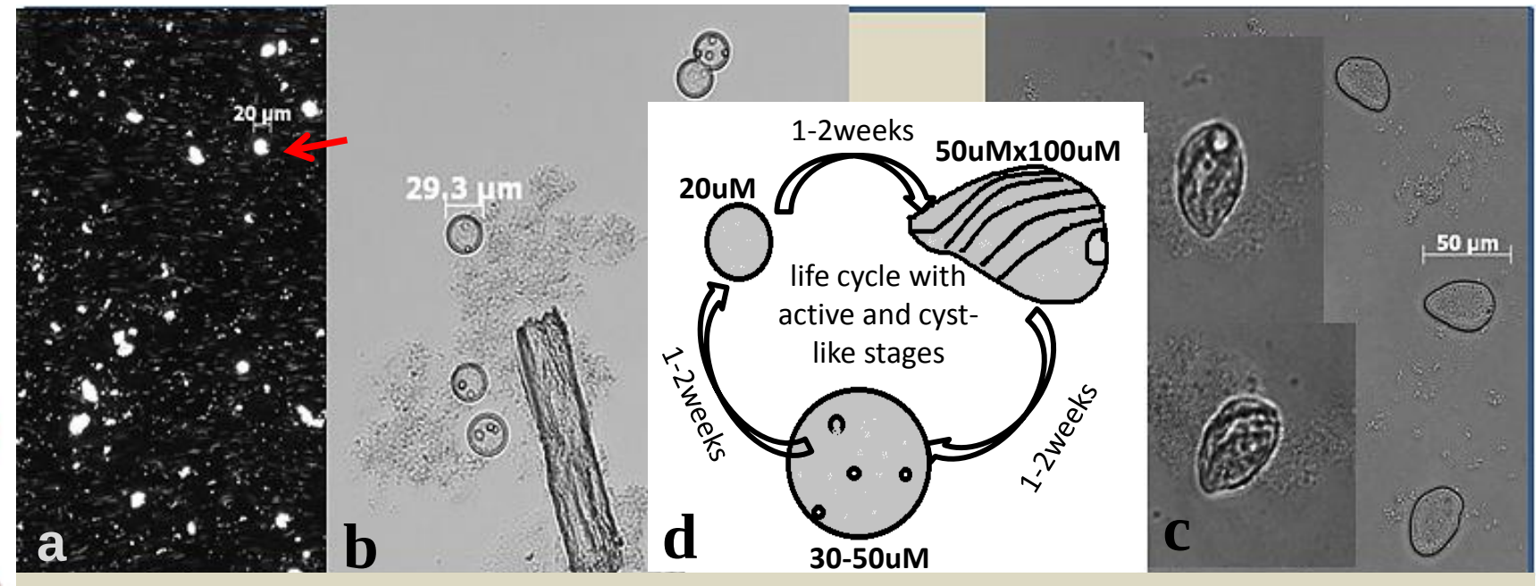
Rarefaction curve for maximum transcriptome coverage (different fungi libraries)



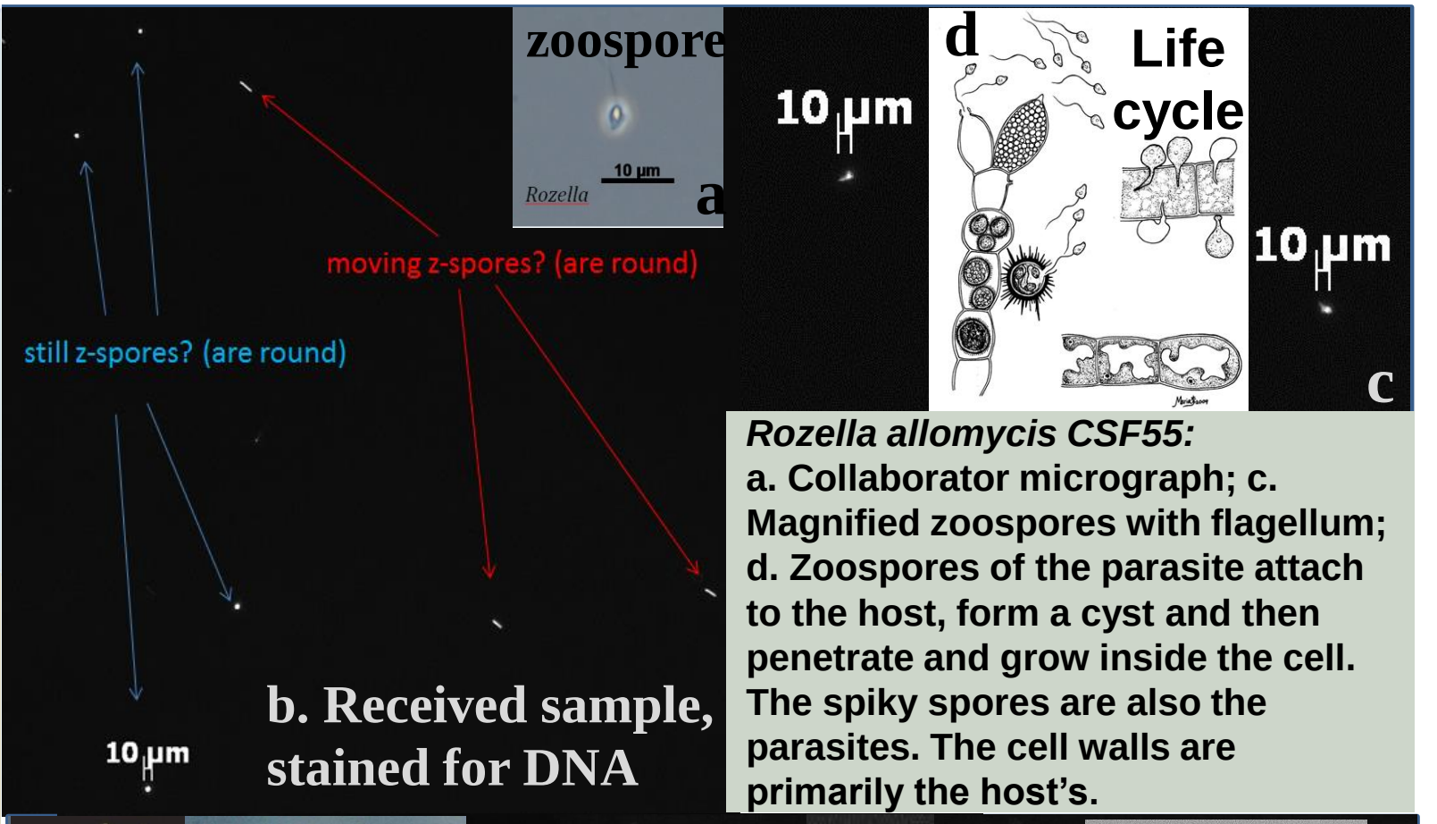
Samples

Single Cell Eukaryote Sequencing at JGI



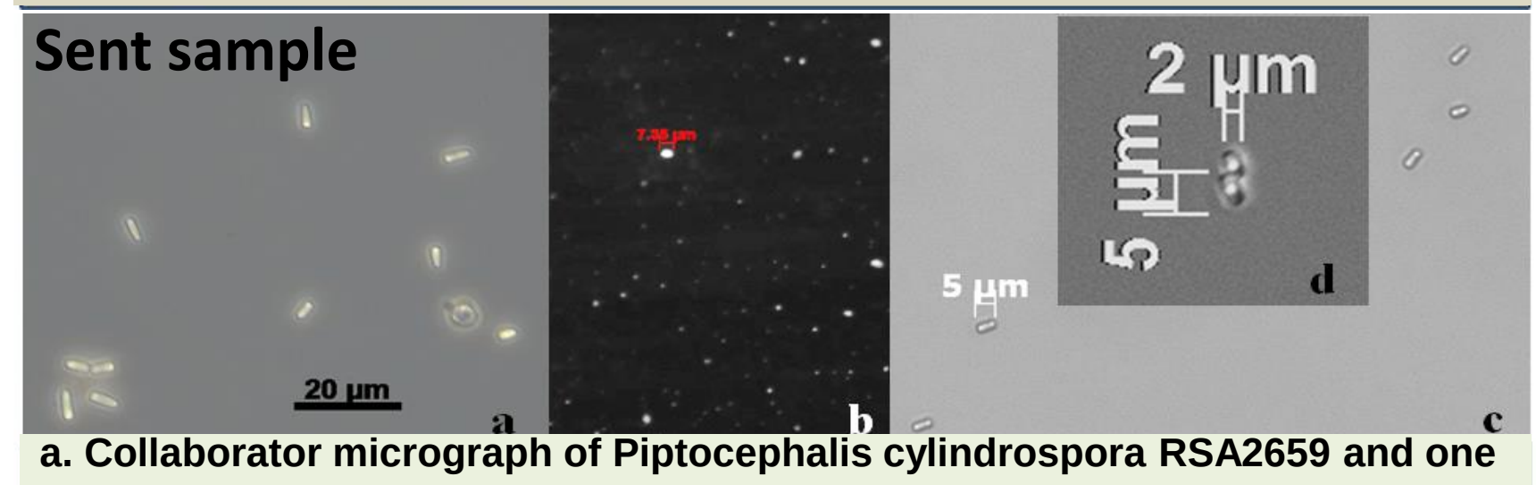


a. Compost sample from JBEI, enriched with microcrystalline cellulose, stained for DNA. b. Protist forming cysts- intermediate form; c. Protist active form, moving and feeding around or on the microcrystalline cellulose. d. Life cycle as observed at JGI.

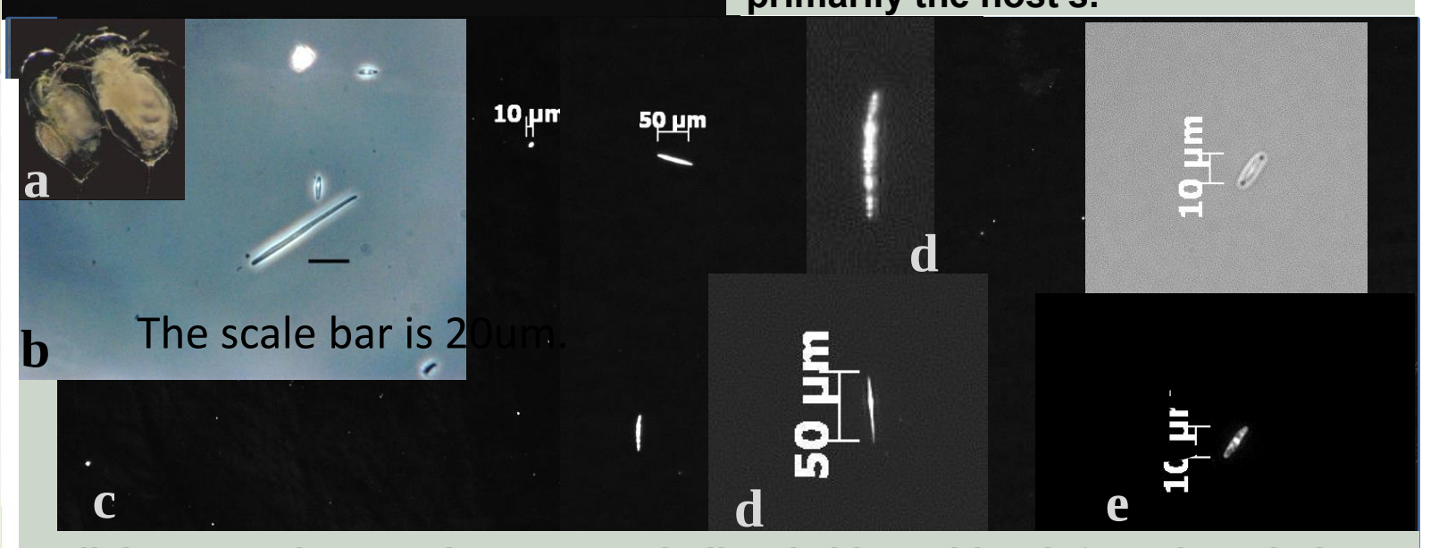

zoospore
Rozella
a
10 µm
still z-spores? (are round)
moving z-spores? (are round)
b. Received sample, stained for DNA
10 µm
d
Life cycle
10 µm
c

Rozella allomycis CSF55: a. Collaborator micrograph; c. Magnified zoospores with flagellum; d. Zoospores of the parasite attach to the host, form a cyst and then penetrate and grow inside the cell. The spiky spores are also the parasites. The cell walls are primarily the host's.

Sent sample



a. Collaborator micrograph of Piptocephalis cylindrospora RSA2659 and one host cell. b.Received sample, stained for DNA shows a heterogeneous composition of the target and other smaller cells. c. Bright field does not detect smaller cells. d. Overlap of BF and FL shows two spores of the parasite with bright nucleus.



Collaborator micrographs: a. *Metschnikowia biscupidata* infected *Daphnia* on right and uninfected on left. b. Ascospore and the yeast cell of the parasite. Received sample: c. Yeast (10um) and ascospore (50um) cells stained for DNA; d. Ascospore magnified; e. yeast cell BL and FL.

METHODS: SEQUENCE ANALYSIS TOOLS for SINGLE CELL

Genome Assembly

Co-Assembly Strategy Comparison for Compost Protist on Normalized Data

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

Co-Assembly Strategy Comparison for *Piptocephalis cylindrospora*

assembler	number of contigs	contig N50	assembled genome size
metagenome pipeline	5987	3.0 KB	9 MB
SPAdes	6102	7.3 KB	10.9 MB

Several assembly strategies were tested using normalized and raw data for single cells and co-assemblies. The current assembly strategy for these projects is to use SPAdes without normalization. This is the same approach that is used now on microbial single cell projects at JGI.

Annotation

Protist Analysis: Annotation pipeline was run on 47675 scaffolds with length > 500bp. For gene prediction we used ab initio method - fgenesh, with parameters specifically trained for ciliates, as well as protein-homology based methods, like genewise and fgenesh++, using alternative genetic code 6.

Fungal Analysis: For P.cylindrospora was used JGI eukaryotic annotation pipeline on a combined assembly of 3 single cells.

Transcriptome Analysis

Preprocessing: Read1 from the fastq files was extracted and all statistics were calculated from only read1 data. Reads were trimmed for the primer sequences followed by Illumina artifacts**.**

% transcriptome mapped: Reads were mapped to the reference transcriptome. Number of reads that mapped to the transcriptome was represented as a percentage of total number of reads generated.

% Transcriptome covered: Reads were mapped to the reference transcriptome. Absolute number of bases in the transcriptome covered by reads was extracted and represented as a percentage of the entire transcriptome length.

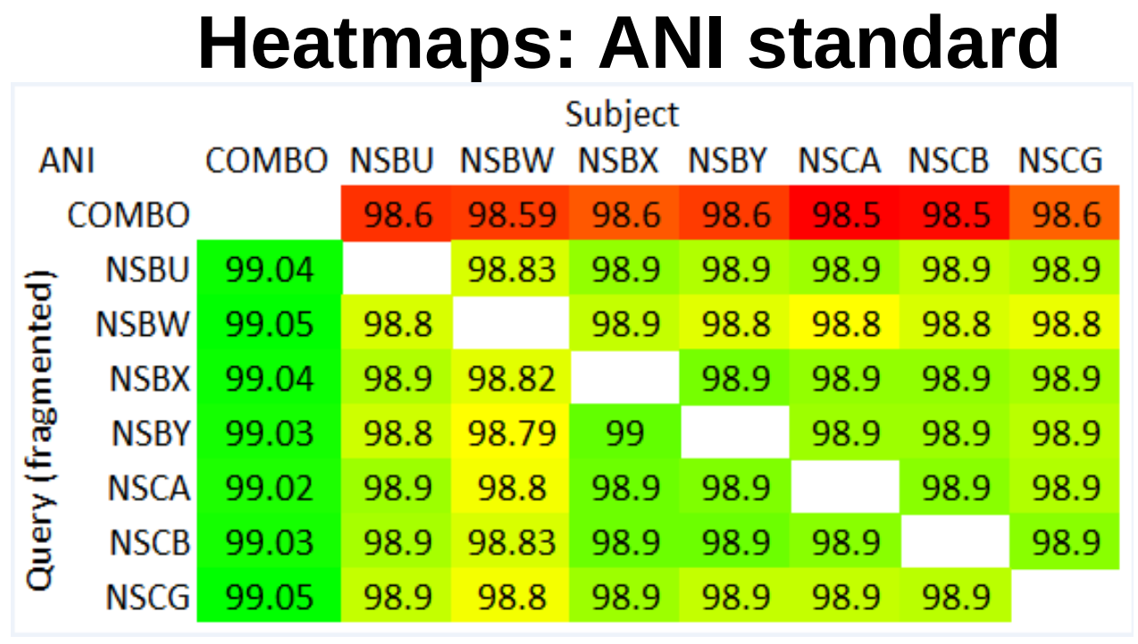
Transcript distribution plot: For each transcript, the number of reads mapping at every base position was calculated. This number was averaged across all the transcripts after normalizing the transcripts to a length of 100 bases. This plot shows if the reads were evenly distributed across the entire length of the transcript.

% transcripts with at least 1 read mapped: Transcripts were binned based on their lengths. For each bin, numbers of reads mapped to the transcripts were calculated. Percentage of transcripts within the bin having at least 1 read is calculated and plotted. This plot shows how many transcripts at a given length had at least 1 read mapped to it.

RESULTS: GENOME ANALYSIS

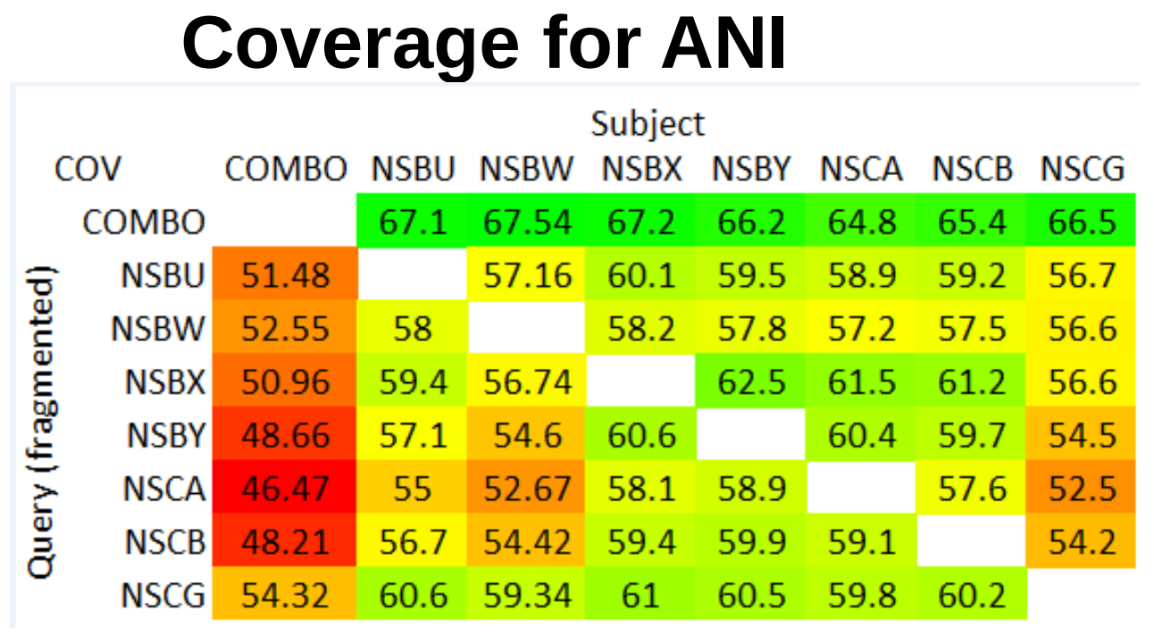
Protist rDNA (18S) 1753bp HiSeq sequence has 99% Identity with Platyophrya vorax

Heatmaps: ANI standard



ANI	COMBO	NSBU	NSBW	NSBX	NSBY	NSCA	NSCB	NSCG
COMBO		98.6	98.59	98.6	98.6	98.5	98.5	98.6
NSBU	99.04		98.83	98.9	98.9	98.9	98.9	98.9
NSBW	99.05	98.8		98.9	98.8	98.8	98.8	98.8
NSBX	99.04	98.9	98.82		98.9	98.9	98.9	98.9
NSBY	99.03	98.8	98.79	99		98.9	98.9	98.9
NSCA	99.02	98.9	98.8	98.9	98.9		98.9	98.9
NSCB	99.03	98.9	98.83	98.9	98.9	98.9		98.9
NSCG	99.05	98.9	98.8	98.9	98.9	98.9	98.9	

Coverage for ANI



COV	COMBO	NSBU	NSBW	NSBX	NSBY	NSCA	NSCB	NSCG
COMBO		67.1	67.54	67.2	66.2	64.8	65.4	66.5
NSBU	51.48		57.16	60.1	59.5	58.9	59.2	56.7
NSBW	52.55	58		58.2	57.8	57.2	57.5	56.6
NSBX	50.96	59.4	56.74		62.5	61.5	61.2	56.6
NSBY	48.66	57.1	54.6	60.6		60.4	59.7	54.5
NSCA	46.47	55	52.67	58.1	58.9		57.6	52.5
NSCB	48.21	56.7	54.42	59.4	59.9	59.1		54.2
NSCG	54.32	60.6	59.34	61	60.5	59.8	60.2	

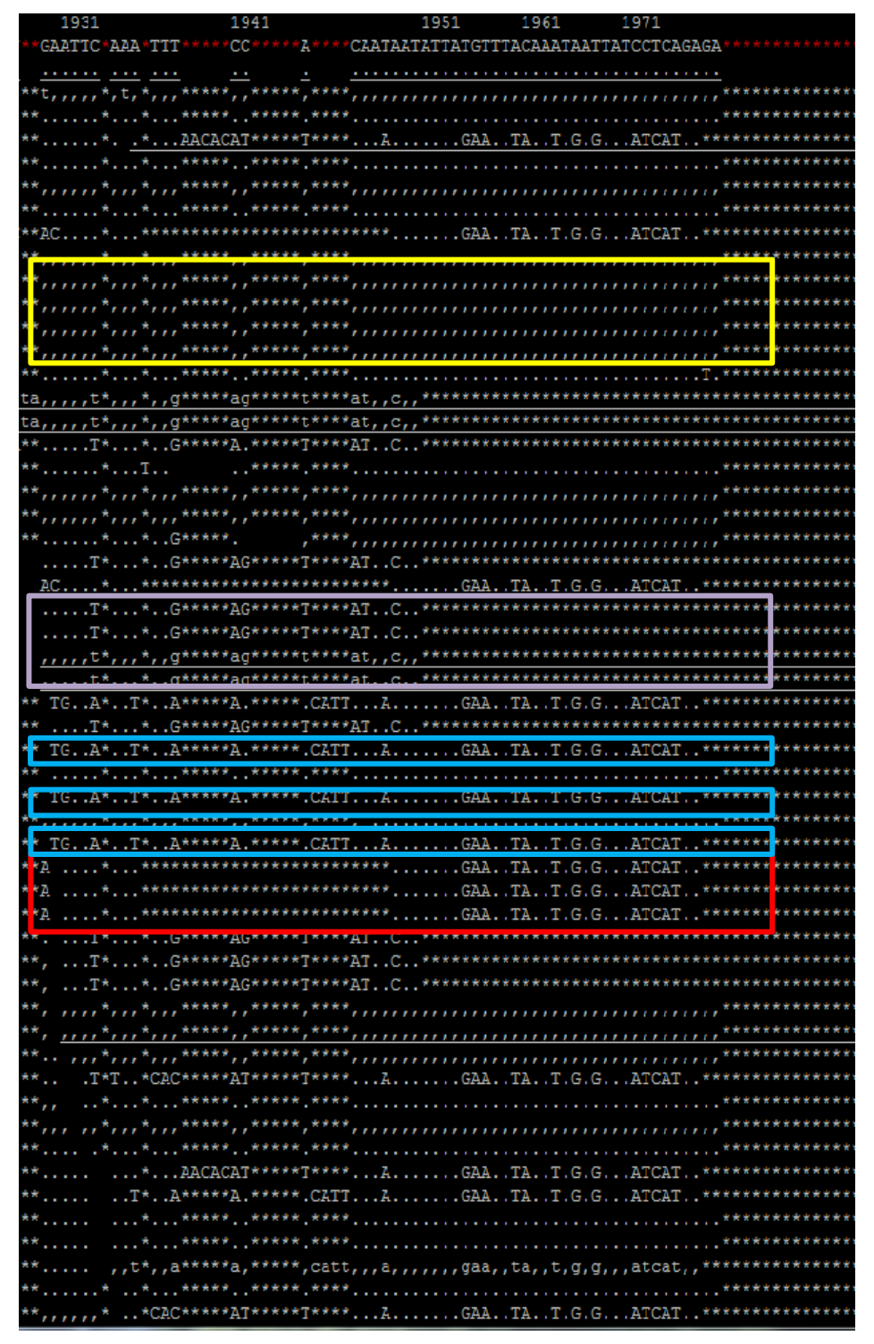
Protist: ANI stands for average nucleotide identity. The coverage heatmap shows the percentage of the genomes that were used for the ANI calculation, i.e. had hits above the cutoff (>70% identity over >70% of the fragment, fragment size was 1020 bp).

Fungal Single Cell Assembled Genomes

Organism	GC%	20mer uniqueness at 1mln reads 100cells	20mer uniqueness at 1mln reads 1cell	Assembled Genome Size MB
Piptocephalis cylindrospora RSA2659	51	NA	10-20%	4.9 (1 cell)
Rozella allomycis CSF55	35	90%	40%	20 (100cell); 7 (1 cell)
Caulochytrium protosteloides	60-70	30%	5%-10%	13 (100cell); 1 (1cell)
Metschnikowia bicuspidata, yeast	50	80%	60%	In progress

Rozella allomycis polymorphism

At least 4 different strains



Annotation

Protist Analysis: Preliminary analysis based on PFAM domains, predicted on all possible potential ORFs, indicated that most of the scaffolds are from some unknown ciliate, which uses alternative genetic code, where TAA and TAG codons code for glutamine Q (translation table 6). Pipeline predicted 40,072 gene models, with ~65% of models having homology to KEGG database proteins and ~61% to Swissprot proteins. ~45% of genes have at least one Pfam domain and ~56% are complete (from start codon to stop codon). Closest species with sequenced genomes to this protist are ciliates Paramecium tetraurelia and Tetrahymena thermophila, whith whom it shares 4839 and 4765 orthologs respectively (~44-45% percent identity on amino acid level), based on bidirectional BLAST hits. Completeness of genome based on CEGMA analysis of core eukaryotic genes was estimated at 94.3%. **Fungal Analysis:** *Piptocephalis cylindrospora RSA2659* assembly filtered to 8.2 Mb in 1000 contigs indicates 3300 genes with median length of 1074. (median: exon length 216bp; intron 82bp, transcript length of 924bp and 2050 spliced genes. Gene density of 403.02 Mbp. Based on CEGMA analysis of core genes, completeness of genome is estimated at 75.5%

CONCLUSIONS

- Several modifications to the existing pipeline for single cell (prokaryote) were tested in order to obtain quality data for single cell eukaryotes.
- Tested modifications affect following major parts of the pipeline: Single Cell Isolation Steps; Single Cell Genome Recovery; Genome Assembly and Annotation. Implemented modifications show good results.
- One of the bottlenecks in single cell eukaryote analysis is the scarcity of rDNA data in the form of curated databases, this area needs further development.
- A new capability for unicellular eukaryotes has been under development and preliminary results indicate that single cell eukaryote transcriptomics could be used as a complementing step for the single cell eukaryote pipeline. One best method has been determined and together with few other methods are currently being tested on single cells for their performance consistency .

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

ACKNOWLEDGEMENTS:

QC; Sequencing; and RQC groups at JGI; Library group for providing assistance and supplies; Patrick Schwientek for providing assistance with data analysis and generating the heatmaps for the protist.