2-2020

De novo Sequencing, Assembly, and Annotation of the Transcriptome for the Free-Living Testate Amoeba *Arcella intermedia*

Giulia M. Ribeiro
University of Sao Paulo, Brazil

Alfredo L. Porfírio-Sousa
University of Sao Paulo, Brazil

Xyrus X. Maurer-Alcalá
Smith College

Laura A. Katz
Smith College, lkatz@smith.edu

Daniel J.G. Lahr
Smith College

Recommended Citation
Ribeiro, Giulia M.; Porfírio-Sousa, Alfredo L.; Maurer-Alcalá, Xyrus X.; Katz, Laura A.; and Lahr, Daniel J.G., "De novo Sequencing, Assembly, and Annotation of the Transcriptome for the Free-Living Testate Amoeba *Arcella intermedia*" (2020). Biological Sciences: Faculty Publications, Smith College, Northampton, MA.
https://scholarworks.smith.edu/bio_facpubs/134

The Journal of Eukaryotic Microbiology

ISOP Published by the International Society of Protistologists

Journal of Eukaryotic Microbiology ISSN 1066-5234

ORIGINAL ARTICLE

De novo Sequencing, Assembly, and Annotation of the Transcriptome for the Free-Living Testate Amoeba *Arcella intermedia*

Giulia M. Ribeiro[a], Alfredo L. Porfírio-Sousa[a], Xyrus X. Maurer-Alcalá[b,c,1], Laura A. Katz[b] & Daniel J.G. Lahr[a]

a Department of Zoology, Institute of Biosciences, University of São Paulo, Matao Street, Travessa 14 Cidade Universitaria, São Paulo 05508-090, São Paulo, Brazil

b Department of Biological Sciences, Smith College, 10 Elm Street, Northampton, Massachusetts 01063

c Program in Organismic and Evolutionary Biology, University of Massachussetts Amherst, 230 Stockbridge Road, Amherst, Massachusetts 01002-9316

Keywords

Gene discovery; next generation sequencing; pathways description.

Correspondence

D.J.G. Lahr, Rua do Matão, Travessa 14, 101, Cidade, São Paulo, SP 05588-090, Brazil

Telephone number: +55-11-3091-0948;

e-mail: dlahr@ib.usp.br

Received: 1 October 2019; revised 19 November 2019; accepted December 22, 2019.

Early View publication February 12, 2020

doi:10.1111/jeu.12788

ABSTRACT

Arcella, a diverse understudied genus of testate amoebae is a member of Tubulinea in Amoebozoa group. Transcriptomes are a powerful tool for characterization of these organisms as they are an efficient way of characterizing the protein-coding potential of the genome. In this work, we employed both single-cell and clonal populations transcriptomics to create a reference transcriptome for *Arcella*. We compared our results with annotations of *Dictyostelium discoideum*, a model Amoebozoan. We assembled a pool of 38 *Arcella intermedia* transcriptomes, which after filtering are composed of a total of 14,712 translated proteins. There are GO categories enriched in *Arcella* including mainly intracellular signal transduction pathways; we also used KEGG to annotate 11,546 contigs, which also have similar distribution to *Dictyostelium*. A large portion of data is still impossible to assign to a gene family, probably due to a combination of lineage-specific genes, incomplete sequences in the transcriptome and rapidly evolved genes. Some absences in pathways could also be related to low expression of these genes. We provide a reference database for *Arcella,* and we highlight the emergence of the need for further gene discovery in *Arcella.*

ARCELLA is an abundant and diverse genus of testate amoebae, present in freshwater environments all around the globe (Meisterfeld 2002; Mitchell et al. 2008; Ogden and Hedley 1980). *Arcella* belongs to the major amoeboid lineage Tubulinea in the Amoebozoa, along with the well-known *Amoeba proteus*. Being one of the most diverse testate amoeba genus, a great deal of work has focused on their taxonomy, systematics and ecology (Gomaa et al. 2012). Their distinctive characteristic is a completely organic secreted shell, composed of hexagonal units (alveoli) and generally dome-shaped (Andrey and Yuri 2006) (Fig. 1). They are also capable of thriving in stressful environments like contaminated, eutrophic, and even relatively saline waters (Escobar et al. 2008; Patterson and Kumar 2000; Reinhardt et al. 1998; Roe and Patterson 2014; Roe et al. 2010).

Understanding *Arcella*'s metabolic functioning can bring new insights on their physiology, ecology, and evolution. However, the lack of genomic sequence data of arcellinids, as well as the majority of Amoebozoa (the key exceptions being slime-molds such as *Dictyostelium* (Eichinger et al. 2005) and *Physarum polycephalum* (Schaap et al. 2015), pathogenic and nonpathogenic *Entamoeba* species (Loftus et al. 2005), and *Acanthamoeba* (Clarke et al. 2013)), represents a substantial hurdle common to many protist groups. RNA-seq experiments are an attractive alternative that allow the identification of genes through sequencing of active transcripts in the cell and, in the absence of an annotated genome, are powerful for characterization of nonmodel organisms (Eldem et al. 2017).

Here, we present a compiled set of *Arcella* transcripts assembled in an attempt to create a reference transcriptome. We analyzed single cells at different stages in order to increase the diversity of metabolic representation in the

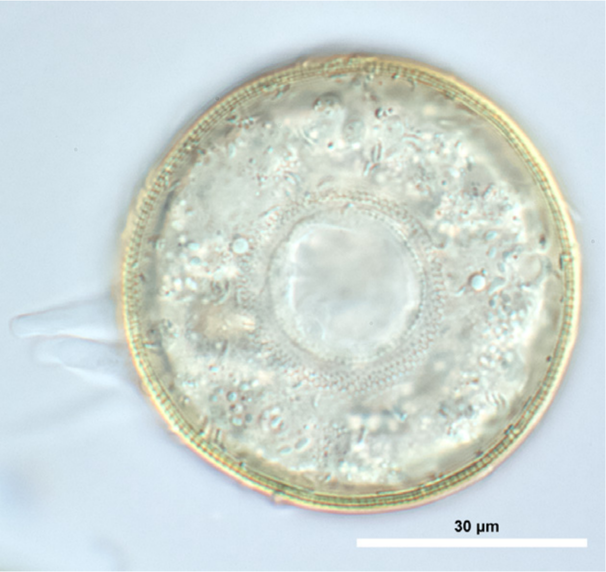

Figure 1 Organism of the study—*Arcella intermedia* LEP isolate 6, Magnification—630×.

dataset. We analyze the data using three annotation strategies (KEGG, EggNOG, and Blast2GO), concentrating only on the known metabolic pathways for eukaryotes. We compared our results with annotations of *Dictyostelium discoideum*, a model organism in Amoebozoa.

MATERIALS AND METHODS

RNA-seq experiment

We isolated an *Arcella intermedia* from a small artificial pool in University of São Paulo in 2013 (Coordinates—23.565720, −46.730512) (Fig. 1). We cultured in 12 ml of filtered and autoclaved water from their original environment in flasks with added carbon source of cereal grass media (Fisher Scientific S25242). We grew cultures at 24 °C, in the dark and until saturation for whole-culture RNA extraction.

Whole-culture mRNA extraction

We suspended cells, filtered in a mesh of 5-µm size, and collected in a sterile tube. We then centrifuged cells, collected the pellet, and subjected it to cell lysis for RNA isolation using TRIzol (Thermo Fisher Scientific, Waltham, MA, EUA, catalog 15596026) according to manufacturer's protocol. We accessed quality by Qubit (Thermo Fisher Scientific) and Bioanalyzer (Agilent Technologies, Santa Clara, CA). Once total RNA qualities and quantities were checked, the material was stored at −80 °C in 50-µl nuclease-free water (Life Technologies, Carlsbad, CA), to prevent total RNA degradation. We constructed libraries using TruSeq (Illumina, San Diego, CA) kit following manufacturer's instructions. The libraries were sequenced on a NextSeq Illumina platform at the CEFAP at Medical Biosciences Institute of University of São Paulo.

Single-cell mRNA extraction

With a single-cell RNA isolation protocol (Picelli et al. 2014), we selected three individuals from each of three different cultures, at four different time steps, totaling 36 individual samples. The time steps correspond to the traditional phases of growth (lag, log (exponential), stationary, and decline). Each cell was washed three times in sterilized water to dilute bacterial carryover. We then constructed libraries using NexteraXT kit (Illumina, EUA), following the manufacturer's instructions, which were sequenced on a NextSeq Illumina platform, at CEFAP facility in Medical Biosciences Institute of University of São Paulo. The 36 transcriptomes for single cells and the two whole-culture transcriptomes are deposited in SRA under BioProject PRJNA515423.

Transcriptome processing

We examined the quality of the reads using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). And we first processed the reads to remove adapter sequences using BBtools (http://sourceforge.net/projects/bbmap), then with a minimum quality score of 28 and minimum length of 75-bp. We used rnaSPAdes to pool and assemble all 38 transcriptomes (http://bioinf.spbau.ru/en/spades_3_9). The assembled transcriptome was rendered in a series of custom python scripts (available in http://github.com/maurerax/KatzLab/tree/HTS-Processing-PhyloGenPipeline) (Cerón-Romero et al. 2019; Maurer-Alcalá et al. 2018a,b). These script-processing steps included filtering of sequences by size (200-bp), and then, we performed a BLAST search against GenBank databases for removal of rDNA and bacterial transcripts, sequence identification (orthologous group) using Usearch and OrthoMCL database and after all, translating and removing partial sequences (Fig. S1). From these scripts, we were able to collect information about the size of the contigs, identity of the ortholog species found by Usearch and possible carryover of rDNA sequencing. Additionally, to avoid contamination in the final transcriptome dataset, we built a database with genomes for the species we identified rDNA. And then, we performed a BLAST search of this database on our transcriptome. We removed strings with an e-value of 1 e^{-100} and identity > 60%. After all the filtrations, we analyzed the completeness of the transcriptome by BUSCO analysis using an eukaryotic database (Simão et al. 2015; Waterhouse et al. 2017). This resulting set of transcripts was considered the reference transcriptome for *A. intermedia* and is used for subsequent analyzes.

The translated *A. intermedia* ORFs were searched against three different databases: EggNOG (http://eggnogdb.embl.de/#/app/home), Blast2GO (Conesa and Götz 2008) and BlastKOALA (http://www.kegg.jp/blastkoala/). EggNOG database assigns proteins of the transcriptome to the respective ortholog group based on best hits to their database and alignment based search (Jensen et al. 2007). EggNOG-mapper tool was used with HMMER (hmmer.org/) mapping mode and eukaryotes in taxonomic scope. Blast2GO (https://www.blast2go.com/) uses the

Table 1. Summary of the results after each of the filtering and processing steps

Arcella intermedia transcriptome database stats	
Number of reads	330,000,000
Number of reads after trimming	209,000,000
Assembled contigs	48,511
Number of contigs after size filtering	40,028
Number of contigs after rDNA filtering	39,903
Number of contigs after Bacteria + Archaea filtering	37,165
Number of mapped OGs	14,580
Number of annotated transcripts (EggNOG, KEGG, Blast2GO)	11,769
Number of annotated transcripts (EggNOG)	11,546
Number of transcripts with Gene onthology (EggNOG)	7,427
Number of annotated transcripts (Koala KEGG)	6,966
Number of annotated transcripts (Blast 2 GO)	11,033

BLAST algorithm to identify similar sequences (Conesa and Götz 2008). Blast2GO searches were made with default parameters. We characterized gene functions using BlastKOALA (https://www.kegg.jp/blastkoala/) using default settings for eukaryotes. We integrated the results for the three annotation strategies in order to have more complete results. To have a better basis for *Arcella* inferences, we compare the results of our translated transcriptome with the translated *D. discoideum* AX4 genome data (*GCF*_000004695.1). We passed this database translated by the same annotation procedure applied to *Arcella*.

RESULTS AND DISCUSSION

Arcella transcriptome sequencing and processing

We obtained and characterized a pooled set of *A. intermedia* transcriptome data (Table S1). The Illumina NextSeq sequencing platform generated 330 millions of reads; and after trimming, a total of 209 millions of clean reads remained (Table 1), which were assembled into 48,511 contigs with an average length of 600-bp (Table S1). The number of assembled contigs was refined after each of the processing steps (Table 1 and Fig. 2a). After filtering by size criteria, 40,028 contigs remained, and after rDNA removal, 39,903 contigs remained. In addition, the main component of the rDNA present was bacterial (Fig. S3). We removed the prokaryotic rDNA carryover leaving 37,165 contigs. The eukaryotic carryover was insignificant in the dataset, as it only included about 200 possible contigs, most of which are short or with uncertain identity. Because they did not impact downstream analyses, they were all removed. A total of 22,453 contigs remained without a orthologous identification, opening space for gene discovery. Postfiltering, we analyzed the transcriptome completeness using BUSCO against translated ORFs at the transcriptome and found 267 of 303 (88.1%)

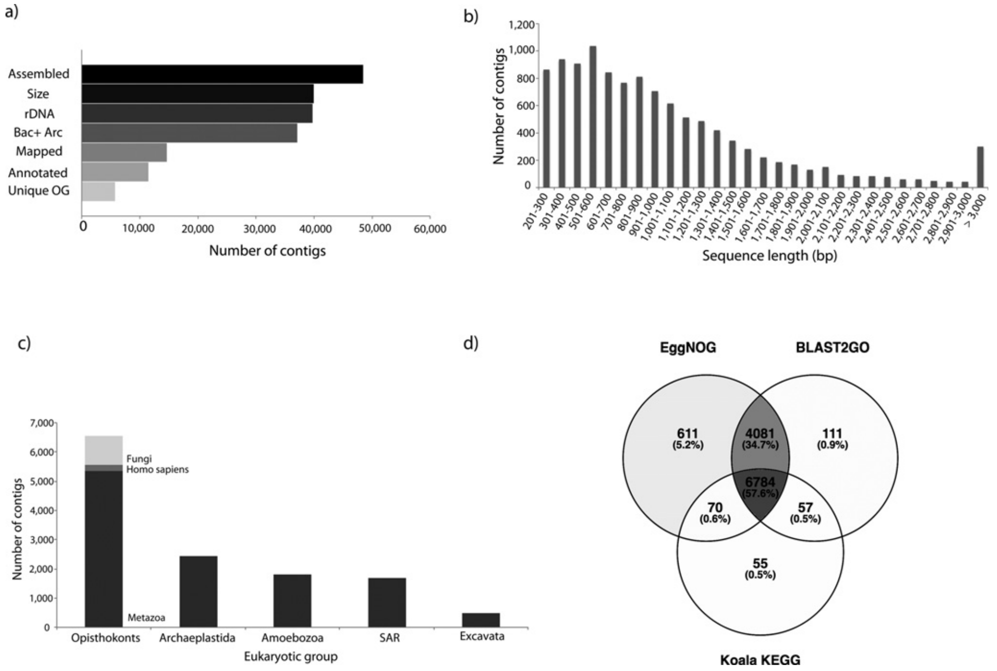

Figure 2 Summary of the resulting contigs after all processing steps. (**a**) Number of contigs remaining after each transcriptome processing step. (**b**) Length distribution of the contigs. The average length is 600-bp, and around 10% of mapped contigs have more than 2,000-bp. (**c**) Species hits distribution of mapped contigs in the main eukaryotic supergroups. Horizontal axis represents taxonomic supergroups, and color gradients inside Opisthokonta group represent relevant inner species and groups. (**d**) Venn diagram with portion of annotated contigs in three different strategies. About 57.6% of transcripts were mapped by all three methods.

core eukaryotic genes represented. Altogether, the high number of core eukaryotic genes and transcripts found indicates the relative completeness of our *A. intermedia* transcriptome.

After orthologous identification using Usearch and OrthoMCL and translation of the contigs, a total of 14,580 ORFs were identified, comprising 5,798 unique homologous groups. Most contigs are common between clonal populations and single-cell data. In addition, single-cell transcriptomes were responsible for adding 200 exclusive genes in the final dataset. The lengths of the smallest to largest ORFs ranged from around 200- to 10,000-bp (Fig. 2b). As *A. intermedia* is a member of an understudied clade, many components of the transcriptome lack homologs in other Amoebozoa and instead have a substantial number of identities into Opisthokonta (including animals and fungi) and plants (Fig. 2c). This is likely the result of uneven distribution of species available in the databases that include only a few species of Amoebozoa with completely sequenced genomes. We used the annotated *Dictyostelium* genome (accession *GCF*_000004695.1) as our reference to aid in the gene ontology (GO) characterization. We combined results of the three different annotation strategies (EggNOG (http://eggnogdb.embl.de/#/app/home), Blast2GO (Conesa and Götz 2008), and BlastKOALA (http://www.kegg.jp/blastkoala/)) (Fig. 2d). From the total of the 14,580 translated ORFs, EggNOG annotated 11,546 (79%) (Fig. 2d). More than half of these annotated ORFs had an e-value score lower than 1 e^{-50}, and none of the annotated ORFs had a score >1 e^{-10}. Lower score values reflect a strong identity between the sequences generated in the assembly and those present in the databases. About 57.6% of the translated ORFs were annotated across the different gene ontology methods (Eggnog, Blast2Go, BlastKOALA).

We integrated Gene ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. Gene ontology is a standardized nomenclature describing biological process, molecular function, and cellular components. *Dictyostelium discoideum* is a slime mold amoeba, which was used here as the expected model for the constitution of an amoeba metabolism. Although still is phylogenetically distant from *Arcella*, it is the closest organism that have available genome data (Eichinger et al. 2005). Eggnog generated GO terms analysis for functional characterization of the ORFs, for 7,427 ORFs in *Arcella*, compared to 5,676 for *Dictyostelium*. Not all the ORFs have available gene ontology (GO terms), however, when present have more than one class in different levels. Classifying the *Arcella* GO terms in the main GO classes, 1,231 were from cellular components, 2,328 for molecular function, and 8,922 for biological processes (Fig. 3). The two-level classification and a qualitative selection of GO terms showed a similar pattern between *Arcella* and *Dictyostelium* (Fig. 3a–c). Many GO categories have a similar ratio of ORFs present between *Arcella* and *Dictyostelium*. However, several GO categories associated with intracellular signaling were enriched in *Arcella* (signal transduction, MAPK cascade, calcium mediated signaling), all with a ratio of at least two (Fig. 3c). We hypothesize that higher ratios of ORFs present suggest important roles of these pathways in *Arcella*.

KEGG pathways assign genes into pathways, and we assigned unique sequences using KEGG orthology (KO) identifiers in KEGG mapper. We identified ORFs for all major KEGG metabolic pathways in *A. intermedia* (Fig. 4). We mapped 6,966 ORFs in 212 pathways from KEGG (Fig. 4). There are four categories of functions represented: metabolism (33%); genetic information processing (25%); environmental information processing; (19%) and cellular processes (23%). Predominantly, the mapped ORFs were enriched in translation and folding, followed by sorting and degradation, both in genetic information processing category. Among cellular processes, transport and catabolism were predominant, followed by cell growth and death. Signal transduction is enriched in environmental information processing category, and amino acid and carbohydrate metabolisms are both enriched in metabolic processes category. As in the analysis of GO terms, the KEGG pathways also show signaling processes with more genes identified in *Arcella*. These results demonstrate that *Arcella* transcriptome covered the conserved mechanisms known for eukaryotes.

Arcella's identified pathways are predominantly distributed in metabolic processes and gene regulation machineries

We assigned ORFs in 22 KEGG pathways (Fig. 4), and additionally, we identified similarities between *Arcella* and *Dictyostelium* KEGG pathways distribution (Table 2 and Fig. S2). In the next subsections, we will show the main similarities between the two organisms.

Metabolic flexibility of Arcella

Carbohydrate metabolism. Based on KEGG pathways, we identified the majority of enzymes involved in the carbohydrate metabolism. *Arcella* has the ability to process more than one type of monosaccharide (hexokinase, fructokinase, glucokinase, ADP-dependent glucokinases), such as glucose and fructose and also produces glycogen as a carbohydrate reserve (phosphoglucomutase, UTP–glucose-1-phosphate uridylyltransferase, glycogen synthase), similar to what was described for *Dictyostelium* (Williamson et al. 1996; Wright and Albe 1994) and *Acanthamoeba castelanii* (Bowers and Korn 1968). Although some enzymes are absent in *Arcella* transcriptome, this is likely due to incompleteness of transcriptome data and/or rapid rates of evolution that make it difficult to identify homologs for annotation.

Amino acid metabolism. Our data shows the catabolism of amino acids concentrated into TCA-cycle intermediates. Amino acids are important energy sources; they usually can replace carbohydrates as energy substrates. Analyzing KEGG pathways of amino acid metabolism, the enzymes for the following conversions are all present: L-serine/L-

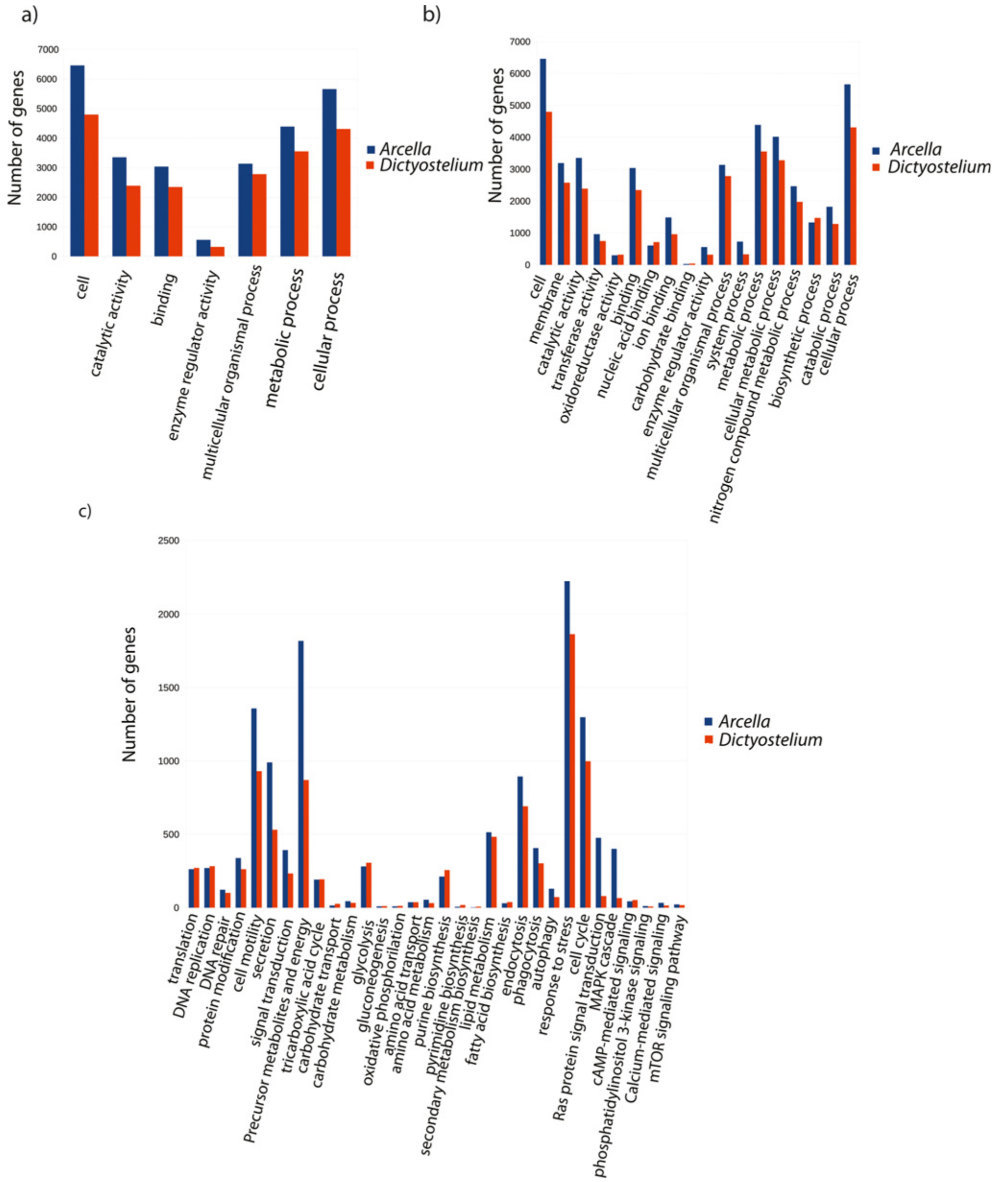

Figure 3 Functional annotation of *Arcella intermedia* transcriptomes using Go terms. Go terms distribution is in three different levels comparing *Arcella intermedia* and *Dictyostelium discoideum*. (**a**) Level 2, (**b**) level 3, and (**c**) a selection of the main Go terms of interest.

threonine ammonia-lyase that converts serine to pyruvate. Aspartate aminotransferase that converts aspartate to oxaloacetate and intermediate of TCA cycle. Glutamate synthase that converts glutamate in 2-oxoglutarate. Leucine, valine, and isoleucine after transamination into the cytosol are oxidized in succinyl-CoA and acetyl-CoA (Fig. 5).

Arcella intermedia does not have any evidence for the presence of an urea cycle; the absence of urea cycle has been demonstrated for several eukaryotic microbes (Ragan 2012). Only diamine oxidase and ornithine-oxo-acid transaminase, enzymes from the entry of this cycle, were identified. The urea cycle, especially in metazoans, is used to process the excess of toxic nitrogen (Krebs 1973; Krebs and Henseleit 1932). *Dictyostelium*, which also lacks the urea cycle pathway, continuously releases nitrogen metabolites in the environment (Payne and Loomis 2006). We hypothesize that *Arcella* as *Dictyostelium* does not use the urea cycle metabolic pathway.

Based on KEGG pathways, *A. intermedia* has the capacity of metabolizing at least 15 of the 20 essential amino acids (alanine, aspartate, asparagine, glutamine, cysteine, tryptophan, glycine, valine, tyrosine, histidine, threonine, leucine, isoleucine, glutamate, arginine, lysine) (Fig. 5). Microbial eukaryotes commonly present losses in amino acid metabolism enzymes and must acquire those amino acids from their food sources. For example, *Dictyostelium* has lost the synthesis capacity for 11 essential amino acids (Payne and Loomis 2006).

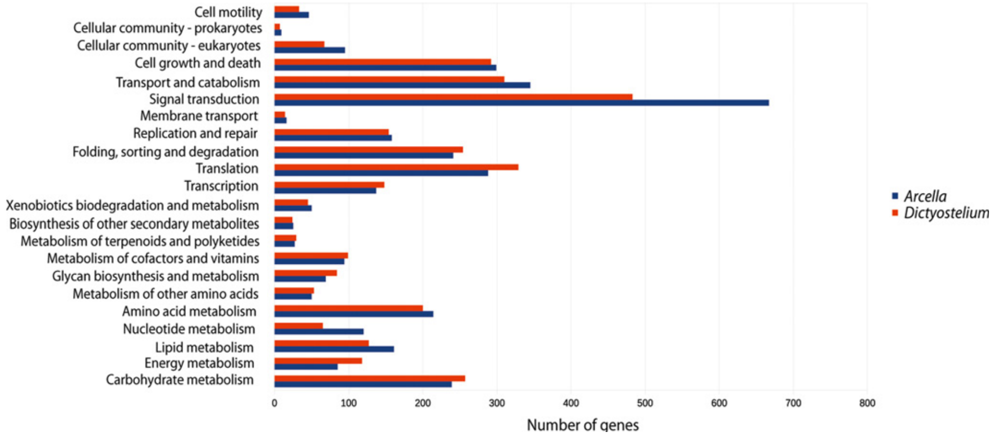

Figure 4 Functional annotation of *Arcella intermedia* transcriptomes using KEGG pathways. Bar chart showing KEGG pathways represented at transcriptomes comparing *A. intermedia* and *Dictyostelium discoideum*.

Signaling complexity of Arcella. The best represented KEGG pathways were RAS, PI3K-AKT, and AMPK (Fig. 5). RAS targets multiple effectors and regulates many signaling pathways, as PI3K, MAPK. PI3K-AKT pathway mainly regulates metabolism, protein synthesis, and the cell cycle process (Manning and Cantley 2007). AMP-activated serine/threonine-protein kinase (AMPK) is an important regulator of energy metabolism, signaling the cell's energy condition. This signaling regulates the production and consumption of energy. The pathway can be activated under conditions of nutrient deprivation or stress (Shaw 2009).

In the processes of signal transduction in *Arcella*, as membrane receptors, we found calcium channels, a few receptors (glutamate, adiponectin, tyrosine-kinase), and also G protein receptors (ADRB2). Receptors can play many roles in signaling pathways. *Arcella*, unlike *Dictyostelium*, shows signaling pathways regulated not only by G protein receptors, but also calcium signaling pathways (ATP2B, CAV1.2, CHRNA7). The intracellular signaling components that initiate signaling cascades found were G proteins, RAS proteins, lipid modifying enzymes that generate other second messengers (the products of phospholipase C and PI3 kinase), and kinase proteins. Calcium is probably acting as secondary messengers in *Arcella*. Calcium action can be concentrated in the cytoplasm regulating kinases (PKC) and signaling pathways, but can also have actions in the nucleus through calmodulin regulating gene expression (Tandoğan and Ulusu 2005).

We found parts of many signaling pathways represented in *Arcella* transcriptome. However, we were unable to resolve the entire pathway in many cases. For example, MAPK-ERK pathways which are important for cell growth and differentiation in many organisms have an atypical pathway in *Dictyostelium* (Hadwiger and Nguyen 2011). The differences can be related to the protein structure and also to components of the pathway. Because we are working with transcriptome data, we cannot say if the proteins could not be identified because of differential evolution between genes or if they have different pathways performing the same functions.

Gene regulation machineries identified are mostly related to translation and protein processing. *Arcella* have almost all components of genetic information processing, however, constituted by both prokaryotic and eukaryotic machineries. In all categories, we identified on average 70% of the constituent proteins. Almost all enzymes related to the formation of ribosomes, to the ubiquitin-proteasome system or DNA replication, were present. *Arcella intermedia* has genes for the formation of at least 19 tRNAs (Table 2). Part of the DNA repair machineries corresponds to the prokaryotic type on KEGG. This pattern is common between *Arcella* and *Dictyostelium,* and many eukaryotic repair machineries are actually derived from archaea and are conserved between eukaryotes and prokaryotes (Hofstatter et al. 2016; Makarova et al. 2014). In addition, phylogenetic information can demonstrate if the sequences detected belong to a prokaryotic carryover and also the evolutionary history of these machineries.

Functioning of *Arcella* in significant processes related to a variety of cellular functions

Endocytosis/Phagocytosis

Clathrin-mediated endocytosis is characterized as a cooperative process resulting in the internalization of concentrated particles (McPherson et al. 2008). The machinery of clathrin recruits a series of proteins and cofactors. Our database shows that *A. intermedia* have members of the epsin (eps15) and adaptin (AP-2) complex family being recruited. Epsin action results in tubulation or even vesiculation as a result of membrane bending (Gleisner et al. 2016). Adaptins (AP-2) are responsible for the recognition of the particles ingested and retention inside the vesicle (Boehm and Bonifacino 2001).

We found in *A. intermedia* at least seven RAB proteins modulating vesicle trafficking: Rab5 mediates fusion of endosomes (Hoffenberg et al. 2000); Rab11 controls slow recycling pathways (Grant and Donaldson 2009), while Rab35 regulates fast recycling (Kouranti et al. 2006); Rab 7 regulates transport to the lysosome (Zhang et al. 2009).

Table 2. Function comparison using KEGG pathways. We identified the number of sequences in each of the KEGG pathways and compared it between *A. intermedia* and *D. discoideum* which is a model organism

Function	Pathway	Presence *Arcella intermedia*	Presence *Dictyostelium discoideum*
Metabolism			
Carbohydrate metabolism	Glycolysis	21	26
	TCA cycle	19	21
Energy metabolism	Oxidative phosphorylation	41	61
Lipid metabolism	Fatty acid synthesis and degradation	30	28
Nucleotide metabolism	Purine metabolism	28	43
	Pyrimidine metabolism	22	22
Amino acid metabolism	Biosynthesis of amino acids	65	47
Genetic information processing			
Transcription	RNA pol	23	24
	Basal transcription factors	21	28
	Spliceosome	97	96
Translation	Ribosome	93	106
	Aminoacyl-tRNA biosynthesis	26	27
	RNA transport	75	89
	mRNA surveillance pathway	43	46
	Ribosome biogenesis in eukaryotes	56	61
Folding, sorting and degradation	Protein export	17	20
	Protein processing in endoplasmic reticulum	70	74
	Ubiquitin mediated proteolysis	63	61
	Proteasome	35	35
Replication and repair	DNA replication	31	31
	Base excision repair	20	20
	Nucleotide excision repair	31	31
	Mismatch repair	18	19
	Homologous recombination	26	23
	Nonhomologous end-joining	8	7
Environmental information processing			
Membrane transport		15	11
Signal transduction	Ras signaling	45	25
	MAPK signaling	48	24
	PI3K-AKT	42	33
	AMPK	38	34
	mTOR	50	43
	cAMP	27	15
	Calcium	22	12
	Phosphatidylinositol	25	27
Cellular processes			
Transport and catabolism	Endocytosis	77	65
	Phagosome	33	30
	Peroxisome	45	32
	Autophagy	18	22
Cell growth and death	Cell cycle	54	56
	Apoptosis	23	16

We also see RAB8 and RAB10, responsible for regulation of transport of some specific types of particles (Hutagalung and Novick 2011). In general, endocytosis pathway seems to be nicely represented at the transcriptome, with most part of the components present. Endocytosis/phagocytosis pathways are expected to be important in amoeboid organisms, as they use for feeding of large particles, communication, and transport.

The core cell cycle proteins found in Arcella are similar to those in Dictyostelium

Arcella cell cycle complement proteins at KEGG were nearly identical to the proteins set identified in *Dictyostelium* (Table 2). However, the cell cycle of *Dictyostelium* is known to be noncanonical. This is because in starvation, the formation of a prestalk structure involves divergent signaling and proteins (Weeks and Weijer 1994).

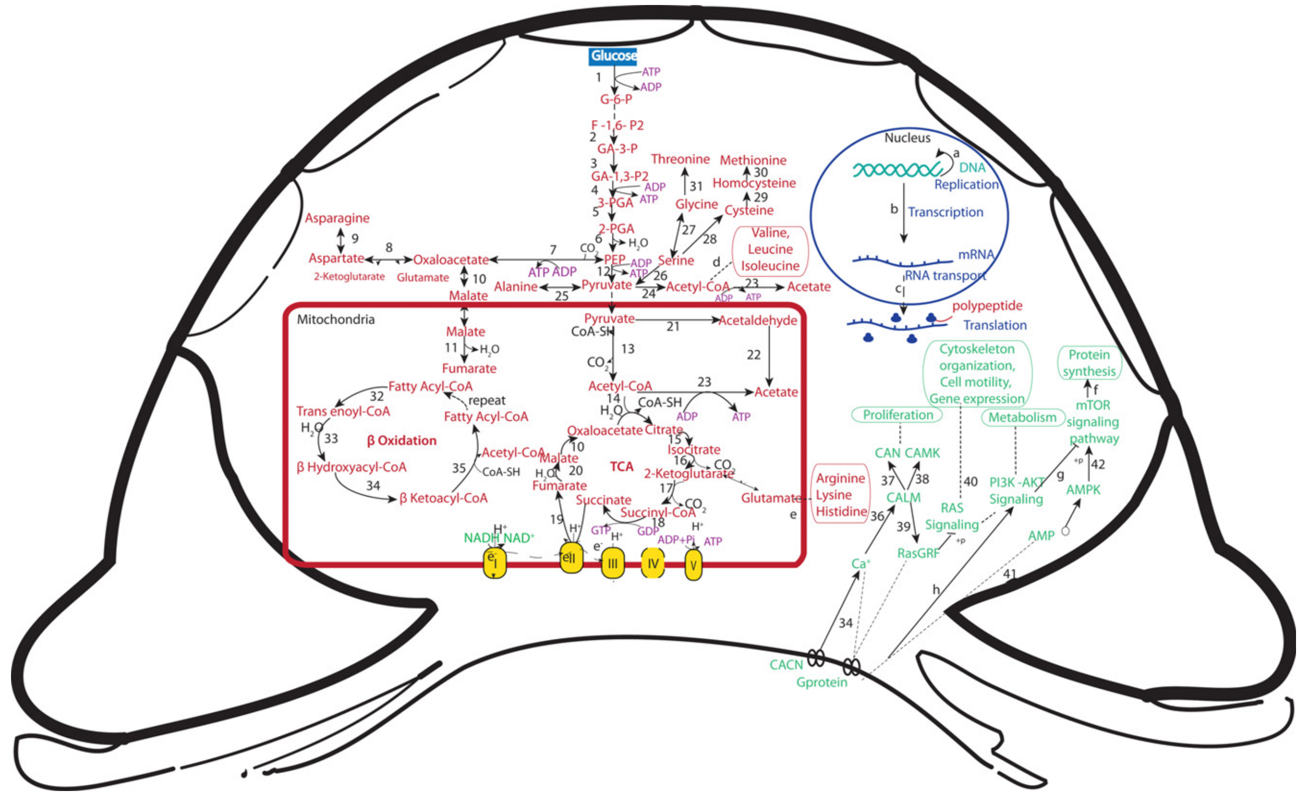

Figure 5 Representation of *Arcella intermedia* metabolism with main pathways based on RNA-seq analysis. Colors correspond to the main KEGG categories: metabolism (red), genetic processing information (blue), environmental signaling and metabolism (green). Enzyme names: [1] hexokinase; [2] aldolase; [3] glyceraldehyde phosphate dehydrogenase; [4] phosphoglycerate kinase; [5] phospho-glycerate mutase; [6] enolase; [7] phosphoenolpyruvate carboxykinase; [8] aspartate aminotransferase; [9] asparagine synthase; [10] malate dehydrogenase; [11] fumarate hydratase; [12] pyruvate kinase; [13] pyruvate dehydrogenase complex; [14] citrate synthase; [15] aconitate hydratase; [16] isocitrate dehydrogenase; [17] 2-oxoglutarate dehydrogenase; [18] succinyl-CoA synthetase; [19] succinate dehydrogenase; [20] fumarate hydratase; [21] pyruvate decarboxylase; [22] aldehyde dehydrogenase; [23] Acetyl-CoA synthase (ADP forming); [24] pyruvate:ferodoxin oxidoreductase; [25] alanine transaminase; [26] L-serine ammonia-lyase; [27] glycine hidroximetil transferase; [28] cystathione synthase + lyase; [29] cystathione synthase; [30] homocysteine methyltransferase; [31] threonine aldolase; [32] acetyl-CoA dehydrogenase; [33] enoyl-CoA hydratase; [34] 3-hydroxiacil dehydrogenase; [35] acetyl-CoA transferase; [CACN] calcium voltage-gated channel; [36] calmodulin; [37] serine/threonine-protein phosphatase; [38] calcium/calmodulin dependent protein kinase; [39] ras-specific guanine nucleotide-releasing factor; [40] ras-related protein; [41] 5′-AMP-activated protein kinase; [42] serine/threonine-protein kinase mTOR. Letters a, b, c, d, e, f, and g are representing pathways composed by a set of enzymes: [a] replication; [b] transcription; [c] RNA transport; [d] valine, leucine, and isoleucine metabolism; [e] arginine, lysine, and histidine metabolism; [f] mTor signaling pathway; [g] PI3K-AKT signaling pathway; [h] G protein cascade.

In replication pathway, *Arcella* presents two transcription factors of metazoa (Rb and E2F), distinct from *Dictyostelium* (E2F and DP 1,2). There are 6 types of origins of replication (ORC) in metazoans and yeast. *Arcella* has only two (Orc 1, Orc 2) and *Dictyostelium* three (Orc 1, 2, and 4). Cdc6, which is the control point to progression to the S-phase of the cell cycle, is also present in *Arcella*. Other components related to the regulation of transcription for metazoa are absent in *Arcella* transcriptome, as p107/130 and E2F4/5. *Arcella* possesses cyclin A/B and H; these proteins control cell cycle progression. *Dictyostelium* has only cyclin A and H. *Arcella* possesses the CDKs correspondent to metazoa, CDK4,6 and CDK2, CDK7. *Dictyostelium* also lacks some CDKs; only CDK2 and CDK7 were identified. *Arcella* has all components of cohesin and separin, proteins related to chromosomal division in eukaryotes. We did not find any of the known securin, which is the regulator of separin (Rao et al. 2001). We also did not identify securin and separin in *Dictyostelium*. If in *Dictyostelium* (an organism with genomic data available) we have not found these proteins, we can infer that other protein is developing the same function; or, the sequences in these organisms are already too different to be detected. The transcriptome of *Arcella intermedia* does not contain obvious markers involved with cell death processes, though apoptotic markers have been observed in unicellular eukaryotes, including *Dictyostelium* (Deponte 2008); However, as we are working with transcriptomic data, if the cell is not dying, we probably will not find the proteins related to the process present.

CONCLUSIONS AND PERSPECTIVES

Here, we presented an annotated transcriptome for *A. intermedia*, a lineage of testate amoeba member of the order Arcellinida (Amoebozoa). The annotation was generated from pooled analyses of 38 transcriptomes in different stages of life. Because *Arcella* cannot be cultured axenically, exclusion of genetic material carryover (e.g. from food sources) is difficult; therefore, we used a conservative approach in our annotations. Based on our dataset, we analyzed well-known genes and pathways and also mapped 267 of 303 (88.1%) target core eukaryotic genes identified by BUSCO. Culture data provide a more complete set of genes then single-cell transcriptomes; however, single-cell experiments contribute with some specific genes. We analyzed KEGG pathways components for entire transcriptome and for each category the main components found were carbohydrate and amino acid metabolism, for the metabolism category; formation and processing of proteins for the genetic information processing; signal transduction for environmental signaling; and transport/catabolism and cell growth/death for cellular processes. In comparison with *Dictyostelium*, a model organism with a well annotated publicly available genome, *Arcella* data have a compatible set of proteins in KEGG. In all categories on average, 70% of the constituent proteins were present. The results presented in this work indicate the relative completeness of our *A. intermedia* transcriptome at least for the known eukaryotic genes expected. In this work, we provided a reference database for *A. intermedia*, which could be helpful in future ecology, physiology, or evolutionary studies.

ACKNOWLEDGMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brazil (CAPES)—Finance Code 001. Funding for this work was also provided by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) fellowships to DJGL (#2013/04585); GMR (#2017/16077-3, #2016/14402-1); APS (#2016/14317-4).

LITERATURE CITED

Andrey, T. & Yuri, M. 2006. Morphology, biometry and ecology of *Arcella gibbosa* Penard 1890 (Rhizopoda, Testacealobosea). *Protistology*, 4(3): 279–294.

Boehm, M. & Bonifacino, J. S. 2001. Adaptins: the final recount. *Mol. Biol. Cell*, 12(10):2907–2920.

Bowers, B. & Korn, E. D. 1968. The fine structure of Acanthamoeba castellanii: I. the trophozoite. *J. Cell Biol.*, 39(1):95–111.

Cerón-Romero, M. A., Maurer-Alcalá, X. X., Grattepanche, J.-D., Yan, Y., Fonseca, M. M. & Katz, L. 2019. Phylotol: a taxon/gene-rich phylogenomic pipeline to explore genome evolution of diverse eukaryotes. *Mol. Biol. Evol.*, 36:1831–1842.

Clarke, M., Lohan, A. J., Liu, B., Lagkouvardos, I., Roy, S., Zafar, N., Bertelli, C., Schilde, C., Kianianmomeni, A., Bürglin, T. R., Frech, C., Turcotte, B., Kopec, K. O., Synnott, J. M., Choo, C., Paponov, I., Finkler, A., Heng Tan, C. S., Hutchins, A. P., Weinmeier, T., Rattei, T., Chu, J. S., Gimenez, G., Irimia, M., Rigden, D. J., Fitzpatrick, D. A., Lorenzo-Morales, J., Bateman, A., Chiu, C. H., Tang, P., Hegemann, P., Fromm, H., Raoult, D., Greub, G., Miranda-Saavedra, D., Chen, N., Nash, P., Ginger, M. L., Horn, M., Schaap, P., Caler, L. & Loftus, B. J. 2013. Genome of Acanthamoeba castellanii highlights extensive lateral gene transfer and early evolution of tyrosine kinase signaling. *Genome Biol.*, 14(2):R11.

Conesa, A. & Götz, S. 2008. Blast2go: a comprehensive suite for functional analysis in plant genomics. *Int. J. Plant Genomics*, 2008:619832.

Deponte, M. 2008. Programmed cell death in protists. *Biochim. Biophys. Acta*, 1783(7):1396–1405.

Eichinger, L., Pachebat, J., Glöckner, G., Rajandream, M.-A., Sucgang, R., Berriman, M., Song, J., Olsen, R., Szafranski, K., Xu, Q., Tunggal, B., Kummerfeld, S., Madera, M., Konfortov, B. A., Rivero, F., Bankier, A. T., Lehmann, R., Hamlin, N., Davies, R., Gaudet, P., Fey, P., Pilcher, K., Chen, G., Saunders, D., Sodergren, E., Davis, P., Kerhornou, A., Nie, X., Hall, N., Anjard, C., Hemphill, L., Bason, N., Farbrother, P., Desany, B., Just, E., Morio, T., Rost, R., Churcher, C., Cooper, J., Haydock, S., van Driessche, N., Cronin, A., Goodhead, I., Muzny, D., Mourier, T., Pain, A., Lu, M., Harper, D., Lindsay, R., Hauser, H., James, K., Quiles, M., Madan Babu, M., Saito, T., Buchrieser, C., Wardroper, A., Felder, M., Thangavelu, M., Johnson, D., Knights, A., Loulseged, H., Mungall, K., Oliver, K., Price, C., Quail, M. A., Urushihara, H., Hernandez, J., Rabbinowitsch, E., Steffen, D., Sanders, M., Ma, J., Kohara, Y., Sharp, S., Simmonds, M., Spiegler, S., Tivey, A., Sugano, S., White, B., Walker, D., Woodward, J., Winckler, T., Tanaka, Y., Shaulsky, G., Schleicher, M., Weinstock, G., Rosenthal, A., Cox, E. C., Chisholm, R. L., Gibbs, R., Loomis, W. F., Platzer, M., Kay, R. R., Williams, J., Dear, P. H., Noegel, A. A., Barrell, B. & Kuspa, A. 2005. The genome of the social amoeba dictyostelium discoideum. *Nature*, 435(7038):43.

Eldem, V., Zararsiz, G., Taşçi, T., Duru, I. P., Bakir, Y. & Erkan, M. 2017. Transcriptome analysis for non- model organism: current status and best-practices. *In:* Applications of RNA-Seq and Omics Strategies-from Microorganisms to Human Health. IntechOpen, Rijeka, p. 55–71.

Escobar, J., Brenner, M., Whitmore, T. J., Kenney, W. F. & Curtis, J. H. 2008. Ecology of testate amoebae (thecamoebians) in subtropical florida lakes. *J. Paleolimnol.*, 40(2):715–731.

Gleisner, M., Kroppen, B., Fricke, C., Teske, N., Kliesch, T.-T., Janshoff, A., Meinecke, M. & Steinem, C. 2016. Epsin n-terminal homology domain (enth) activity as a function of membrane tension. *J. Biol. Chem.*, 291:19953–19961.

Gomaa, F., Todorov, M., Heger, T. J., Mitchell, E. A. & Lara, E. 2012. SSU rRNA phylogeny of Arcellinida (Amoebozoa) reveals that the largest Arcellinid genus, Difflugia Leclerc 1815, is not monophyletic. *Protist*, 163(3):389–399.

Grant, B. D. & Donaldson, J. G. 2009. Pathways and mechanisms of endocytic recycling. *Nat. Rev. Mol. Cell Biol.*, 10(9):597.

Hadwiger, J. A. & Nguyen, H.-N. 2011. Mapks in development: insights from dictyostelium signaling pathways. *Biomol. Concepts*, 2(1–2):39–46.

Hoffenberg, S., Liu, X., Nikolova, L., Hall, H. S., Dai, W., Baughn, R. E., Dickey, B. F., Barbieri, M. A., Aballay, A., Stahl, P. D. & Knoll, B. J. 2000. A novel membrane-anchored rab5 interacting protein required for homotypic endosome fusion. *J. Biol. Chem.*, 275(32):24661–24669.

Hofstatter, P. G., Tice, A. K., Kang, S., Brown, M. W. & Lahr, D. J. 2016. Evolution of bacterial recombinase A (recA) in eukaryotes explained by addition of genomic data of key microbial lineages. *Proc. R. Soc. B.*, 283(1840):20161453. https://doi.org/10.1098/rspb.2016.1453

Hutagalung, A. H. & Novick, P. J. 2011. Role of Rab GTPases in membrane traffic and cell physiology. *Physiol. Rev.*, 91(1):119–149.

Jensen, L. J., Julien, P., Kuhn, M., von Mering, C., Muller, J., Doerks, T. & Bork, P. 2007. eggnog: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res.*, 36(Suppl 1):D250–D254.

Kouranti, I., Sachse, M., Arouche, N., Goud, B. & Echard, A. 2006. Rab35 regulates an endocytic recycling pathway essential for the terminal steps of cytokinesis. *Curr. Biol.*, 16(17):1719–1725.

Krebs, H. A. 1973. The discovery of the ornithine cycle of urea synthesis. *Biochem. Educ.*, 1(2):19–23.

Krebs, H. & Henseleit, K. 1932. Formation of urea in the animal body. *Hoppe-Seyler's Z. Physiol. Chem.*, 210:33–66.

Loftus, B., Anderson, I., Davies, R., Alsmark, U. C. M., Samuelson, J., Amedeo, P., Roncaglia, P., Berriman, M., Hirt, R. P., Mann, B. J., Nozaki, T., Suh, B., Pop, M., Duchene, M., Ackers, J., Tannich, E., Leippe, M., Hofer, M., Bruchhaus, I., Willhoeft, U., Bhattacharya, A., Chillingworth, T., Churcher, C., Hance, Z., Harris, B., Harris, D., Jagels, K., Moule, S., Mungall, K., Ormond, D., Squares, R., Whitehead, S., Quail, M. A., Rabbinowitsch, E., Norbertczak, H., Price, C., Wang, Z., Guillén, N., Gilchrist, C., Stroup, S. E., Bhattacharya, S., Lohia, A., Foster, P. G., Sicheritz-Ponten, T., Weber, C., Singh, U., Mukherjee, C., El-Sayed, N. M., Petri Jr, W. A., Clark, C. G., Embley, T. M., Barrell, B., Fraser, C. M. & Hall, N. 2005. The genome of the protist parasite entamoeba histolytica. *Nature*, 433(7028):865.

Makarova, K. S., Krupovic, M. & Koonin, E. V. 2014. Evolution of replicative DNA polymerases in Archaea and their contributions to the eukaryotic replication machinery. *Front. Microbiol.*, 5:354.

Manning, B. D. & Cantley, L. C. 2007. Akt/pkb signaling: navigating downstream. *Cell*, 129(7):1261–1274.

Maurer-Alcalá, X. X., Knight, R. & Katz, L. A. 2018a. Exploration of the germline genome of the ciliate chilodonella uncinata through single-cell omics (transcriptomics and genomics). *MBio*, 9(1):e01836-17.

Maurer-Alcalá, X. X., Yan, Y., Pilling, O. A., Knight, R. & Katz, L. A. 2018b. Twisted tales: insights into genome diversity of ciliates using single-cell 'omics'. *Genome Biol. Evol.*, 10(8):1927–1938.

McPherson, P. S., Ritter, B. & Wendland, B. 2008. Clathrin-Mediated Endocytosis. *In:* Madame Curie Bioscience Database. Landes Bioscience, Austin, TX. p. 2000–2013.

Meisterfeld, R. 2002. Order Arcellinida Kent, 1880. *In:* Lee, J. J., Leedale, G. F. & Bradbury, P. (ed.), The Illustrated Guide to the Protozoa, Vol. 2, Society of Protozoologists, Allen Press In., Lawrence, KA. p. 827–860.

Mitchell, E. A., Charman, D. J. & Warner, B. G. 2008. Testate amoebae analysis in ecological and paleoecological studies of wetlands: past, present and future. *Biodivers. Conserv.*, 17 (9):2115–2137.

Ogden, G. & Hedley, R. H. 1980. An atlas of freshwater testate amoebae. *Soil Sci.*, 130(3):176.

Patterson, R. T. & Kumar, A. 2000. Assessment of arcellacean (thecamoebian) assemblages, species, and strains as contaminant indicators in James Lake, Northeastern Ontario, Canada. *J. Foramin. Res.*, 30(4):310–320.

Payne, S. H. & Loomis, W. F. 2006. Retention and loss of amino acid biosynthetic pathways based on analysis of whole-genome sequences. *Eukaryot. Cell*, 5(2):272–276.

Picelli, S., Faridani, O. R., Björklund, Å. K., Winberg, G., Sagasser, S. & Sandberg, R. 2014. Full-length RNA-seq from single cells using smart-seq2. *Nat. Protoc.*, 9(1):171.

Ragan, M. 2012. A biochemical phylogeny of the protists. Academic Press INC., New York.

Rao, H., Uhlmann, F., Nasmyth, K. & Varshavsky, A. 2001. Degradation of a cohesin subunit by the n-end rule pathway is essential for chromosome stability. *Nature*, 410(6831):955.

Reinhardt, E. G., Dalby, A. P., Kumar, A. & Patterson, R. T. 1998. Arcellaceans as pollution indicators in mine tailing contaminated lakes near cobalt, Ontario, Canada. *Micropaleontology*, 44 (2):131–148.

Roe, H. M. & Patterson, R. T. 2014. Arcellacea (testate amoebae) as bio-indicators of road salt contamination in lakes. *Microb. Ecol.*, 68(2):299–313.

Roe, H. M., Patterson, R. T. & Swindles, G. T. 2010. Controls on the contemporary distribution of lake thecamoebians (testate amoebae) within the greater Toronto area and their potential as water quality indicators. *J. Paleolimnol.*, 43(4):955–975.

Schaap, P., Barrantes, I., Minx, P., Sasaki, N., Anderson, R. W., Bénard, M., Biggar, K. K., Buchler, N. E., Bundschuh, R., Chen, X., Fronick, C., Fulton, L., Golderer, G., Jahn, N., Knoop, V., Landweber, L. F., Maric, C., Miller, D., Noegel, A. A., Peace, R., Pierron, G., Sasaki, T., Schallenberg-Rüdinger, M., Schleicher, M., Singh, R., Spaller, T., Storey, K. B., Suzuki, T., Tomlinson, C., Tyson, J. J., Warren, W. C., Werner, E. R., Werner-Felmayer, G., Wilson, R. K., Winckler, T., Gott, J. M., Glöckner, G. & Marwan, W. 2015. The physarum polycephalum genome reveals extensive use of prokaryotic two-component and metazoan-type tyrosine kinase signaling. *Genome Biol. Evol.*, 8(1):109–125.

Shaw, R. J. 2009. Lkb1 and amp-activated protein kinase control of mTOR signalling and growth. *Acta Physiol.*, 196(1):65–80.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. 2015. Busco: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19):3210–3212.

Tandoğan, B. & Ulusu, N. N. 2005. Importance of calcium. *Turk. J. Med. Sci.*, 35(4):197–201.

Waterhouse, R. M., Seppey, M., Simão, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E. V. & Zdobnov, E. M. 2017. Busco applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.*, 35(3):543–548.

Weeks, G. & Weijer, C. J. 1994. The dictyostelium cell cycle and its relationship to differentiation. *FEMS Microbiol. Lett.*, 124 (2):123–130.

Williamson, B. D., Favis, R., Brickey, D. A. & Rutherford, C. L. 1996. Isolation and characterization of glycogen synthase in dictyostelium discoideum. *Dev. Genet.*, 19(4):350–364.

Wright, B. E. & Albe, K. R. 1994. Carbohydrate metabolism in dictyostelium discoideum: I. model construction. *J. Theor. Biol.*, 169(3):231–241.

Zhang, M., Chen, L., Wang, S. & Wang, T. 2009. Rab7: roles in membrane trafficking and disease. *Biosci. Rep.*, 29(3):193–209.

[1] Present address: Institute of Cell Biology, University of Bern, Baltzerstrasse 4, 3012, Bern, Switzerland.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure S1. Summary of the methods applied in the study.

Figure S2. Comparison of the metabolism of *Arcella* (red) and *Dictyostelium* (blue).

Figure S3. Classification of the SSU rDNA sequences found in *Arcella intermedia* transcriptomes.

Table S1. Properties of the samples produced in this work, which contains the accession number, condition, number of the culture bottle and if is single-cell or total culture sampling.