Current Biology 24, 11–18, January 6, 2014 ©2014 Elsevier Ltd All rights reserved http://dx.doi.org/10.1016/j.cub.2013.11.027

Article

The Genome of the Foraminiferan *Reticulomyxa filosa*

Gernot Glöckner,[1,2,*] Norbert Hülsmann,[3] Michael Schleicher,[4] Angelika A. Noegel,[1] Ludwig Eichinger,[1] Christoph Gallinger,[4] Jan Pawlowski,[5] Roberto Sierra,[5] Ursula Euteneuer,[4] Loic Pillet,[5] Ahmed Moustafa,[6] Matthias Platzer,[7] Marco Groth,[7] Karol Szafranski,[7] and Manfred Schliwa[4]

[1]Institute for Biochemistry I, Medical Faculty, Center for Molecular Medicine Cologne (CMMC), Cologne Excellence Cluster on Cellular Stress Responses in Aging-Associated Diseases (CECAD), University of Cologne, Joseph-Stelzmann-Strasse 52, 50931 Köln, Germany

[2]Leibniz Institute of Freshwater Ecology and Inland Fisheries, IGB Müggelseedamm 301, 12587 Berlin, Germany

[3]Protozoology, Freie Universität Berlin, Königin-Luise-Strasse 1–3, 14195 Berlin

[4]Institute for Anatomy and Cell Biology, Ludwig-Maximilians-Universität München, Schillerstrasse 42, 80336 München, Germany

[5]Department of Genetics and Evolution, University of Geneva, 4 Boulevard D'Yvoy, 1205 Genève, Switzerland

[6]Department of Biology and Biotechnology Graduate Program, American University in Cairo, New Cairo 11835, Egypt

[7]Genome Analysis, Fritz Lipmann Institute, Beutenbergstrasse 11, 07745 Jena, Germany

Summary

Background: Rhizaria are a major branch of eukaryote evolution with an extensive microfossil record, but only scarce molecular data are available. The rhizarian species *Reticulomyxa filosa*, belonging to the Foraminifera, is free-living in freshwater environments. In culture, it thrives only as a plasmodium with thousands of haploid nuclei in one cell. The *R. filosa* genome is the first foraminiferal genome to be deciphered.

Results: The genome is extremely repetitive, and the large amounts of identical sequences hint at frequent amplifications and homologous recombination events. Presumably, these mechanisms are employed to provide more gene copies for higher transcriptional activity and to build up a reservoir of gene diversification in certain gene families, such as the kinesin family. The gene repertoire indicates that it is able to switch to a single-celled, flagellated sexual state never observed in culture. Comparison to another rhizarian, the chlorarachniophyte alga *Bigelowiella natans*, reveals that proteins involved in signaling were likely drivers in establishing the Rhizaria lineage. Compared to some other protists, horizontal gene transfer is limited, but we found evidence of bacterial-to-eukaryote and eukaryote-to-eukaryote transfer events.

Conclusions: The *R. filosa* genome exhibits a unique architecture with extensive repeat homogenization and gene amplification, which highlights its potential for diverse life-cycle stages. The ability of *R. filosa* to rapidly transport matter from the pseudopodia to the cell body may be supported by the high diversification of actin and kinesin gene family members.

Introduction

The eukaryote tree of life consists of six to eight major branches [1], among which is the supergroup Rhizaria that contains the majority of skeleton-building protists [2]. In most of the branches, whole genomes of several species were deciphered and analyzed, enabling comparative genomics studies. These analyses yielded invaluable insights into eukaryote evolution. However, only one Rhizarian genome is currently available, that of *Bigelowiella natans* [3], making it the most poorly sampled of the major branches of eukaryotes. To fill this gap, we report the genome of *Reticulomyxa filosa*, a representative of the Foraminifera, the most species-rich clade of Rhizaria. Foraminifera are unicellular but can develop huge multinuclear cells with extraordinarily varied multichambered calcareous tests. They are best known as microfossils used widely as paleostratigraphic and paleoecological indicators [4]. Extant Foraminifera play a key ecological role as the most abundant and diversified component of marine meiofauna and the major contributor to the global carbon cycle. Foraminifera are also renowned for their distinctive pseudopodia, which form large, dynamic networks and display rapid transport processes enabling dynamic interactions with their environment [5]. Recent studies associate these particular features to the unusual and highly divergent β-tubulin [6], but the genomic basis of foraminiferal movement is largely unknown. Little is also known about the intriguing characteristics of other foraminiferal genes, including the hypervariable rRNA genes that are commonly used to study the phylogeny and diversity of this group [7].

R. filosa (Figure 1) represents the best known, but not the only, freshwater species of the otherwise marine Foraminifera. It thrives on the bottom of lakes and streams. The stationary thicker center part of the plasmodium is mostly hidden in the ground, whereas the slender peripheral pseudopodia spread over the substrate and are active in food uptake and transport to the center. Most strains of this species stem from aquarium tanks and garden ponds and were isolated independently several times between 1937 and 1984 [8–10]; the sequenced strain originates from Lake Möwensee near Fürstenberg/Havel (Brandenburg, Germany), where it was discovered in 1993 [11]. *R. filosa* forms giant net-like plasmodia, often up to 10 cm or more in diameter, consisting of numerous thick veins in the central area and slender reticulopodia in the periphery, which exhibit a distinctive bidirectional streaming behavior typical for Foraminifera. The plasmodial stage is haploid and contains hundreds or thousands of nuclei with a diameter of about 5 μm. The nuclear divisions occur synchronously and are performed as closed mitosis [11]. Under adverse conditions, the plasmodia undergo encystation.

As a representative of the so far poorly sampled supergroup of Rhizaria, *R. filosa* provides insights into Rhizaria evolution and general eukaryote conditions. Thus, it allows the narrowing down of the minimally required eukaryote gene complement. Furthermore, it helps to describe the specificity of Foraminifera; in particular, it contributes to a better

*Correspondence: gernot.gloeckner@uni-koeln.de

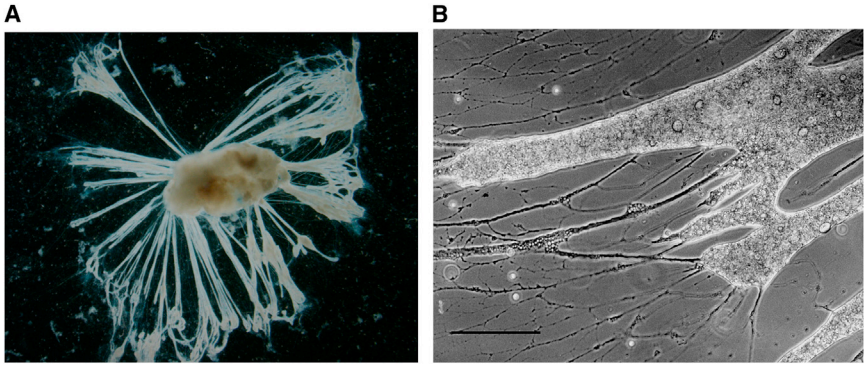

Figure 1. *Reticulomyxa filosa*, *Bochum* Strain

(A) An acellular syncytium (plasmodium) with thousands of nuclei. Very rapid plasma streams transport food to the cell body.

(B) Periphery of the central area with thicker strands and filose reticulopodia.

See also Figure S3.

understanding of the mechanisms responsible for reticulopodial motility, molecular interactions in the cytoskeleton, and the unusual fast machinery for organelle transport. Accordingly, *R. filosa* has already been used as a model system for the analysis of cellular movement and transport [12–14]. The genome information thus will make such analyses much easier and can also help to elucidate Foraminifera-specific evolutionary developments and features, such as the complex life cycle, endoreplication, and nuclear dimorphism [15]. Moreover, it has been hypothesized that all Rhizaria are derived from a common ancestor, which possessed a photosynthetic endosymbiotic partner [16]. The scars of this evolutionary history should be present even in heterotrophic members of this clade.

Results

The Genome

From a *R. filosa* culture, we isolated nuclei using a standard yeast protocol [17]. We then produced 1.6 Gb of raw sequences from this DNA with the Roche/454 GS FLX Titanium platform and 300 million 76 bp Illumina paired-end reads, totaling nearly 23 Gb of raw sequences. Assemblies constructed with different programs yielded around 100 Mb of contig sequences (see Table S1 available online). Yet, previous analysis using flow cytometry showed that the *R. filosa* genome size should be in the range of 400 Mb. We remapped all raw reads from the Illumina sequencer back to the assembled contigs larger than 5 kb to obtain a measure of overall coverage. This approach yielded a mean contig coverage value of around 70, accounting for a genome size of around 320 Mb. To get an impression of the true coverage distribution in the genome, we completed the sequences of 15 fosmid clones using Sanger technology and mapped the raw sequencing reads of the whole genome to these genomic regions. We found that the overall sequence coverage per fosmid varied widely, from 110 times to more than 24,000, but that the mean coverage of most bases in the fosmids was also around 70, whereas some specific regions or even whole fosmids were heavily covered (Figure 2). We extracted highly covered regions from the fosmid sequences and found that these consist mostly of short sequences, often directly or indirectly repeated. The 20 most highly covered regions of all fosmids comprise 1.415 bases (mean 67 bases) yet account, according to their coverage, for more than 16 Mb of the *R. filosa* genome. This indicates not only that the *R. filosa* genome is littered with repetitive sequences but also that many repeats are highly similar, since they can be readily mapped despite a rigorous identity threshold. Also, long-range duplications and/or amplifications seem to exist, which makes it unlikely that the fosmid sequences can unambiguously be connected to yield a contiguous genome sequence. We also mapped more than 16,000 fosmid end sequences to the assembly and found that only 25% of these could be mapped over their entire length to it. However, shorter stretches of the fosmid end sequences had counterparts in the genome assembly, further supporting our notion that the *R. filosa* genome contains a lot of simple repeats. Thus, the current *R. filosa* genome assembly is most likely the best obtainable with presently available sequencing techniques.

Completeness

Since the genome is highly repetitive and the assembly size differs considerably from the previously estimated size, we laid emphasis on the analysis of the completeness of the assembly in terms of coding capacity. The transcriptome data indicate that the protein-coding part of the genome is well represented in our assembly. 93% of the available 1,630 expressed sequence tag (EST) sequences and nearly all of our RNA-seq transcriptome data (99.5%) could be mapped to the genome assembly. A core eukaryotic genes mapping approach (CEGMA) analysis using cegma_v2.4 [18] with the *R. filosa* and *B. natans* genomes revealed that the completeness of both genomes is alike and that most of the CEGMA gene set is also found in the RNA-seq data. Furthermore, we conducted an analysis of biochemical pathways of the primary metabolism using KEGG [19, 20]. All primary metabolism genes are present in our predicted protein data set, as expected for a free-living species (Table S2).

We observed a number of contigs in our assembly with only half the coverage of the other contigs. An analysis of these contigs revealed that they were derived from a *Rickettsia*-like bacterium. Indeed, all of these contigs sum up to a little more than 1 Mb, which is in the normal genome size range of this genus. *Rickettsia* species are able to enter the nucleus of eukaryotes and thrive there [21]. We found this DNA in different preparations of nuclei from culture samples separated by long culturing periods, indicating stable maintenance of this parasite. These contigs were removed from the assembly prior to further analysis of the *R. filosa* genome.

We also screened the predicted coding genes for the presence of transposon-associated domains (transposases, endonucleases, and reverse transcriptases) to get an estimation of how widespread such genomic components are in this genome and identified 114 such domains. We then obtained the sequence coverage of the respective contigs and found that they are not or are only slightly (up to five times) overrepresented in the raw data. Thus, we conclude that transposons are not very active in the genome. Moreover, many of the

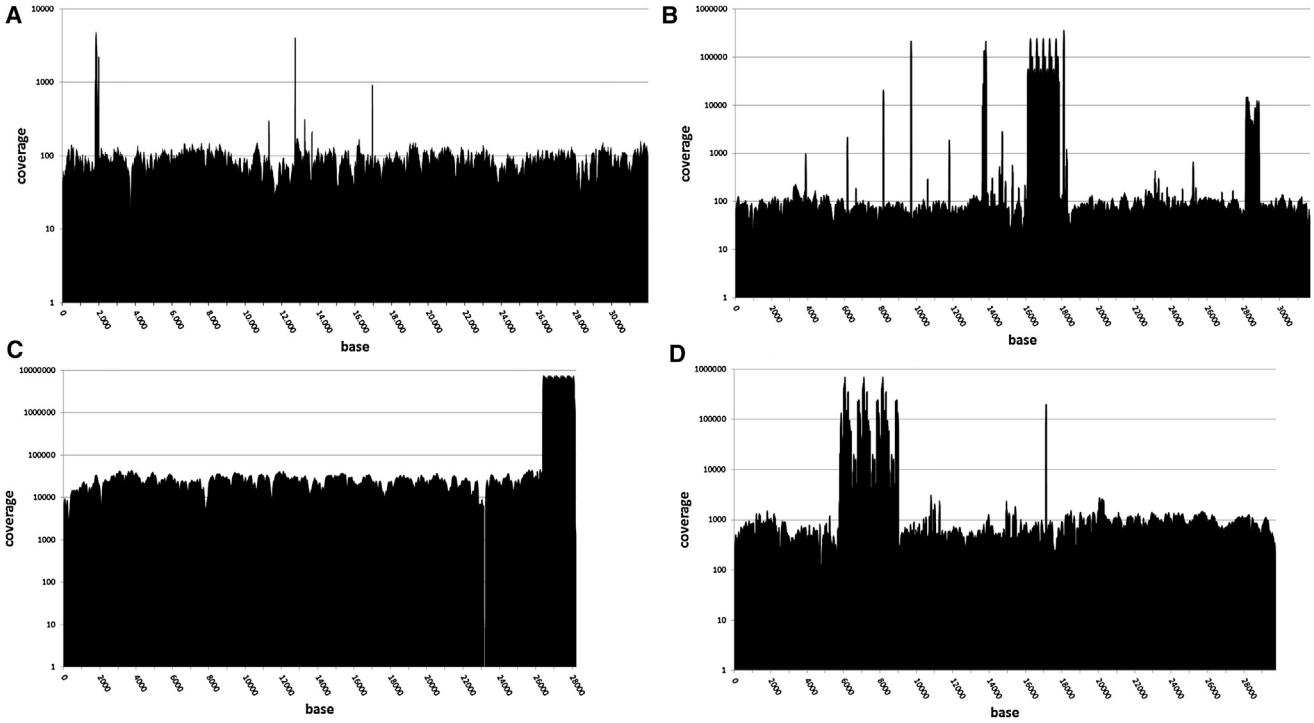

Figure 2. Coverage of Fosmid Clones with Illumina Sequencing Raw Reads at Base Resolution

(A and B) Fosmids Rf17-F-a-01b12 and Rf13-F-a-02a02, respectively, with a mean coverage of 72 and a few repeated segments.

(C) A fosmid with the highest observed overall coverage and long-range similarities to another fosmid (FSTFLPJ01C7H2U).

(D) A fosmid (Rf7-F-a-02e07) consisting almost entirely of simple repeats.

domains found are likely encoded by the *Rickettsia*-like bacterium, since they reside on contigs with a lower coverage than the *R. filosa* genome, and some have *Rickettsia*-like genes as neighbors.

Gene Repertoires and Potential Pseudogenes

We found a complete set of tRNA genes (Table 1). Gene prediction in fragmented genomic assemblies often yields higher numbers of protein-coding DNA sequence (CDS) predictions than in complete genome sequences. Indeed, we predicted 40,433 CDS in this assembly (Table 1). 6,256 of the predicted protein coding genes are supported by our transcriptomic data, but transcriptional activity is not correlated with the prediction score (Figure S1). For example, some high-scoring genes coding for potential flagellar apparatus and meiosis proteins have no transcript support (see below). Since we could not discern between pseudogenes, false-positive predicted genes, and true genes without functional analysis, we used the whole set of 40,433 predicted genes for the further analyses. Of these, 14,151 contained domains identifiable by similarity to the Pfam database [22]. The most prominent domain was the WD40 repeat domain (PF00400), which functions in coordinating multiprotein complex assemblies (Table S3). The high AT content of the genome with 65% is reflected in the coding sequence (62.9%) by runs of A-rich codons encoding stretches of asparagine and lysine like in other A/T-rich genomes [23].

Amplification of specific gene families may contribute to the success of a species in its specific habitat [24]. The OrthoMCL software package groups proteins into families according to their similarity [25]. In this way, we identified 4,368 families with two or more members. Many of the largest families cannot be assigned to a known function. Some of these families might be derived from false-positive gene predictions in repetitive elements, but even in these families with undefined functions, we found transcribed members. Yet, among the twenty largest families, there are five families associated with signaling and the cytoskeleton, emphasizing a prominent role of these functions for *R. filosa*.

Furthermore, we analyzed the functional capabilities of *R. filosa* in comparison to other eukaryote lineages. *B. natans*, a chlorarachniophyte alga (Cercozoa) within the Rhizaria, is currently the most closely related species with a genome sequence available (http://genome.jgi.doe.gov/Bigna1/Bigna1.home.html). As an outgroup, we used *Naegleria gruberi*, a free-living heterotrophic species from the supergroup Excavata, for the global analysis of common functional capabilities in unicellular species [26]. For the detailed analysis of genes and gene families, we also compared the *R. filosa* gene set to members of all other supergroups.

Signaling Components, Phagocytosis, and Adhesion

The *R. filosa* G protein-coupled receptors (GPCRs) are highly divergent from those of other organisms, and this diversity is also reflected by a high divergence within families. *R. filosa* encodes only 32 seven-transmembrane-domain receptors (7TMDRs). On the other hand, the genome encodes an overwhelming number of G protein α and β subunits, 154 and 32 respectively, which could account for differentiated responses to the environment (Table S4).

A comparison to *B. natans* and *N. gruberi* revealed that, on the whole, most signaling components are equally represented in *R. filosa* and *B. natans*, and the comparison to *N. gruberi* shows that Rhizaria have their fair share of these components inherited from the last common ancestor (LCA) of all eukaryotes. The same is true for components involved

Table 1. Predicted Genes in the *R. filosa* Genome

		Number	Mean Length (Bases)	Per Gene
Protein-coding genes		40,443	957	
	exons	101,686	335	2.51
	introns	61,253	85	1.51
	single-exon genes	12,305	690	
tRNA genes	complete	41		
	pseudogenes	6		

See also Table S1 and Figure S1.

in adhesion and phagocytosis. Yet, some signaling domains are underrepresented or missing in *R. filosa*. The X and Y domains of phospholipase C (PF00387 and PF00388, respectively), the BLUF (PF04940), and the nitrate/nitrite-sensing protein (PF08376) could not be identified, and the IQ motif (PF00612) seems to be underrepresented (Table S4).

Molecular Motors and Cytoskeleton

R. filosa attracted the attention of cell biologists several years ago because it exhibited one of the fastest movements of particles within a cell observed so far. Transport can be readily observed in its reticulate pseudopod net, where food, organelles, and other particles can be identified [9, 27, 28]. We wanted to know whether this ability for rapid transport is correlated to features of the molecular motors, which are encoded in its genome. We therefore compiled a comprehensive overview by thoroughly analyzing the genes coding for motor proteins. Initially, we screened the predicted genes for the presence of motor domains and reconstructed the underlying genes using the gene predictions and mRNA information. All three kinds of motor-domain-containing proteins are present (Table 2). Myosin and dynein gene families have nearly the same number of members as in other eukaryotes. We found ten different dynein heavy chains, which can be positioned in the respective proposed nine categories [29] by phylogenetic reconstruction (Figure S2A). Only IAD group 3 has two members. The rest of the predicted domains (seven) have low signals and/or are incomplete, so we assume them to be pseudogenes. We found 11 myosin genes, of which one is presumably a pseudogene. This number of myosins is comparable to that of other organisms [30]. A phylogenetic analysis revealed that nine of them group into two clusters of four and five members, respectively. However, the search for kinesin domains yielded a very high number of potential coding genes (Table 2; Table S5). Further analyses revealed that half of the total of 86 genes might be pseudogenes. This is underlined by the fact that we could not find transcriptional activity for these genes in our data set. Whether this is due to pseudogenization, inactivity in a certain lifestyle, or undersampling of transcriptome data remains elusive. Moreover, two are clearly derived from a reverse transposition event, since they contain no introns and are associated with the reverse transcriptase domain of a transposon. We reconstructed a phylogeny with the set of domains at least 200 amino acids long together with a representative kinesin set (http://www.cellbio.duke.edu/kinesin/). Many domains clustered with the known kinesin families, but we noticed also two large *R. filosa*-specific families. Taken together, the kinesins appear to have undergone frequent pseudogenization events, and *R. filosa* has an extended set of kinesins compared to other organisms [31].

R. filosa possesses an actin cytoskeleton with an amplified actin family (Table 2; Figure S2B). In other organisms, actin copies can be highly identical [32]. Due to assembly limitations, we cannot exclude that such identical actin gene copies exist in the *R. filosa* genome. Interestingly, *R. filosa* and other Foraminifera are uncommon, as their detected actin family members are much more diverse than in other species. This might hint at an "amplification with degradation" mechanism consistent with our observation of the high repetitiveness of the genome. Other common components of the actin cytoskeleton are also present (Table 2).

Flagellar and Meiosis Genes

The *R. filosa* life cycle is presently unknown. Years of culture in several laboratories, including a 12-year period by Nauss [10], revealed nothing but the plasmodial form. Therefore, it came as a surprise that we found a large set of genes coding for flagellar proteins in the *R. filosa* genome (Table S6). A study that led to an initial definition of required genes for flagellar and amoeboid locomotion was performed with *N. gruberi*. Of the 156 proteins described in this study as being associated with eukaryote flagella [26], 82 have clearly identifiable orthologs in *R. filosa*. A further ten are likely orthologs with less similarity. Many of the detected orthologs have functions only in flagella. Thus, the basic requirements to produce a flagellar apparatus are smaller than estimated in the previous study [26], the detected orthologs are only remnants of a life stage with a functional flagella, or other proteins fulfill analogous functions in *R. filosa*. Furthermore, some genes encoding meiosis-related proteins are also present. For other Foraminifera, it was shown that they reproduce sexually via flagellated gametes [33]. Thus, *R. filosa* possibly has the capability to generate flagellated cells like other species with multinucleated life stages [34, 35]. Most of the detected genes are not transcribed (Table S6), which is in accordance with the vegetative life stage we examined.

Transcription Factors

Sophisticated transcriptional regulation and its changes contribute significantly to speciation and establishment of novel evolutionary lineages [36]. We screened the *R. filosa* protein set for proteins of the transcriptional machinery and for specific transcription factors (TFs) for TF domain-containing proteins using InterProScan [37]. As expected, the complete basic eukaryote transcriptional machinery is present, but TFs are scarce, with only 43 members. The most abundant TFs are jumonji proteins comprising more than half of all TFs (23), followed by homeobox and cold-shock domain proteins. The scarcity of TFs could be due to limits in detectability but could also hint at a limited regulatory repertoire.

Comparison to Other Eukaryotes

Rhizaria comprise such diverse organisms as cercozoans, radiolarians, foraminiferans, gromiids, and the parasitic plasmodiophorids and haplosporidians. Most of these groups are not represented by cultivated species. Thus, phylogenetic reconstructions within this group profit from extensive EST data sets [38]. We compared available EST data to the *R. filosa* protein set. Less than half of each EST set has a counterpart in *R. filosa*, the nonmatching ESTs being either noncoding parts or species specific transcripts. This finding emphasizes the genomic diversity within the Rhizaria.

The current view of the phylogeny holds that Rhizaria form a supergroup with Stramenopiles and Alveolates, the SAR group. In the supplement, we show the position of *R. filosa* within the SAR supergroup (Figure S3). The Cercozoa and

Table 2. Motor Proteins and Actin Cytoskeleton

Gene Family	Members	Organization	Pseudogenes
Motors			
myosin	10	2 cluster (4 and 5 members); 1 Misato	1
dynein	10	1 member in each family; second member in IAD 3/4	7
kinesin	49 (41)	10 RF-specific cluster 1; 11 orphan cluster 1; 6 kinesin-14; 2 kinesin-3; 1 kinesin-13; 3 orphan cluster 2; 3 RF-specific cluster 2; 2 orphan cluster 3; 3 orphan cluster 4; 2 CENP-E; 2 kinesin-2; 4 kinesin-5	37 (45)
Actin Cytoskeleton			
actin	5		ND
actin-related	13		22; many short additional hits
Cap1/2	1		16
fimbrin	4		–
formin	potentially 5 genes with FH2 domains		>15
Ste20-like kinase	1		–
tubulin	>10		>30

Pseudogene numbers are only estimates based on partial similarities. Kinesin numbers in parentheses are the likely true values for kinesin genes and pseudogenes. Family assignments were made based on the kinesin domain only. Actin and actin-related pseudogenes are not discernable (ND) and therefore are listed only under actin-related pseudogenes. See also Tables S5 and S6 and Figure S2.

Foraminifera, to which belong the two rhizarian species *B. natans* and *R. filosa* with deciphered genomes, are separated by a long evolutionary distance. A comparison of the genomic features of the two species shows that *R. filosa* has an extended gene repertoire mainly due to gene family expansions (Table 3). Thus, a comparison of gene families rather than genes would yield the common "inventions" of this whole lineage. We found that 244 gene families evolved in the Rhizaria, which are stably inherited and might be the founding toolkit for this lineage (Figure 3). Interestingly, members of some of these families contain known domains, e.g., for cyclic nucleotide binding and hydrolyzing enzymes, which are likely involved in signaling cascades (Table S7).

Horizontal Gene Transfer

Horizontal gene transfer (HGT) plays an important role in the evolution of eukaryote species and clades [40]. We performed a global BLAST analysis (threshold p 10^{-10}) followed by phylogenetic tree reconstructions to analyze which proteins from *R. filosa* and *B. natans* are affiliated with specific supergroups (Figure 4). Only a small number of proteins appear to be shared between the two Rhizaria and other supergroups and bacteria, while most of potential HGT genes are not orthologous between *R. filosa* and *B. natans*. Due to the low similarity threshold chosen, this analysis provides the upper limit of HGT. To define the lower limits, we analyzed some aspects of HGT in the Rhizaria in more detail.

Prokaryote-to-Eukaryote Transfer

The transfer of genes from prokaryotes to eukaryotes, especially recent events, can be readily detected. We manually inspected the more than 1,000 potential bacterial HGTs defined by the automated analysis and used stronger selection criteria. In total, we found 17 well-supported genes shared between Rhizaria and four genes that are present only in *R. filosa* (Table S8) and were likely introduced by HGT. Due to the strict search criteria, this is likely the lower threshold of the total number of transferred bacterial genes.

Eukaryote-to-Eukaryote Transfer

A definite eukaryote-to-eukaryote HGT is only detectable if a branch-specific invention was affected or an HGT event makes a gene tree incongruent with the species tree. This ensures that the HGT is discernible from vertical transmission over long evolutionary distances. We estimated from the automated analysis and the manual inspection of bacterial HGTs that no more than 150 eukaryote genes have been transferred to or from *R. filosa*. We noted one potential early HGT event between the eukaryote supergroups Opisthokonta and Rhizaria. All currently known genomes of Opisthokonta besides Ecdysozoa and both Rhizaria, but not the whole SAR group, encode a Churchill domain protein (Figure S4). Using pattern and PSI blast searches, we detected that the only other species outside Opisthokonta and Rhizaria having this domain encoded in its genome is *Acanthamoeba castellanii*, which has a pronounced history of HGT [52].

Photosynthesis in the Last Common Ancestor of Foraminifera?

It was speculated that early in evolution, Rhizaria obtained a photosynthetic eukaryote, which was lost differentially in different lineages [53]. Acquiring the photosynthetic machinery involves transfer of genes of the endosymbiont to the nucleus. These genes are often associated with the photosynthetic activity [54]. All 45 proteins found this way could also be found in other cellular compartments such as mitochondria (Table S9; Supplemental Experimental Procedures). Thus, they may not be strictly related to the chloroplasts maintenance or activity, which is congruent with previous transcriptomic analysis of other foraminiferan species [55].

Discussion

The genome of *R. filosa* provides insights in a so-far neglected eukaryote crown group. Foraminifera are generally difficult to cultivate, and therefore genomic material is not easy to obtain. Thus, *R. filosa*, due to its cultivability, gave us a unique possibility to study a genome of this lineage and will serve as the reference genome for this entire branch of eukaryote evolution. Removal of DNA from culture contaminants, which we achieved by nuclei isolation, is crucial for subsequent analyses. This foraminiferal genome boasts short repeats, pseudogenization, and vast gene family expansions. The presence of long-range amplifications and tandem arrays makes a complete assembly of

Table 3. Comparison of Features of the Two Available Rhizaria Genomes

Feature	*R. filosa*	*B. natans*
Size (Mb)	~320	~100
Assembled size (Mb)	103	95
G/C content (%)	35	45
Predicted CDS	40,433	21,708
Unique (without counterparts in other species)	29,352	13,722
Shared between Rhizaria (genes)	7,994	4,215

See also Tables S2, S3, and S4.

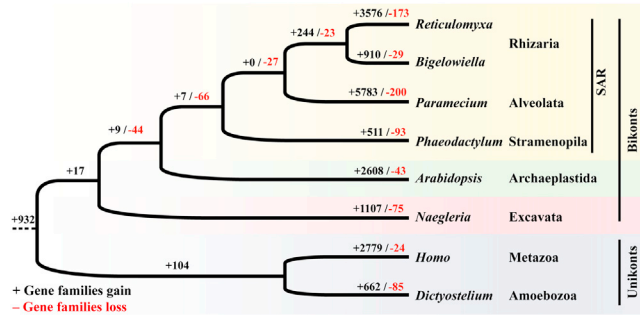

Figure 3. Gene Family Gains and Losses in Members of Major Eukaryote Lineages

The tree is schematically drawn based on the tree in Burki et al. [39] and Figure S3. Gains are calculated as gene families shared in the respective branch or trunk to the exclusion of others. Losses are calculated in respect to the origin (asterisk). Thus, if for example Paramecium is deleted from the tree, 200 additional gene families must be added to the 932 gene families at the asterisk. See also Figure S4 and Table S7.

the genome impossible. Gene duplications followed by pseudogenization could be associated with gene activity and/or functions since we noted that in some gene families (kinesins, actins) this is more pronounced than in others. The *R. filosa* genome analysis revealed several interesting aspects that foster our understanding of eukaryote genome evolution. Presumably, the LCA of eukaryotes possessed a flagellum, but the *R. filosa* protein set suggests that a smaller set than previously thought might be sufficient for this function. Our analysis also indicates that inclusion of more genomes from all eukaryote branches is required to properly define gene sets of the LCA. Several gene families are shared between the distantly related *B. natans* and *R. filosa*. This common gene set of presumably all Rhizaria is enriched with functions involved in signal transduction and regulation, indicating the importance of these functions for the establishment of new lineages.

Our detailed analysis of bacterial HGT revealed that these events are relatively rare in Rhizaria and that most occurred either in the LCA of Rhizaria or later in parallel in the examined species. Our finding of a common domain in two well-separated eukaryote branches points to a HGT event early in eukaryote evolution, since both known rhizarian genomes are separated by approximately, or more than, one billion years. Independent or sequential take-up would be another possibility to explain our finding. Given the unclear relationships of eukaryote supergroups, the possibility exists that this protein was invented early by the LCA of Rhizaria and Opisthokonta and was lost in some branches.

The search for genes required for photosynthesis showed that all photosynthetic organisms in this analysis (*B. natans*, *A. thaliana*, and *P. tricornutum*) share 319 gene families not found in nonphotosynthetic organisms, including *R. filosa*. This is a comparably high number for otherwise unrelated organism, as the complete analysis of shared and lost families showed (Figure 3). Since we could not detect traces of photosynthetic genes acquired via endosymbiontic gene transfer, we conclude that the common ancestor of Rhizaria was heterotrophic.

Experimental Procedures

The *R. filosa* clone whose nuclear genome we sequenced was originally isolated in 1993 and cultivated in commercial table water containing wheat germ flakes as food. The culture should be clonal, since only one plasmodium was used to set it up. The original culture contained common freshwater bacterial species, ciliates, and other undefined protozoa. We minimized the number of these contaminants by incubation with antibiotics and washing steps (Supplemental Experimental Procedures). The cells were harvested, and ~10 million nuclei per preparation for sequencing were isolated using a standard yeast protocol [17]. The nuclei were broken up, and the DNA was converted to sequencing libraries for Illumina and Roche/454 next-generation sequencing machines and sequenced on the respective machines. A fosmid library was constructed using the pSMART FOS fosmid cloning kit from Lucigen, and fosmid ends were sequenced on an ABI3700 sequencing machine. RNA was isolated from whole cells of the same culture with the QIAGEN RNA isolation kit and converted to cDNA using the Evrogen Mint 2cDNA kit. This cDNA was converted to a Roche/454 sequencing library and sequenced on a 454 FLX sequencing machine.

Assembly of the sequences was performed using ABySS (http://www.bcgsc.ca/platform/bioinfo/software/abyss/), CLC (http://www.clcbio.com/), and Newbler (http://www.454.com/) software. The outcome of different methods is listed in Table S1. For all further analyses, we used the assembly termed "merged assembly" in Table S1. The isolation of nuclei successfully discriminated against nonnuclear DNA in the preparation (Supplemental Experimental Procedures). Completeness of the assembly was assessed using transcript data and by the analysis of genes and gene families.

Gene prediction was performed using a trained version of Geneid [56]. Training of the prediction algorithm was conducted with available EST and transcriptome data. In this way, we predicted 40,433 protein-coding genes with a score value $\geq$ 20.

Accession Numbers

All sequence data can be accessed via http://genome.imb-jena.de/reti_blast/BlastReti.cgi. The whole-genome shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession ASPP00000000; the version described in this paper is version ASPP01000000. Further additional information on the genome analysis can be found at http://www.uni-koeln.de/med-fak/biochemie/biomed1/filosa/03_supplement.shtml.

Supplemental Information

Supplemental Information includes four figures, nine tables, and Supplemental Experimental Prodecures and can be found with this article online at http://dx.doi.org/10.1016/j.cub.2013.11.027.

Author Contributions

G.G. and M. Schliwa designed the study. N.H. and U.E. provided material and expert knowledge on the organism. M. Schleicher, A.A.N., L.E., C.G., J.P., R.S., U.E., L.P., A.M., M.P., M.G., K.S., M. Schliwa, and G.G. analyzed the data. M. Schliwa, J.P., and G.G. wrote the manuscript. All authors read the manuscript and provided comments and suggestions for improvements.

Acknowledgments

This work was supported by the BMBF Excellence Initiative to M. Schliwa; Swiss National Science Foundation grant 31003A-140766 to J.P., R.S., and L.P.; and DFG grants SFB670 to A.A.N. and L.E. and SFB914 to M. Schleicher. The authors are thankful to José Fahrni for the photo of the *Reticulomyxa filosa* cell and to Franz Schwarz for expert assistance with lab work.

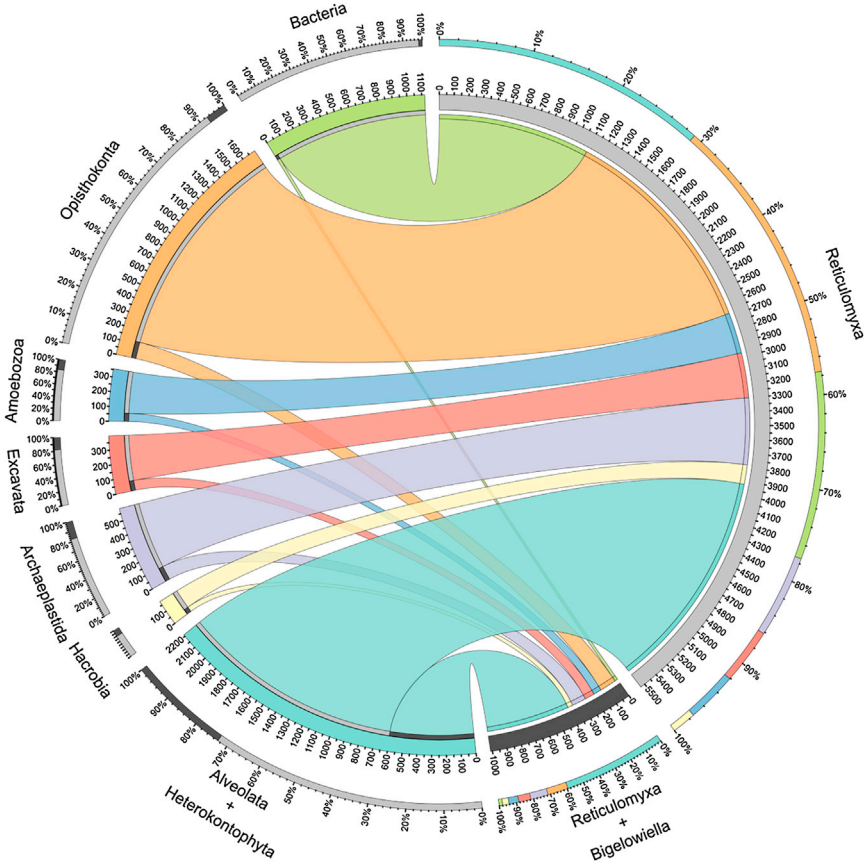

Figure 4. Phylogenomic Classification of *Reticulomyxa* Proteins

The *R. filosa* predicted protein models were phylogenomically analyzed as described previously [41, 42]. The phylogenomic reference database included RefSeq [43] (release 55) and additional sequence data for unicellular eukaryotes from different repositories, e.g., NCBI nr, JGI [44], and dbEST [45], with a total of about 24 million protein models from more than 300,000 species and strains. Multiple sequence alignments were computed using multiple sequence comparison by log expectation (MUSCLE) [46]. The generated alignments were tuned using trimAl [47] with the "gappyout" option. The approximate-maximum-likelihood phylogenies were inferred from the trimmed alignments with the WAG model [48] of amino acid evolution under the gamma model with 20 rate categories using FastTree [49]. The reliability of each split in a tree was estimated using the Shimodaira-Hasegawa (SH) test [50] with 1,000 resampling replicates. The trees were classified into the different topological categories with bootstrap support $\geq$ 70% using PhyloSort [51]. Inner tracks represent the *R. filosa* genome and its phylogenetic sister lineages. The numbers on the inner tracks are the shared proteins. The outer tracks are the percentages of the contributions. The widths of the connecting ribbons are proportional to the number of shared proteins. *R. filosa* proteins shared with *B. natans* are depicted separately to emphasize common affiliations of Rhizaria. See also Table S8 and Table S9.

Received: June 25, 2013
Revised: October 9, 2013
Accepted: November 12, 2013
Published: December 12, 2013

References

1. Adl, S.M., Simpson, A.G., Lane, C.E., Lukeš, J., Bass, D., Bowser, S.S., Brown, M.W., Burki, F., Dunthorn, M., Hampl, V., et al. (2012). The revised classification of eukaryotes. J. Eukaryot. Microbiol. *59*, 429–493.
2. Nikolaev, S.I., Berney, C., Fahrni, J.F., Bolivar, I., Polet, S., Mylnikov, A.P., Aleshin, V.V., Petrov, N.B., and Pawlowski, J. (2004). The twilight of Heliozoa and rise of Rhizaria, an emerging supergroup of amoeboid eukaryotes. Proc. Natl. Acad. Sci. USA *101*, 8066–8071.
3. Curtis, B.A., Tanifuji, G., Burki, F., Gruber, A., Irimia, M., Maruyama, S., Arias, M.C., Ball, S.G., Gile, G.H., Hirakawa, Y., et al. (2012). Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. Nature *492*, 59–65.
4. Murray, J. (2006). Ecology and Applications of Benthic Foraminifera (Cambridge: Cambridge University Press).
5. Travis, J.F., and Bowser, S.S. (1991). The motility of Foraminifera. In Biology of Foraminifera, J.J. Lee and R.O. Anderson, eds. (London: Academic Press), pp. 91–156.
6. Habura, A., Wegener, L., Travis, J.L., and Bowser, S.S. (2005). Structural and functional implications of an unusual foraminiferal beta-tubulin. Mol. Biol. Evol. *22*, 2000–2009.
7. Bowser, S.S., Habura, A., and Pawlowski, J. (2006). Molecular evolution of Foraminifera. In Genomics and Evolution of Microbial Eukaryotes, L.A. Katz and D. Bhattacharya, eds. (New York: Oxford University Press), pp. 78–93.
8. Hülsmann, N. (1984). Biology of the genus Reticulomyxa (Rhizopoda). J. Protozool. *34*, 55a.
9. Koonce, M.P., Euteneuer, U., McDonald, K.L., Menzel, D., and Schliwa, M. (1986). Cytoskeletal architecture and motility in a giant freshwater amoeba, Reticulomyxa. Cell Motil. Cytoskeleton *6*, 521–533.
10. Nauss, R.N. (1949). *Reticulomyxa filosa* gen. et spec. nov., a new primitive plasmodium. Bull. Torrey Bot. Club *76*, 161–173.
11. Hülsmann, N. (2006). Bewegung, Nahrungsaufnahme und Fortpflanzung bei *Reticulomyxa filosa* (Rhizopoda), Begleitpublikation zum Film C 1639 (Göttingen: IWF Wissen und Medien gGmbH). http://www.filmarchives-online.eu.
12. Ashkin, A., Schütze, K., Dziedzic, J.M., Euteneuer, U., and Schliwa, M. (1990). Force generation of organelle transport measured in vivo by an infrared laser trap. Nature *348*, 346–348.
13. Euteneuer, U., Koonce, M.P., Pfister, K.K., and Schliwa, M. (1988). An ATPase with properties expected for the organelle motor of the giant amoeba, Reticulomyxa. Nature *332*, 176–178.
14. Koonce, M.P., Tong, J., Euteneuer, U., and Schliwa, M. (1987). Active sliding between cytoplasmic microtubules. Nature *328*, 737–739.
15. Parfrey, L.W., and Katz, L.A. (2010). Genome dynamics are influenced by food source in Allogromia laticollaris strain CSH (Foraminifera). Genome Biol. Evol. *2*, 678–685.
16. Burki, F., Shalchian-Tabrizi, K., and Pawlowski, J. (2008). Phylogenomics reveals a new 'megagroup' including most photosynthetic eukaryotes. Biol. Lett. *4*, 366–369.
17. Aris, J.P., and Blobel, G. (1991). Isolation of yeast nuclei. Methods Enzymol. *194*, 735–749.
18. Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics *23*, 1061–1067.
19. Aoki-Kinoshita, K.F., and Kanehisa, M. (2007). Gene annotation and pathway mapping in KEGG. Methods Mol. Biol. *396*, 71–91.
20. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C., and Kanehisa, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. *35* (Web Server issue), W182–W185.

21. Burgdorfer, W., Anacker, R.L., Bird, R.G., and Bertram, D.S. (1968). Intranuclear growth of Rickettsia rickettsii. J. Bacteriol. *96*, 1415–1418.
22. Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., et al. (2012). The Pfam protein families database. Nucleic Acids Res. *40* (Database issue), D290–D301.
23. Szafranski, K., Lehmann, R., Parra, G., Guigo, R., and Glöckner, G. (2005). Gene organization features in A/T-rich organisms. J. Mol. Evol. *60*, 90–98.
24. Heidel, A.J., Lawal, H.M., Felder, M., Schilde, C., Helps, N.R., Tunggal, B., Rivero, F., John, U., Schleicher, M., Eichinger, L., et al. (2011). Phylogeny-wide analysis of social amoeba genomes highlights ancient origins for complex intercellular communication. Genome Res. *21*, 1882–1891.
25. Li, L., Stoeckert, C.J., Jr., and Roos, D.S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. *13*, 2178–2189.
26. Fritz-Laylin, L.K., Prochnik, S.E., Ginger, M.L., Dacks, J.B., Carpenter, M.L., Field, M.C., Kuo, A., Paredez, A., Chapman, J., Pham, J., et al. (2010). The genome of Naegleria gruberi illuminates early eukaryotic versatility. Cell *140*, 631–642.
27. Orokos, D.D., and Travis, J.L. (1997). Cell surface and organelle transport share the same enzymatic properties in Reticulomyxa. Cell Motil. Cytoskeleton *38*, 270–277.
28. Schliwa, M., Shimizu, T., Vale, R.D., and Euteneuer, U. (1991). Nucleotide specificities of anterograde and retrograde organelle transport in Reticulomyxa are indistinguishable. J. Cell Biol. *112*, 1199–1203.
29. Wickstead, B., and Gull, K. (2007). Dyneins across eukaryotes: a comparative genomic analysis. Traffic *8*, 1708–1721.
30. Kollmar, M. (2006). Thirteen is enough: the myosins of Dictyostelium discoideum and their light chains. BMC Genomics *7*, 183.
31. Kollmar, M., and Glöckner, G. (2003). Identification and phylogenetic analysis of Dictyostelium discoideum kinesin proteins. BMC Genomics *4*, 47.
32. Joseph, J.M., Fey, P., Ramalingam, N., Liu, X.I., Rohlfs, M., Noegel, A.A., Müller-Taubenberger, A., Glöckner, G., and Schleicher, M. (2008). The actinome of Dictyostelium discoideum in comparison to actins and actin-related proteins from other organisms. PLoS ONE *3*, e2654.
33. Myers, E.M. (1940). Observations on the origin and fate of flagellated gametes in multiple tests of discorbis (Foraminifera). J. Mar. Biol. Assoc. U. K. *24*, 201–225.
34. Barrantes, I., Glöckner, G., Meyer, S., and Marwan, W. (2010). Transcriptomic changes arising during light-induced sporulation in Physarum polycephalum. BMC Genomics *11*, 115.
35. Glöckner, G., Golderer, G., Werner-Felmayer, G., Meyer, S., and Marwan, W. (2008). A first glimpse at the transcriptome of Physarum polycephalum. BMC Genomics *9*, 6.
36. Lorenz, P., Dietmann, S., Wilhelm, T., Koczan, D., Autran, S., Gad, S., Wen, G., Ding, G., Li, Y., Rousseau-Merck, M.F., and Thiesen, H.J. (2010). The ancient mammalian KRAB zinc finger gene cluster on human chromosome 8q24.3 illustrates principles of C2H2 zinc finger evolution associated with unique expression profiles in human tissues. BMC Genomics *11*, 206.
37. Zdobnov, E.M., and Apweiler, R. (2001). InterProScan—an integration platform for the signature-recognition methods in InterPro. Bioinformatics *17*, 847–848.
38. Burki, F., Kudryavtsev, A., Matz, M.V., Aglyamova, G.V., Bulman, S., Fiers, M., Keeling, P.J., and Pawlowski, J. (2010). Evolution of Rhizaria: new insights from phylogenomic analysis of uncultivated protists. BMC Evol. Biol. *10*, 377.
39. Burki, F., Okamoto, N., Pombert, J.F., and Keeling, P.J. (2012). The evolutionary history of haptophytes and cryptophytes: phylogenomic evidence for separate origins. Proc. Biol. Sci. *279*, 2246–2254.
40. Keeling, P.J., and Palmer, J.D. (2008). Horizontal gene transfer in eukaryotic evolution. Nat. Rev. Genet. *9*, 605–618.
41. Moustafa, A., Beszteri, B., Maier, U.G., Bowler, C., Valentin, K., and Bhattacharya, D. (2009). Genomic footprints of a cryptic plastid endosymbiosis in diatoms. Science *324*, 1724–1726.
42. Moustafa, A., Bhattacharya, D., and Allen, A.E. (2010). iTree: a high-throughput phylogenomic pipeline. Proceedings of the Biomedical Engineering Conference (CIBEC), 2010 5th Cairo International, 103–107.
43. Pruitt, K.D., Tatusova, T., and Maglott, D.R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. *35* (Database issue), D61–D65.
44. Grigoriev, I.V., Nordberg, H., Shabalov, I., Aerts, A., Cantor, M., Goodstein, D., Kuo, A., Minovitsky, S., Nikitin, R., Ohm, R.A., et al. (2012). The genome portal of the Department of Energy Joint Genome Institute. Nucleic Acids Res. *40* (Database issue), D26–D32.
45. Boguski, M.S., Lowe, T.M., and Tolstoshev, C.M. (1993). dbEST—database for "expressed sequence tags". Nat. Genet. *4*, 332–333.
46. Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797.
47. Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*, 1972–1973.
48. Whelan, S., and Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. *18*, 691–699.
49. Price, M.N., Dehal, P.S., and Arkin, A.P. (2009). FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol. Biol. Evol. *26*, 1641–1650.
50. Shimodaira, H., and Hasegawa, M. (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. *16*, 1114.
51. Moustafa, A., and Bhattacharya, D. (2008). PhyloSort: a user-friendly phylogenetic sorting tool and its application to estimating the cyanobacterial contribution to the nuclear genome of Chlamydomonas. BMC Evol. Biol. *8*, 6.
52. Clarke, M., Lohan, A.J., Liu, B., Lagkouvardos, I., Roy, S., Zafar, N., Bertelli, C., Schilde, C., Kianianmomeni, A., Bürglin, T.R., et al. (2013). Genome of Acanthamoeba castellanii highlights extensive lateral gene transfer and early evolution of tyrosine kinase signaling. Genome Biol. *14*, R11.
53. Keeling, P.J. (2010). The endosymbiotic origin, diversification and fate of plastids. Philos. Trans. R. Soc. Lond. B Biol. Sci. *365*, 729–748.
54. Nowack, E.C., Vogel, H., Groth, M., Grossman, A.R., Melkonian, M., and Glöckner, G. (2011). Endosymbiotic gene transfer and transcriptional regulation of transferred genes in Paulinella chromatophora. Mol. Biol. Evol. *28*, 407–422.
55. Pillet, L., and Pawlowski, J. (2013). Transcriptome analysis of foraminiferan *Elphidium margaritaceum* questions the role of gene transfer in kleptoplastidy. Mol. Biol. Evol. *30*, 66–69.
56. Blanco, E., Parra, G., and Guigo, R. (2007). Using Geneid to identify genes. Curr. Protoc. Bioinfomatics *4*, 4.3. http://dx.doi.org/10.1002/0471250953.bi0403s18.