Phylogenomics Places Orphan Protistan Lineages in a Novel Eukaryotic Super-Group

Matthew W. Brown1,2, Aaron A. Heiss3,4, Ryoma Kamikawa5, Yuji Inagaki6,7, Akinori Yabuki8, Alexander K. Tice1,2, Takashi Shiratori6, Ken-Ichiro Ishida6, Tetsuo Hashimoto6,7, Alastair G.B. Simpson3,* and Andrew J. Roger9,*

1Department of Biological Sciences, Mississippi State University, USA
2Institute for Genomics, Biocomputing & Biotechnology, Mississippi State University, USA
3Department of Biology, and Centre for Comparative Genomics and Evolutionary Bioinformatics, Dalhousie University, Halifax, Nova Scotia, Canada
4Department of Invertebrate Zoology and Sackler Institute for Comparative Genomics, American Museum of Natural History, New York, New York, USA
5Graduate School of Human and Environmental Studies, Graduate School of Global Environmental Studies, Kyoto University, Japan
6Graduate School of Life and Environmental Sciences, University of Tsukuba, Ibaraki, Japan
7Center for Computational Sciences, University of Tsukuba, Ibaraki, Japan
8Japan Agency for Marine-Earth Science and Technology (JAMSTEC), Yokosuka, Kanagawa, Japan
9Department of Biochemistry and Molecular Biology, and Centre for Comparative Genomics and Evolutionary Bioinformatics, Dalhousie University, Halifax, Nova Scotia, Canada

*Corresponding authors: E-mails: alastair.simpson@dal.ca; andrew.roger@dal.ca.
Accepted: January 18, 2018

Data deposition: All new transcriptomic data have been deposited at the National Center for Biotechnology Information under BioProjects PRJNA401035, as detailed in supplementary table S1, Supplementary Material online. All single gene alignments, masked and unmasked, and phylogenomic matrices are available in supplementary file, Supplementary Material online Brown et al.2017.CRuMs.tgz.

Abstract

Recent phylogenetic analyses position certain “orphan” protist lineages deep in the tree of eukaryotic life, but their exact placements are poorly resolved. We conducted phylogenomic analyses that incorporate deeply sequenced transcriptomes from representatives of collodictyonids (diphylleids), rigifilids, Mantamonas, and ancyromonads (planomonads). Analyses of 351 genes, using site-heterogeneous mixture models, strongly support a novel super-group-level clade that includes collodictyonids, rigifilids, and Mantamonas, which we name “CRuMs”. Further, they robustly place CRuMs as the closest branch to Amorphea (including animals and fungi). Ancyromonads are strongly inferred to be more distantly related to Amorphea than are CRuMs. They emerge either as sister to malawimonads, or as a separate deeper branch. CRuMs and ancyromonads represent two distinct major groups that branch deeply on the lineage that includes animals, near the most commonly inferred root of the eukaryote tree. This makes both groups crucial in examinations of the deepest-level history of extant eukaryotes.

Key words: eukaryote tree of life, concatenated phylogenetic analysis, protist, site-heterogeneous models.

Introduction

Our understanding of the eukaryote tree of life has been revolutionized by genomic and transcriptomic investigations of diverse protists, which constitute the overwhelming majority of eukaryotic diversity (Burki 2014; Simpson and Eglit 2016). Phylogenetic analyses of super-matrices of proteins typically show a eukaryote tree consisting of five-to-eight “super-groups” that fall within three even-higher-order assemblages: 1) Amorphea (Amoebozoa plus Obazoa, the latter including animals and fungi), 2) Diaphoretickes (primarily Sar, Archaeplastida, Cryptista, and Haptophyta), and 3) Excavata (Discoba...
Material online. Poorly aligned sites were excluded using described in the supplementary text, Supplementary Material online. Extensive efforts to exclude contamination and paralogs, as alignment of 97,002 amino acid (AA) sites with < 25% missing data for both 61- and 64-taxon data sets (supplementary table S2, Supplementary Material online). Results of experimental methods for culturing, nucleic acid extraction, and Illumina sequencing are described in the supplementary text, Supplementary Material online.

**Materials and Methods**

Details of experimental methods for culturing, nucleic acid extraction, and Illumina sequencing are described in the supplementary text, Supplementary Material online.

**Phylogenomic Data Set Construction**

A reference data set of 351 aligned proteins described in (Kang et al. 2017) was used as the starting point for the current analysis, from which 61 or 64 taxa representing diverse eukaryotes were selected (see supplementary table S2, Supplementary Material online). Extensive efforts were made to exclude contamination and paralogs, as described in the supplementary text, Supplementary Material online. Poorly aligned sites were excluded using BMGE (Criscuolo and Gribaldo 2010), resulting in an alignment of 97,002 amino acid (AA) sites with < 25% missing data for both 61- and 64-taxon data sets (supplementary table S2, Supplementary Material online).

**Phylogenomic Tree Inference**

Maximum likelihood (ML) trees were inferred using IQ-Tree v. 1.5.5 (Nguyen et al. 2015). The best-fitting available model based on the Akaike Information Criterion (AIC) was the LG + C60 + F + I model with class weights optimized from the data set and four discrete gamma (I) categories. ML trees were estimated under this model for both 61- and 64-taxon data sets. We then used this model and best ML tree under the LG + C60 + F + I model to estimate the “posterior mean site frequencies” (PMSF) model (Wang et al. 2017) for both 61 (fig. 1) and 64 (supplementary fig. S1, Supplementary Material online) taxon data sets. This LG + C60 + F + I - PMSF model was used to re-estimate ML trees, and for a bootstrap analysis of the 61-taxon data set, with 100 pseudoreplicates (fig. 1). AU topology tests under the LG + C60 + F + I were conducted with IQ-Tree to evaluate whether trees recovered by the Bayesian analyses or alternative placements (see supplementary table S1, Supplementary Material online, for hypotheses tested) of the orphan taxa could be rejected statistically.

Bayesian inferences were performed using PhyloBayes-MPI v1.6j (Rodrigue and Lartillot 2014), under the CAT-GTR+I model, with four discrete I categories. For the 61-taxon analysis, 6 independent Markov chain Monte Carlo chains were run for ~4,000 generations, sampling every second generation. Two sets of two chains converged at 800 and 2,000 generations, which were, respectively, used as the burnin, with the largest discrepancy in posterior probabilities (PPs) (maxdiff < 0.05. The topologies of the converged chains are presented in supplementary figures S3 and S4, Supplementary Material online, and are mapped upon figure 1. For the 64-taxon analysis, four chains were run for ~3,000 generations. Two chains converged at ~200 generations, which was used as the burnin (maxdiff = 0) and the posterior probabilities are mapped upon the ML tree in supplementary figure S1, Supplementary Material online.

**Fast-Site Removal and Gene Subsampling Analyses**

For fast site removal, rates of evolution at each site of the 61-taxon data set were estimated with DetE (Susko et al. 2003) under the LG model using discrete gamma probability estimation. A custom Python script was then used to remove fastest evolving sites in 4,000-site steps. Random subsampling of 20%, 40%, 60%, or 80% of the genes in the 61-taxon data set was conducted using a custom Python script, with the number of replicates as given in figure 2B. In both cases each step or subsample was analyzed using 1,000 UFBOOT replicates in IQ-Tree under the LG + C60 + F + I - PMSF model.

**Results**

Using a custom phylogenomic pipeline plus manual curation, we generated a data set of 351 orthologs. The data set was...
filtered of paralogs and potential cross-contamination by visualizing each protein’s phylogeny individually, then removing sequences whose positions conflicted with a conservative consensus phylogeny (as in Tice et al. 2016; Kang et al. 2017) (supplementary methods, Supplementary Material online). We selected data-rich species to represent the phylogenetic diversity of eukaryotes. Our primary data set retained 61 taxa, with metamonads represented by two short-branching taxa (Trimastix and Paratrimastix). We also analyzed a 64-taxon data set containing three additional longer branching metamonads. Maximum likelihood (ML) and Bayesian analyses were conducted using site-heterogeneous models; LG + C60 + F + C - PMSF as implemented in IQ-TREE (Wang et al. 2017) and CAT-GTR model in PHYLOBAYES-MPI, respectively.

Such site-heterogeneous models are important for deep-level phylogenetic inference with numerous substitutions along branches (Lartillot et al. 2007; Le et al. 2008; Wang et al. 2008, 2017; Pisani et al. 2015).

Our analyses of both 61- and 64-taxon data sets robustly recover well-accepted major groups including Sar, Discoba, Metamonada, Obazoaa, and Amoebozoa (fig. 1 and supplementary fig. S1, Supplementary Material online). Cryptista (e.g., cryptomonads and close relatives) branches with Haptophyta (fig. 1) in the LG + C60 + F + P - PSMF analyses as well as in one set of two converged PHYLOBAYES-MPI chains under the CAT-GTR model (supplementary fig. S2, Supplementary Material online). However another pair of converged chains places Haptophyta as sister to Sar while Cryptista nests within Archaeplastida (supplementary fig. S3, Supplementary Material online), which is largely consistent with some other recent phylogenomic studies (Burki et al. 2016). Excavata was never monophyletic, with Discoba forming a clan with Diaphoretickes taxa (Sar, Haptophyta, Archaeplastida + Cryptista) and Metamonada grouping with Amorphea plus the four orphan lineages targeted in this study (see below).
Malawimonads, which are morphologically similar to certain metamonads and discobids (Simpson 2003), also branch among the “orphans” (see below).

Phylogenies of both data sets place all four orphan taxa near the base of Amorphea (fig. 1 and supplementary fig. S1, Supplementary Material online). The uncertain position of the eukaryotic root (discussed earlier) therefore makes it unclear which bipartitions are truly clades, and which could be interrupted by the root. To allow efficient communication, we discuss the phylogenies as if the orphan taxa all lie on the Amorphea side of the root. We will also consider Amorphea as previously circumscribed (Adl et al. 2012): the least-inclusive clade or clan containing Amoebozoa and Opisthokonta.

Three of the orphan lineages are specifically related in our trees (fig. 1 and supplementary fig. S1, Supplementary Material online). In both 61- and 64-taxon analyses, Rigiﬁla ramosa (representing Rigiﬁla) forms a maximally supported clade with the collodictyonids Collodictyon triciliatum and Diphylliea rotans. Mantamonas plastica then branches as their closest relative, with maximal support. This Collodictyon + Rigiﬁla + Mantamonas clade (“CRuMs”) forms the sister group to Amorphea, again with maximal support.

ML analyses and the converged PhyloBayes chains grouped ancyromonads, malawimonads, and CRuMs with Amorphea, with strong bootstrap support and Bayesian posterior probability (fig. 1, 61 taxa; PMSF BS = 98%, PP = 1). Ancyromonads and malawimonads formed a clade in the ML analyses, but with equivocal support (fig. 1, 61 taxa; BS = 77%). Both sets of converged chains of the Bayesian analyses instead grouped malawimonads with CRuMs + Amorphea to the exclusion of ancyromonads (supplementary figs. S2 and S3, Supplementary Material online, PP = 1 for both); however some unconverted chains support...
an ancyromonad + malawimonad clade (data not shown). Lack of convergence among multiple chains using the CATERGTR+\( \Gamma \) model is unfortunately common for large data sets, and often cannot be resolved by increasing the number of generations of Markov chain Monte Carlo within a reasonable time frame (Pisani et al. 2015; Kang et al. 2017). Instead we treat the two topologies recovered in these analyses as candidate hypotheses requiring further investigation.

We conducted approximately unbiased (AU) topology tests on the 61-taxon data set under the LG + C60 + F + \( \Gamma \) mixture model (supplementary table S1, Supplementary Material online). These tests rejected the Phylobayes trees, as well as all trees optimized by enforcing constraints representing plausible alternative relative placements of ancyromonads, malawimonads, and metamonads.

The fastest evolving sites are expected to be the most prone to saturation and systematic error arising from model misspecification in phylogenomic analyses (Philippe et al. 2011). We conducted a “fast-site removal” analysis with the 61-taxon data set and generated ultrafast bootstrap support (UFBOOT) values (Minh et al. 2013) for relevant groups as sites were progressively removed from fastest to slowest (fig. 2A). All groups of interest receive reasonably strong support until \(~44,000–48,000\) sites were removed, when support fell markedly for the ancyromonad + malawimonad clade and the Amorphea + CRuMs + ancyromonad + malawimonad clan. At this point, a notable proportion of the bootstrap trees show malawimonads and/or ancyromonads grouping with metamonads. This decline in support for the ancyromonad + malawimonad group reverses somewhat with further site removal, before support falls again as overall phylogenetic structure is lost when \(~76,000\) sites are removed (fig. 2A).

To evaluate heterogeneity in phylogenetic signals among genes (Inagaki et al. 2009), we also inferred phylogenies from subsamples of the 351 examined genes (61-taxon data set; fig. 2B and C). For each subsample 20–80% of the genes were randomly selected, without replacement, with replication as per figure 2B (giving a >95% probability that a particular gene would be sampled at each level), and UFBOOT support for major clades was inferred (fig. 2C). The “80% retained” replicates gave nearly identical results to the full data set, indicating that there was little stochastic error associated with gene sampling at this level. Support for the CRuMs clade is almost always high when 40% of genes (Inagaki et al. 2009), we also inferred phylogenies from 10(2):427–433 doi:10.1093/gbe/evy014 Advance Access publication January 19, 2018 431

Discussion

Our 351 protein (97,002 AA site) super-matrix places several orphan lineages in two separate clades emerging between Amorphea and all other major eukaryote groups. All methods recover a strongly supported clade comprising the free-swimming collodictyonid flagellates, the idiosyncratic filose protist Rigiﬁla (Rigiﬁila), and the gliding flagellate Mantamonas. This clade is resilient to exclusion both of fast-evolving sites and of randomly selected genes. It is also consistently placed as the immediate sister taxon to Amorphea. This represents the first robust estimate of the positions of these three taxonomically poor but phylogenetically deep clades. Previous phylogenomic analyses placed collodictyonids in various positions, such as sister to either malawimonads or Amoebozoa, but often with low statistical support (Zhao et al. 2012; Cavalier-Smith et al. 2014). Placements of Mantamonas have varied dramatically. A recent phylogenomic study recovered a weak Mantamonas + collodictyonid clade in some analyses, but other analyses in the same study instead recovered a weak Mantamonas + ancyromonad relationship (Cavalier-Smith et al. 2014), and SSU + LSU rRNA gene phylogenies strongly grouped Mantamonas with apusomonads (Glücksman et al. 2011; Yabuki, Ishida, et al. 2013). Our study decisively supports the first of these possibilities. This is the first phylogenomic analysis incorporating Rigiﬁila: Previous SSU + LSU rRNA gene analyses recovered a negligibly supported collodictyonid + rigiﬁila clade, but not a relationship with Mantamonas (Yabuki, Ishida, et al. 2013).

Overall, the hypotheses that 1) collodictyonids, rigiﬁillas, and Mantamonas form a major eukaryote clade, and 2) this clade is sister to Amorphea, are novel, plausible, and evolutionarily important. No name exists for this putative super-group, and it is obviously premature to propose a formal taxon. We suggest the place-holding moniker “CRuMs” (Collodictyonidae, Rigiﬁila, Mantamonas), which is euphonic and evokes the species-poor nature of these taxa.
Whether ancyromonads branch outside Amorphea or within it has been disputed (Paps et al. 2013; Cavalier-Smith et al. 2014). Our study strongly places ancyromonads outside Amorphea, more distantly related to it than are the CRuMs. Ancyromonads instead fall "among" the excavate lineages (Discoba, Metamonada, and Malawimonadidae). Resolving the relationships among "excavates" is extremely challenging (Hampel et al. 2009; Derelle et al. 2015), and this likely contributed to our difficulty in resolving the exact position of ancyromonads vis-à-vis malawimonads. A close relationship between ancyromonads and some/all excavates would be broadly consonant with the marked cytoskeletal similarity between Ancyromonas and "typical excavates" (Heiss et al. 2011). Certainly, our study flags ancyromonads as highly relevant to resolving relationships among excavates.

Both candidate positions for ancyromonads place them at the center of a crucial open question: locating the root of the eukaryote tree. As discussed earlier, the latest analyses (Derelle et al. 2015) locate the root between Discoba + Diaphoretickes ("Diphoda") and a clade including Amorphea, collocytonids, and malawimonads ("Opimoda"). Our phylogenies show the ancyromonad lineage emerging close to this split. One of the two positions we recovered would actually place ancyromonads either as the deepest branch within "Diphoda," or the deepest branch within "Opimoda," or even as sister to all other extant eukaryotes. This demonstrates the profound importance of including ancyromonads in future rooted phylogenies of eukaryotes, using data sets optimized for this purpose.

**Supplementary Material**

**Supplementary data** are available at Genome Biology and Evolution online.

**Acknowledgments**

The authors thank Tom Cavalier-Smith and Ed Glücksmann (Oxford University) for supplying cultures strains B-70 (Ancyromonas sigmoides), NYK3C (Fabomonas tropica), and Bass1 (Mantamonas plastica). The part of this work conducted at Dalhousie University was supported by NSERC Discovery grants awarded to A.G.B.S. (298366-2014) and A.J.R. (2016-06792), respectively. A.J.R. also acknowledges the Canada Research Chairs program for support. This project was supported in part by the National Science Foundation (NSF) Division of Environmental Biology (DEB) grant 1456054 (http://www.nsf.gov), awarded to M.W.B. Mississippi State University’s High Performance Computing Collaboratory provided some computational resources. The part of this work conducted at the University of Tsukuba was supported by grants from the Japan Society for the Promotion of Science (JSPS; 15H05606 and 15K14591 awarded to R.K., 23117006 and 16H04826 awarded to Y.L., 15H04411 awarded to K.I., and 15H05231 to T.H.) and by the “Tree of Life” research project (University of Tsukuba).

**Literature Cited**

Adl SM, et al. 2012. The revised classification of eukaryotes. J Eukaryot Microbiol. 59(5):429–493.

Brugerolle G, Bricheux G, Philippe H, Coffea G. 2002. Collocyton nodulosa and Diphyleia rotans (= Aulacomonas submarina) form a new family of flagellates (Collocytonidae) with tubular mitochondrial cristae that is phylogenetically distant from other flagellate groups. Protist 153(1):59–70.

Burki F. 2014. The eukaryotic tree of life from a global phylogenomic perspective. Cold Spring Harb Perspect Biol. 6(5):a016147.

Burki F, et al. 2016. Untangling the early diversification of eukaryotes: a phylogenomic study of the evolutionary origins of Centrolehida, Haptophyta and Cryptista. Proc Biol Sci. 283:20152802.

Cavalier-Smith T. 2010. Kingdoms Protozoa and Chromista and the eozoan root of the eukaryotic tree. Biol Lett. 6(3):342–345.

Cavalier-Smith T. 2013. Early evolution of eukaryote feeding modes, cell structural diversity, and classification of the protozoan phyla likes, Sulcozoa, and Choanozoa. Eur J Protistol. 49(2):115–178.

Cavalier-Smith T, et al. 2014. Multigene eukaryote phylogeny reveals the likely protozoan ancestors of opisthokonts (animals, fungi, choanozoans) and Amoebozoa. Mol Phylogenet Evol. 81:71–85.

Criscuolo A, Grimaldi S. 2010. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol Biol. 10:210.

Derelle R, et al. 2015. Bacterial proteins pinpoint a single eukaryotic root. Proc Natl Acad Sci U S A. 112(7):E693–E699.

Feuda R, et al. 2017. Improved modeling of compositional heterogeneity supports sponges as sister to all other animals. Curr Biol. 27(24):3864–3870.

Glücksmann E, et al. 2011. The novel marine gliding zooflagellate genus Mantamonas (Mantamonadida ord. n.: Apusozoa). Protist 162(2):207–221.

Hampel V, et al. 2009. Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups." Proc Natl Acad Sci U S A. 106(10):3859–3864.

He D, et al. 2014. An alternative root for the eukaryote tree of life. Curr Biol. 24(4):465–470.

Heiss AA, Walker G, Simpson AGB. 2011. The ultrastructure of Ancyromonas, a eukaryote without supergroup affinities. Protist 162(3):373–393.

Inagaki Y, Nakajima Y, Sato M, Sakaguchi M, Hashimoto T. 2009. Gene sampling can bias multi-gene phylogenetic inferences: the relationship between red algae and green plants as a case study. Mol Biol Evol. 26(5):1171–1178.

Kang S, et al. 2017. Between a pod and a hard test: the deep evolution of amoebae. Mol Biol Evol. 34:2258–2270.

Katz LA, Grant JR, Parfrey LW, Burleigh JG. 2012. Turning the crown upside down: gene tree parsimony roots the eukaryotic tree of life. Syst Biol. 61(4):653–660.

Katz LA, Grant JR. 2015. Taxon-rich phylogenomic analyses resolve the eukaryotic tree of life and reveal the power of subsampling by sites. Syst Biol. 64(3):406–415.

Lartillot N, Brinkmann H, Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol Biol. 7(Suppl 1):S4.

Le SQ, Lartillot N, Gascuel O. 2008. Phylogenetic mixture models for proteins. Philos Trans R Soc Lond B Biol Sci. 363(1512):3965–3976.

Minh BQ, Nguyen MAT, von Haeseler A. 2013. Ultrafast approximation for phylogenetic bootstrap. Mol Biol Evol. 30(5):1188–1195.
Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 32(1):268–274.

Paps J, Medina-Chacón LA, Marshall W, Suga H, Ruiz-Trillo I. 2013. Molecular phylogeny of unikonts: new insights into the position of apusomonads and ancyromonads and the internal relationships of opisthokonts. Protist 164(1):2–12.

Pawlovski J. 2013. The new micro-kingdoms of eukaryotes. BMC Biol. 11:40.

Philippe H, et al. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol. 9(3):e1000602.

Pisani D, et al. 2015. Genomic data do not support comb jellies as the sister group to all other animals. Proc Natl Acad Sci U S A. 112(50):15402–15407.

Rodrigue N, Lartillot N. 2014. Site-heterogeneous mutation-selection models within the PhyloBayes-MPI package. Bioinformatics 30(7):1020–1021.

Simpson AGB. 2003. Cytoskeletal organization, phylogenetic affinities and systematics in the contentious taxon Excavata (Eukaryota). Int J Syst Evol Microbiol. 53(6):1759–1777.

Simpson AGB, Eglit Y. 2016. Protist diversification. In: Kliman RM, editor. Encyclopedia of evolutionary biology. Vol. 3. Amsterdam: Elsevier. p. 344–360.

Susko E, Field C, Blouin C, Roger AJ. 2003. Estimation of rates-across-sites distributions in phylogenetic substitution models. Syst Biol. 52(5):594–603.

Susko E, Roger AJ. 2007. On reduced amino acid alphabets for phylogenetic inference. Mol. Biol. Evol. 24(9):2139–2150.

Tice AK, et al. 2016. Expansion of the molecular and morphological diversity of Acanthamoebidae (Centramoebida, Amoebozoa) and identification of a novel life cycle type within the group. Biol Direct. 11:69.

Torruella G, et al. 2015. Phylogenomics reveals convergent evolution of lifestyles in close relatives of animals and fungi. Curr Biol. 25(18):2404–2410.

Wang H, Minh B, Susko E, Roger AJ. 2017. Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation. Syst Biol. doi: 10.1093/sysbio/syx068.

Wang H-C, Li K, Susko E, Roger AJ. 2008. A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny. BMC Evol Biol. 8:331.

Yabuki A, Eikrem W, Takishita K, Patterson DJ. 2013. Fine structure of Telonema subtills Griessmann, 1913: a flagellate with a unique cytoskeletal structure among eukaryotes. Protist 164(4):556–569.

Yabuki A, Ishida K-I, Cavalier-Smith T. 2013. Rgifila ramosa n. gen., n. sp., a filose apusozoan with a distinctive pellicle, is related to Micronuclearia. Protist 164:75–88.

Zhao S, et al. 2012. Collodictyon–an ancient lineage in the tree of eukaryotes. Mol Biol Evol. 29(6):1557–1568.

Associate editor: Laura Katz