PHYLOTRANSCRIPTOMICS POINTS TO MULTIPLE INDEPENDENT ORIGINS OF MULTICELLULARITY AND CELLULAR DIFFERENTIATION IN THE VOLVOCINE ALGAE

A Thesis
Presented to
The Academic Faculty

By

Charles Ross Lindsey

In Partial Fulfillment
Of the Requirements for the Degree
Master of Science in the
School of Biological Sciences

Georgia Institute of Technology
December 2021

PHYLOTRANSCRIPTOMICS POINTS TO MULTIPLE INDEPENDENT ORIGINS OF MULTICELLULARITY AND CELLULAR DIFFERENTIATION IN THE VOLVOCINE ALGAE

Approved by:

Dr. Frank Rosenzweig, Advisor
School of Biological Sciences
Georgia Institute of Technology

Dr. Matthew Herron
School of Biological Sciences
Georgia Institute of Technology

Dr. Joseph Lachance
School of Biological Sciences
Georgia Institute of Technology

Date of Approval: January 22nd, 2021

ACKNOWLEDGEMENTS

I extend my thanks to my advisors Dr. Frank Rosenzweig and Dr. Matthew Herron for their guidance throughout this project. I would also like to thank Dr. Joseph Lachance for being part of my committee and providing me with helpful discussions throughout the course of this thesis. I am also grateful for useful discussion with and comments on the manuscript provided by Andy Chea, Emily Cook, Sa Geng, Armin Hallmann, Bradley Olson, and James Umen. I thank Olivier De Clerck for kindly providing amino acid alignments. I gratefully acknowledge the core facilities at the Parker H. Petit Institute for Bioengineering and Bioscience at the Georgia Institute of Technology for the use of their shared equipment, services, and expertise.

This work used the Hive cluster, which is supported by the National Science Foundation under grant number OAC-1828187. This research was supported in part through research cyberinfrastructure resources and services provided by the Partnership for an Advanced Computing Environment (PACE) at the Georgia Institute of Technology, Atlanta, Georgia, USA. This project was funded by NASA Astrobiology Grant# NNA17BB05A to Rosenzweig (PI) Herron (co-I); NSF DEB-1723293 to Herron (PI) Rosenzweig (co-PI); Rosenzweig NASA Exobiology Grant # 80NSSC20K0621 to Rosenzweig (PI); Rosenzweig Georgia Tech start-up account DE00000308.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

RNA-seq	RNA-sequencing
ITS	Internal transcribed spacer
ML	Maximum likelihood
BI	Bayesian inference
CB	Coalescence-based
CC	Chlamydomonas Culture Collection
FACHB	Freshwater Algae at the Institute of Hydrobiology
NIES	National Institute for Environmental Studies
SAG	Culture Collection of Algae at the University of Göttingen
UTEX	University of Texas at Austin
QRPMK	QIAGEN RNeasy Plant Mini Kit
TR	TRizol RNeasy
AU	Approximately Unbiased
MLBS	Maximum Likelihood Bootstrap
BPP	Bayesian Posterior Probability
CPP	Coalescent Posterior Probability
PVC	*Pandorina* + *Volvulina* + *Colemanosphaera*
EVP	*Eudorina* + *Volvox* + Pleodorina
LED	Light emitting diode
BLAST	Basic Local Alignment Search Tool
BLASTP	Protein Basic Local Alignment Search Tool

AIC	Akaike Information Center
NCBI	National Center for Biotechnology Institute
SRA	Sequence Read Archive

SUMMARY

The volvocine algae, which include the single-celled species *Chlamydomonas reinhardtii* and the colonial species *Volvox carteri*, serve as a model in which to study the evolution of multicellularity and cellular differentiation. Studies reconstructing the history of this group have by and large relied on datasets of one to a few genes for phylogenetic inference and ancestral character state reconstruction. As a result, volvocine phylogenies lack concordance depending on the number and/or type of genes (i.e., chloroplast vs nuclear) chosen for phylogenetic inference. While multiple studies suggest that multicellularity evolved only once in the volvocine algae, that each of its three colonial families is monophyletic, and that there have been at least three independent origins of cellular differentiation in the group, other studies call into question one or more of these conclusions. An accurate assessment of the evolutionary history of the volvocine algae requires inference of a more robust phylogeny.

We performed RNA sequencing (RNA-seq) on 55 strains representing 47 volvocine algal species and obtained similar data from curated databases on 13 additional strains. We then compiled a dataset consisting of transcripts for 40 single-copy, protein-coding, nuclear genes, and subjected the predicted amino acid sequences of these genes to maximum likelihood, Bayesian inference, and coalescent-based analyses. These analyses show that multicellularity independently evolved at least twice in the volvocine algae and that the colonial family Goniaceae is not monophyletic. Our data further indicate that cellular differentiation arose independently at least four, and possibly as many as six times, within the volvocine algae. Altogether, our results demonstrate that multicellularity and cellular differentiation are

evolutionarily labile in the volvocine algae, affirming the importance of this group as a model system for the study of major transitions in the history of life.

CHAPTER 1:
INTRODUCTION

The evolution of multicellularity is widely considered a major transition in the history of life [1,2,3,4]. Multicellularity not only gave rise to most of the visible life forms on the planet, but also opened the door to cellular differentiation, including that between somatic and reproductive cells, a hallmark feature of sexual reproduction in eukaryotes that exhibit morphological complexity [3, 5, 6]. Questions regarding the evolution of multicellularity and cellular differentiation have been approached using the fossil record [7,8,9], laboratory evolution [10,11,12,13], and comparative approaches that include superimposing cell biology upon molecular phylogeny [14,15,16]. The last of these approaches is predicated on the assumption that the cell biology and molecular phylogeny are mutually informative, an assumption that requires the phylogeny itself to be accurate.

The volvocine green algae have proved especially useful for investigating the major transition leading to multicellularity. The group consists of ~ 50 extant species, which exhibit a range of body plans, cell numbers, sizes, and forms of sexual reproduction. The smallest of these are single-celled (e.g., *Chlamydomonas reinhardtii*); the largest, at up to 3 mm in diameter and up to 50,000 cells, are spheroidal, swimming colonies in the genus *Volvox*. Since the initial "very pleasant sight" of swimming *Volvox* colonies described by Van Leeuwenhoek more than 300 years ago [17], the volvocine algae have come to be accepted as a useful model system in which to address questions related to the origins of multicellularity and cellular differentiation [18, 19]. Multiple species have now had their genomes sequenced [20, 21, 22, 23, 24], and those of unicellular *C. reinhardtii* and multicellular *V. carteri* forma *nagariensis* are well-annotated

[23, 24]. However, the volvocine algae encompass more than two organisms representing alternative forms of life in terms of size and development. Vegetative forms range in characteristic cell number from 1 to ~ 50,000 and exhibit intermediate degrees of complexity likely similar to extinct ancestors. Further, multicellularity and cellular differentiation arose within the volvocine algae much more recently than those traits arose in animals: ~ 220 million years ago [25] versus ~ 600 million years ago [26], respectively.

Evolution of the volvocine algae has sometimes been viewed as a linear progression in size and complexity [27, 28]. Unicellular taxa such as *Chlamydomonas* occupy one end of this continuum, while fully differentiated, multicellular taxa such as *Volvox* occupy the other. This concept, the "volvocine lineage hypothesis", used a streamlined phylogeny of the volvocine algae to help explain how a multicellular species with complete germ-soma differentiation such as *Volvox* might evolve from a unicellular, *Chlamydomonas*-like ancestor. However, morphological and molecular phylogenetic studies suggest that the history of the volvocine algae may be more complicated, as cellular differentiation, different modes of sexual reproduction, and varying body plans appear to have evolved multiple times within the group [29, 30].

Current understanding of the major evolutionary relationships within this group has often been based on the analysis of five chloroplast gene sequences [14, 25, 31,32,33,34,35]. Chloroplast gene-based phylogenies have also been used to carry out ancestral-state reconstructions [14, 29, 30, 36], opening a window on how multicellularity and cellular differentiation evolved within the volvocine algae. Overall, the branching order of most chloroplast gene-based phylogenies is defined by two related groups: (i) a set of unicellular species (e.g., *Chlamydomonas reinhardtii*) that are paraphyletic with respect to (ii) a clade that encompasses the three major families of colonial volvocine algae: Tetrabaenaceae

(*Tetrabaena* and *Basichlamys*), Goniaceae (*Gonium* and *Astrephomene*), and Volvocaceae (*Colemanosphaera*, *Eudorina*, *Pandorina*, *Platydorina*, *Pleodorina*, *Volvox*, *Volvulina*, and *Yamagishiella*) (Fig. 1a, d). In this scheme, the Tetrabaenaceae is a sister group to the clade formed by the Goniaceae and Volvocaceae. Although this framework only takes into account family-level relationships, several conclusions can be drawn. *First*, the colonial species form a clade. *Second*, each of the three families is monophyletic. *Third*, monophyly among the colonial species implies that multicellularity evolved only once within that group with no reversion to unicellularity.

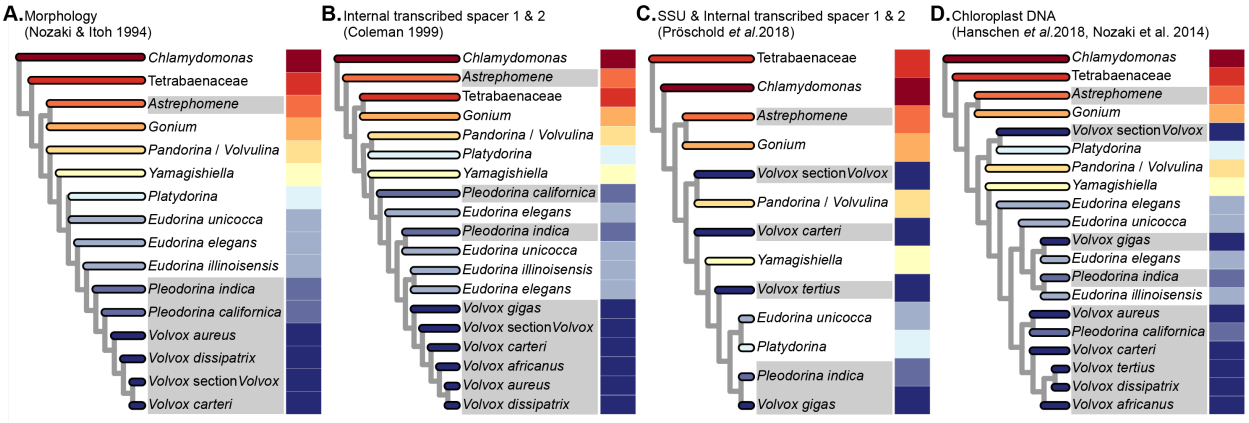


Figure 1. Phylogenies of the volvocine algae are not concordant: Four volvocine green algae phylogenies based on different types of data, displayed in chronological order of their appearance in the literature. Species highlighted in shades of gray exhibit somatic cell differentiation. The varying colors to the right of each phylogeny have been arbitrarily assigned to particular genera and are intended to be used as a visual aid to highlight differences among the phylogenies.

Two recent studies have called into question the monophyly of the colonial volvocine algae (Fig. 1c). Pröschold et al. [37] based their inferences on two datasets: one consisting of SSU *r*DNA sequences plus internal transcribed spacer (ITS) sequences 1 and 2, the other consisting of ITS sequences alone. Nakada et al. [38] used a single-gene 18S *r*RNA dataset. Both

studies inferred that the colonial species are paraphyletic with respect to certain unicells in the genera *Chlamydomonas* and *Vitreochlamys*.

The taxonomic status of the Goniaceae has also been called into question by studies (Fig. 1b) that indicate the group is either not monophyletic [39] or that there is low support for a sister relationship between *Astrephomene* and *Gonium* [33, 37, 38]. Moreover, a number of recent volvocine algal phylogenies leave uncertainty as to how many times cellular differentiation evolved within the group. Chloroplast sequence data suggest at least 3 independent origins of cellular differentiation: in *Astrephomene*, in *Volvox* section *Volvox* (sometimes referred to as *Euvolvox*), and in the *Eudorina*, *Volvox*, *Pleodorina* (EVP) clade (Fig. 1b–d). Within the EVP clade it is unclear whether cellular differentiation in *Pleodorina thompsonii*, *Volvox gigas* and *V. powersii*, and *Pleodorina starrii* and *P. indica* arose independently from that in *V. carteri* (Fig. 1b–d).

The foregoing uncertainties highlight the need for a new and more robust molecular phylogeny of the volvocine algae. These uncertainties may arise from incomplete taxonomic sampling, limited genetic sampling, or both. While five volvocine algal species have had their genomes sequenced, most taxonomically comprehensive phylogenetic inferences about this evolutionarily important group have been constructed using relatively small datasets. Most consist of the sequence of five chloroplast genes [14, 31, 34, 40] representing an aggregate of ~ 6000 nucleotide positions. Others consist of small (≤ 6) multi-gene datasets consisting of chloroplast gene(s), ribosomal molecular markers, or both [37, 38]. Moreover, the use of chloroplast genes in phylogenetic reconstruction can be problematic because they are effectively a single linkage group, they vary little among recently diverged species [41], and they are at

increased risk of incomplete lineage sorting due to the retention of ancestral polymorphisms [42, 43].

Of special concern is the observation that volvocine phylogenies inferred using chloroplast genes (Fig. 1d) conflict with those constructed using nuclear genes (Fig. 1b, c) [37,38,39]. While conflicts between chloroplast and nuclear phylogenies are not unusual [44,45,46], they do foster ambiguity.

Here, we seek to resolve volvocine relationships using taxonomically dense sampling of multiple, unlinked loci. We have adopted a phylotranscriptomic approach that uses a concatenated amino acid alignment of 40 nuclear protein-coding, single-copy genes. We sequenced whole transcriptomes of 55 strains encompassing 47 nominal species and used previously published RNA-Seq data for 9 strains and amino acid alignments for 4 strains that were shared with our group by the De Clerck laboratory. Our goal was to derive a robust phylogeny of the volvocine algae that would enable inferences about the evolution of multicellularity, cellular differentiation, sexual dimorphism, and other traits in this group. Our results represent the most taxonomically comprehensive phylogeny yet produced of the volvocine algae using a nuclear dataset, including all described genera and multiple representatives of all genera that are not monotypic. Our results show that the colonial species do not form a clade, that the Goniaceae are not monophyletic, and that multicellularity has independently evolved at least twice and cellular differentiation at least four times within the volvocine algae.

CHAPTER 2:

RESULTS AND DISCUSSION

2.1 De Novo Transcriptomic Data Makes Possible 40 Single-Gene Alignements

We sampled 68 taxa representing all presumed major lineages of the colonial volvocine algae and 9 of their nearest unicellular relatives. Because the phylogenetic position of *Chlamydomonas reinhardtii* has recently been called into question [37, 38], we used a member of the Trebouxiophyceae, *Chlorella variabilis*, as an outgroup (Table 1). All described volvocine genera were included, with multiple species represented for every genus that is not monotypic. Truly comprehensive taxon sampling was not possible, since several described species, especially in the genus *Volvox*, are no longer available in culture collections. While our main focus was to resolve relationships within the colonial volvocine algae, our study included several closely related unicellular taxa from the genera *Chlamydomonas* and *Vitreochlamys* in order to provide better phylogenetic resolution of the volvocine algae as a whole.

The total number of raw reads generated from RNA sequencing for each species ranged from 25,665,262 to 87,455,695 reads with an average of 60,194,849 reads per species. After quality trimming of the raw reads (see "Methods"), the total number of clean paired-end reads ranged from 20,161,297 to 69,539,684 with an average of 44,416,935 reads per species (Table 1). From the RNA-seq data, we assembled a total of 40 single-gene alignments that were later concatenated to a single alignment representing an aggregate of 12,650 amino acids, equivalent to 37,950 nucleotide positions, with a total of 5972 parsimony-informative sites. Numbers of informative positions in the single-gene alignments ranged from 40 to 446. Trees inferred using maximum likelihood (ML), Bayesian inference (BI), and coalescence-based (CB)

analyses were generally well-supported with some topological differences between the ML and BI analyses relative to the CB analysis, as described below.

Table 1. List of taxa used in this study and summary of sequencing and assembly. Under Strain or Pubmed ID, "CC" refers to Chlamydomonas Culture Collection at the University of Minnesota (CC, USA), "F" refers to Culture Collection of Freshwater Algae at the Institute of Hydrobiology, Chinese Academy of Sciences (FACHB, China), "N" refers to National Institute for Environmental Studies (NIES, Japan), "S" refers to Culture Collection of Algae at the University of Göttingen (SAG, Germany), and "U" refers to The Culture Collection of Algae at the University of Texas at Austin (UTEX, USA). "QRPMK" and "TR" under RNA Extraction Method refer to QIAGEN RNeasy Plant Mini Kit and TRizol RNeasy, respectively. Strains assigned an asterisk represent data from previously published studies, with accession numbers shown in **Supplementary Materials: Table S1 [21, 47]**.

Taxa	Strain or Pubmed ID	RNA Extraction Method	Total # of Raw Reads	Total # of Clean PE Reads	Total # of Contigs	N50	G/C %
Astrephomene gubernaculifera	N-418	QRPMK	61460551	44350080	78105	3894	0.61
A. perforata	N-564	QRPMK	63350620	46811342	72913	3613	0.61
Basichlamys sacculifera	N-566	QRPMK	66205749	48327288	134880	1945	0.69
Colemanosphaera angeleri	F 2363*		28119810	27000176	120763	3348	0.62
C. charkowiensis	F 2326*		27967726	26876865	103286	3382	0.62
	N-3383	TR	58777161	40550696	119848	2523	0.62
Chlamydomonas debaryana	S 11-55a	QRPMK	87455695	64552440	105096	2776	0.67
	S 70.81	QRPMK	50406749	42095108	103962	2858	0.64
C. globosa	S 81.72	TR	73820020	49419928	120062	2079	0.65
C. moewusii	S 11-16f	QRPMK	54728573	40436359	118833	2498	0.6
C. reinhardtii	CC-503*						
C. schloesseri	S 2486	QRPMK	70163393	49255021	64233	3094	0.66
Chlorella variabilis NC64a	20852019*						
Eudorina cylindrica	F 2322*		26442110	25262148	103026	2846	0.64
	N-722	QRPMK	44479173	36297801	105029	2927	0.64
E. elegans	F 2321*						
	N-456	TR	84549961	56756776	122975	2108	0.63
	N-458	QRPMK	67451343	46725012	131335	2944	0.65
	N-568	QRPMK	83433797	69539684	120487	3614	0.64
	N-717	QRPMK	47625429	37616922	89118	2094	0.64
	N-719	QRPMK	57157324	43565312	98180	2976	0.63
	N-720	QRPMK	38134498	31388413	95662	3699	0.61

E. illinoisensis	N-460	QRPMK	62485922	49396583	97014	2615	0.64
E. minodii	N-856	QRPMK	83574258	66413244	128802	4127	0.61
E. peripheralis	N-725	QRPMK	46607143	36528930	100934	2322	0.62
E. unicocca	S 24-1c	TR	54482446	37679863	110391	2267	0.62
Gonium multicoccum	N-737	QRPMK	65984190	44301578	122279	3275	0.64
G. pectorale	N-2863*						
G. octonarium	N-851	TR	60029472	39826211	104857	2302	0.65
G. quadratum	N-653	QRPMK	62431743	48230702	100468	3789	0.64
G. viridistellatum	N-654	QRPMK	43286980	31852410	97991	2837	0.65
Pandorina colemaniae	F 2361*		25181003	23823370	80843	3292	0.62
	N-572	TR	49862534	37589048	84053	2237	0.62
P. morum	F 2362*		29837807	28442854	147990	2674	0.61
	N-890	QRPMK	39337532	24490235	133742	3123	0.61
Platydorina caudata	N-728	QRPMK	48702213	36896556	125206	3503	0.61
Pleodorina indica	N-736	QRPMK	57564627	44082516	192570	2006	0.63
P. japonica	U 2523	QRPMK	64897803	49495315	83662	3499	0.61
P. starrii	N-1362	TR	68790016	45935282	87641	2438	0.64
	N-1363	TR	59466876	44726861	128732	1695	0.63
P. thompsonii	N-4126	QRPMK	81077847	61986735	118940	4005	0.62
Tetrabaena socialis	N-571*		42302450	41232209	62014	1012	0.65
Vitreochlamys aulata	N-878	QRPMK	80710519	54651118	132115	2673	0.65
	S 80.81	QRPMK	41008694	32585611	84834	2100	0.65
V. nekrassovii	S 11-10	QRPMK	65658295	48341108	109360	3304	0.62
V. ordinata	N-882	TR	60631467	42981449	75434	2778	0.69
Volvox africanus	N-863	QRPMK	70739371	52378949	110066	4586	0.55
V. aureus	N-541	QRPMK	80821130	64315625	106245	4065	0.55
V. barberi	N-730	QRPMK	49805303	37197105	152910	2217	0.58
V. carteri f. *kawasakiensis*	N-732	TR	53994112	34239609	86004	1982	0.58
V. carteri f. *nagariensis*	20616280*						
	N-865	QRPMK	65624222	50641434	117314	3947	0.57
V. carteri f. *weismannia*	N-866	QRPMK	25913610	20987137	78230	2636	0.57
V. dissipatrix	N-4128	QRPMK	77137081	51045569	112190	5146	0.55
V. ferrisii	N-3986	QRPMK	49810357	37522413	91045	3281	0.57
V. gigas	N-867	QRPMK	50554912	37592146	151302	2499	0.63
V. globator	S 199.80	QRPMK	67413919	52861449	100861	4005	0.56
V. kirkiorum	N-543	QRPMK	44941633	34650729	89533	3081	0.56
V. obversus	N-868	QRPMK	67394604	50823546	99171	3881	0.58
V. ovalis	N-2569	QRPMK	52242003	35319700	107999	3643	0.53
V. powersii	N-4127	QRPMK	55232784	42410688	140237	2908	0.62
V. tertius	N-544	QRPMK	86222840	62563240	108551	5454	0.54
Volvulina boldii	N-893	QRPMK	25665262	20161297	158265	1593	0.62
V. compacta	F 2337*		27543035	26374210	90261	3550	0.61
	N-582	QRPMK	53045384	38991781	112583	3380	0.63

V. pringsheimii	N-895	TR	82233959	55849890	128304	3122	0.62
V. steinii	S 90-1	QRPMK	46133638	37699635	89515	3635	0.62
Yamagishiella unicocca	F 2364*		30922957	29574420	64415	3609	0.63

2.2 Our results conflict with prior volvocine algal phylogenies in four respects

First, we find that the colonial volvocine algae are paraphyletic with respect to some unicellular species. *Second*, monophyly of the family Goniaceae is not supported. *Third*, section *Volvox* is inferred to be sister to the remaining Volvocaceae. *Fourth*, cellular differentiation independently arose at least four and perhaps as many as six times within the volvocine algae.

2.3 Colonial volvocine algae are not monophyletic

All three of our phylogenetic analyses indicate that the colonial volvocines are not monophyletic (Figs. 2 and 3); further, an *a*pproximately *u*nbiased (AU) test strongly rejected monophyly for this group ($p = 2.82e{-}38$) (Additional file 1: Fig. S1a). These findings represent a major departure from earlier chloroplast gene-based volvocine phylogenies [14, 25, 31,32,33,34, 40, 48], phylogenies based on morphological characters [49, 50], phylogenies inferred using ITS 1 and 2 sequences [39], as well as less taxonomically comprehensive phylogenies inferred using nuclear data [51], all of which suggest that the colonial volvocine algae are monophyletic.

Consistent with Pröschold et al. [37], our results support the view that multicellularity evolved independently in the Tetrabaenaceae and in the Goniaceae + Volvocaceae. In each analytical framework, the Tetrabaenaceae was found to be sister to *Vitreochlamys ordinata* rather than to the Goniaceae + Volvocaceae (*M*aximum *L*ikelihood *B*oot*s*trap [MLBS]

= 100, *B*ayesian *P*osterior *P*robabilities [BPP] = 1.0, Coalescent Posterior Probabilities [CPP] = 1.0). A sister relationship between the Tetrabaenaceae and *V. ordinata* was inferred in 17/39 of our single-gene phylogenies and in 27/39 of our 4-taxa, unrooted, single-gene phylogenies (Fig. 4). These results imply one independent origin of multicellularity in the Tetrabaenaceae and another origin in the Goniaceae + Volvocaceae.

Our results differ in key respects from a recent volvocine algal phylogeny inferred by Zhang et al. [51], which like ours is based on single-copy nuclear genes. Zhang et al. [51] sought to understand the evolutionary relationships between two psychrophilic algae: *Chlamydomonas* sp. ICE-L and *Tetrabaena socialis* N-691. To do so, they constructed a phylogeny consisting of ICE-L, N-691, three colonial *Volvox* strains, and eight unicellular species, including *C. reinhardtii*. Among their conclusions was that *T. socialis* N-691 is sister to the Volvocaceae, which is at odds with results shown in Figs. 2 and 3. These results indicate that the Tetrabaenaceae is sister to *V. ordinata*, and together they are sister to *C. reinhardtii* + Goniaceae + Volvocaceae.

We hypothesized that the lack of concordance between our findings and those of Zhang et al. [51] could be attributed to limited taxon sampling. To test this hypothesis, we first confirmed that *T. socialis* N-691 and *T. socialis* N-571 are conspecific (Additional file 2: Confirming the conspecificity of *Tetrabaena socialis*N-571 and N-691) [52, 53]. Once we confirmed that N-691 and N-571 were conspecific, we were able to replicate the branching order produced by Zhang et al. [51] using our concatenated 40-gene dataset (Additional file 1: Fig. S2a) [51]. For our initial tree, we sampled our strains of *Chlamydomonas reinhardtii*, *C. moewusii*, *T. socialis*, *Volvox aureus*, *V. carteri* f. *nagariensis*, and *V. globator* to match taxa that were used in that study. For an outgroup species, we sampled *Chlorella*

variabilis. Multiple studies have shown that the accuracy of phylogenetic reconstruction can be improved by increasing the number of taxa sampled [54,55,56]. When we added more taxa and performed ML analysis on the new dataset, the three colonial volvocine families were no longer monophyletic. The Tetrabaenaceae were sister to *Vitreochlamys ordinata*, and this clade appeared sister to *C. reinhardtii* + Goniaceae + Volvocaceae (Additional file 1: Fig. S2b) [51]. These analyses confirm that the placement of *T. socialis* N-691 as sister to the Volvocaceae is an artifact of limited taxon sampling. From this, we draw three conclusions: *First*, the colonial volvocine algae are not monophyletic; s*econd*, at least two independent origins of multicellularity occurred within the volvocine algae; *third*, once multicellularity evolved no extant lineage reverted to the ancestral unicellular state (see Figs. 2 and 3).

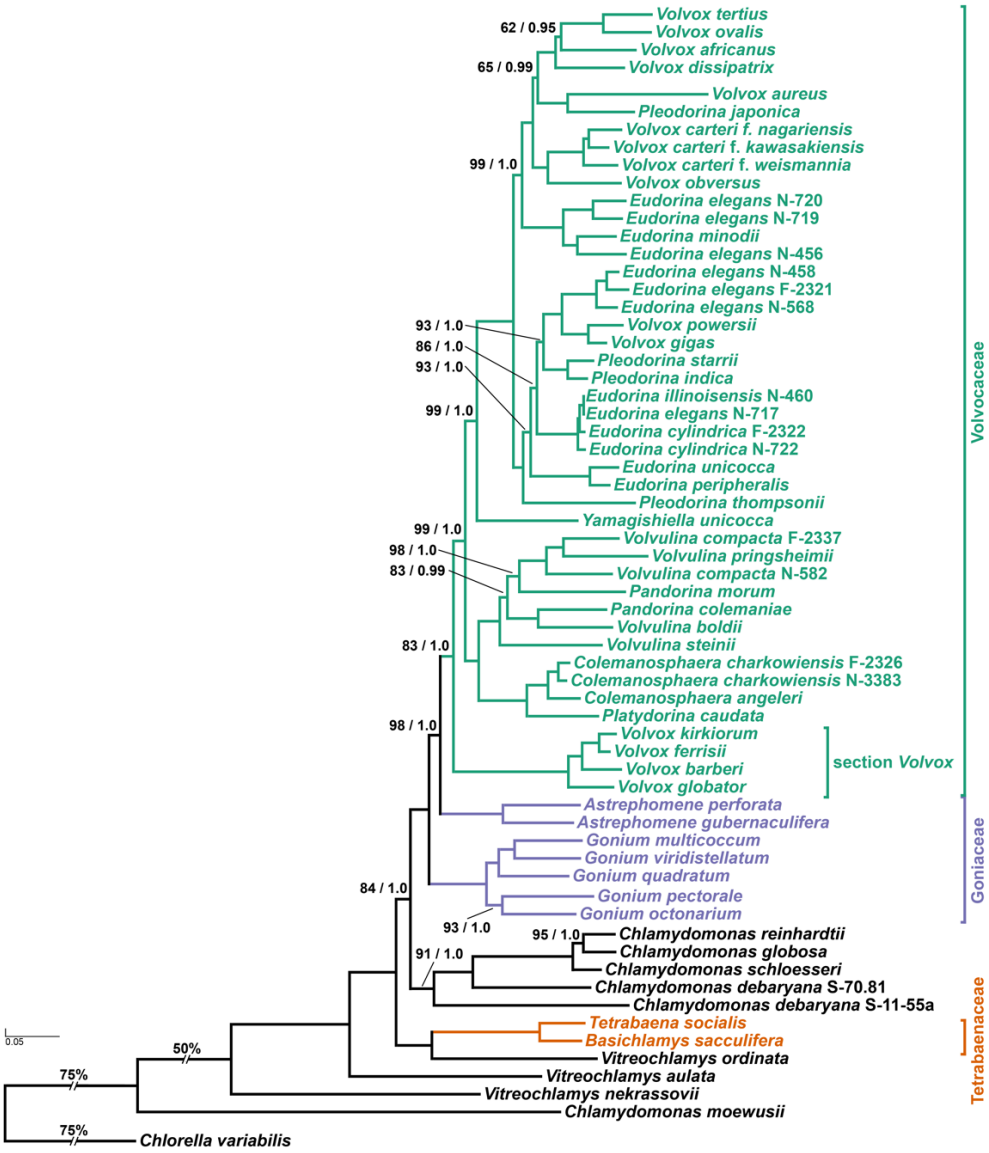


Figure 2. Molecular phylogeny of the colonial volvocine algae (Tetrabaenaceae, Goniaceae, and Volvocaceae) and closely related unicellular taxa represented by *Chlamydomonas* and *Vitreochlamys* with *Chlorella variabilis* as the outgroup. The phylogenetic tree shown is based on a multi-gene dataset of single-copy, protein-coding nuclear genes (12,650 aligned amino acid positions of 68 taxa) inferred using the maximum likelihood method, the branching order of which is identical to that inferred in the Bayesian Inference using MrBayes. Numbers on branches represent bootstrap values and Bayesian posterior probabilities, respectively (all support values not shown are MLBS = 100, BPP = 1.0). Branch lengths correspond to genetic divergence, as indicated by the scale bar. Members of the Tetrabaenaceae, Goniaceae, and Volvocaceae are denoted in orange, purple, and green, respectively; unicellular species are denoted in black.

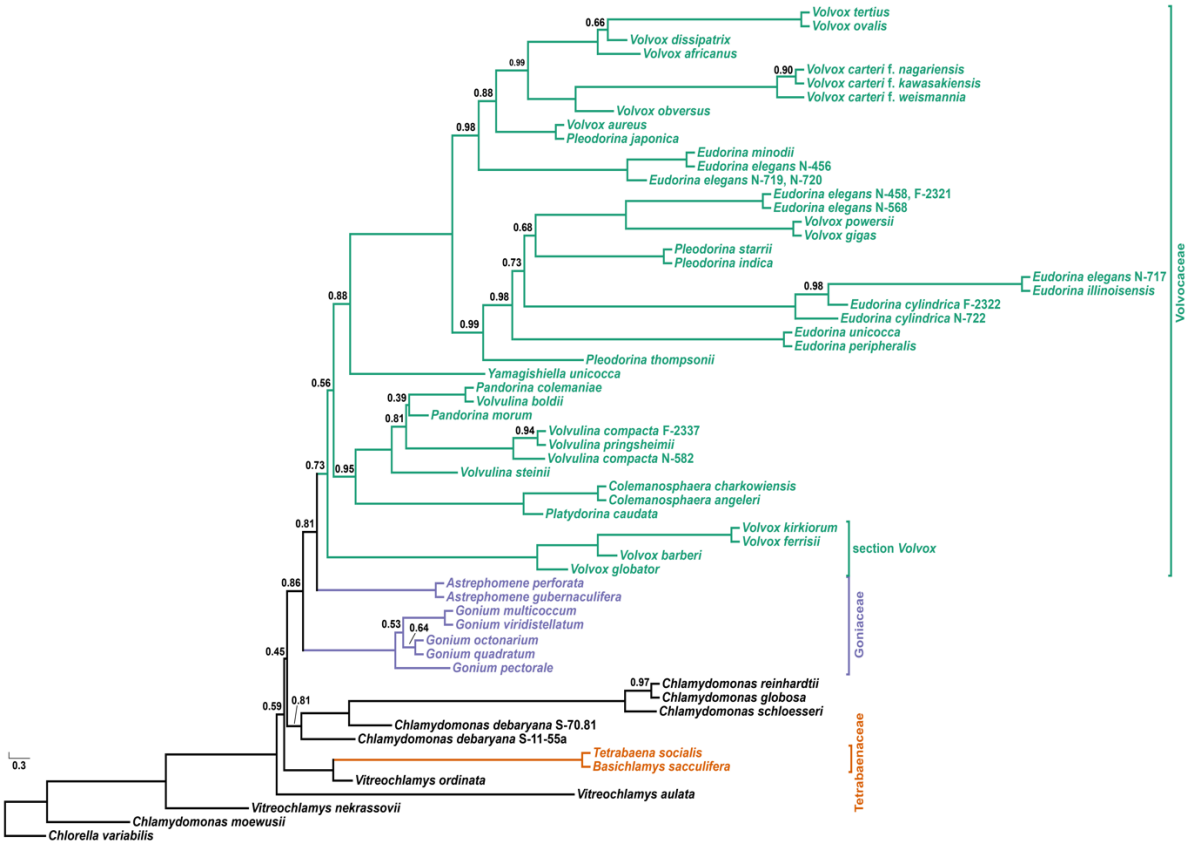


Figure 3. Phylogeny of the volvocine algae inferred using a coalescent-based analysis of 40 single-gene phylogenies. Numbers on branches represent posterior probabilities (support values not shown are CPP = 1.0). Members of the Tetrabaenaceae, Goniaceae, and Volvocaceae are denoted in orange, purple, and green, respectively; unicellular species are denoted in black.

2.4 The family Goniaceae is not monophyletic.

Multiple volvocine phylogenies have concluded that the Goniaceae is monophyletic [14, 25, 29, 31, 33, 37, 38, 49, 50, 57, 58]. Our analyses suggest otherwise (Figs. 2and 3): we find that *Astrephomene* is sister to the Volvocaceae (MLBS = 98, BPP = 1.0, CPP = 0.81) rather than to *Gonium*. This inference is strengthened by observations that 37/40 of our single-gene phylogenies show that *Gonium* and *Astrephomene* are not sister taxa, as do 20/40 of our four-taxon, unrooted phylogenies (Fig. 4). All three of our analyses indicate that *Astrephomene* is monophyletic and sister to the Volvocaceae clade (MLBS = 98, BPP = 1.0, CPP = 0.81),

with *Gonium* sister to *Astrephomene* + Volvocaceae (MLBS = 100, BPP = 1.0, CPP = 0.86). Furthermore, we performed an AU test where the monophyly of the Goniaceae was tested against our finding of paraphyly for the Goniaceae. The null hypothesis, monophyly of the Goniaceae, was rejected (p = 0.0446) (Additional file 1: Fig. S1b). The inferred sister relationship between *Astrephomene* and the Volvocaceae is also consistent with the apparent synapomorphy of zygote germination producing a single gone cell, which is unique to these two taxa [50].

Prior studies have produced mixed results regarding monophyly of the Goniaceae, sometimes with low support values for the relevant relationships. Nozaki and colleagues [59] published four phylogenies inferred using a single chloroplast gene and different inference methods; all four trees either showed low support for monophyly of the Goniaceae or suggested a topology where *Astrephomene* is sister to *Gonium* + Volvocaceae. Coleman [39] inferred a volvocine phylogeny based on ITS-1 and ITS-2 sequences that showed *Astrephomene* sister to Tetrabaenaceae + *Gonium* + Volvocaceae; however, the bootstrap support for this suggested relationship was between 50 and 75%, indicating weak support for the branching order. Other phylogenies suggesting monophyly in the Goniaceae do so with weak or contradictory support [33, 37, 38].

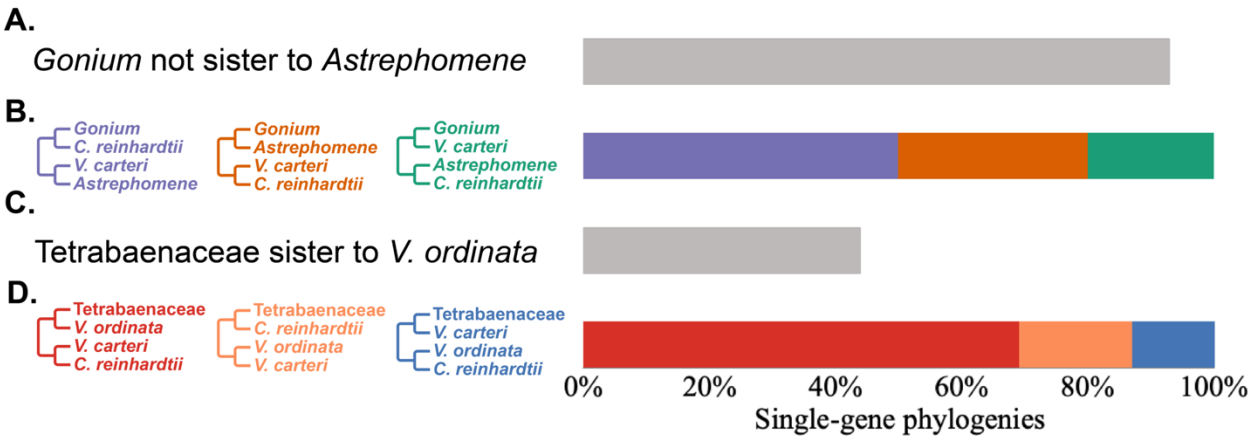


Figure 4. Phylogenetic relationships between *Gonium* and *Astrephomene* and between the Tetrabaenaceae and *Vitreochlamys ordinata*. Four-taxon, unrooted trees were generated by collapsing our single-gene phylogenies. The percentage of single-gene phylogenies representing a specific four-taxon, unrooted tree is represented by the purple, orange, and green bars for trees containing *Gonium* and *Astrphomene*, and red, orange, and blue for trees containing the Tetrabaenaceae. **(A)** percentage of single-gene phylogenies that show *Gonium* not sister to *Astrephomene* represented by the grey bar. **(B)** percentage of four-taxon, unrooted trees representing specific relationships between *Gonium* and *Astrephomene*. **(C)** percentage of single-gene phylogenies that show Tetrabaenaceae sister to *V. ordinata* represented by the grey bar. **(D)** percentage of four-taxon, unrooted trees representing specific relationships between the Tetrabaenaceae and *V ordinata* in four-taxon, unrooted trees. For all relationships involving *V. ordinata*, 39 out of 40 single-gene phylogenies were used due to *V. ordinata* not appearing in one of the inferences. All single-gene phylogenies were inferred using maximum likelihood under the appropriate evolutionary model as estimated by ProtTest.

Our inference that the Goniaceae are not monophyletic is consistent with some — but not all — of the analyses recently reported by Pröschold et al. [37] and Nakada et al. [38]. However, we should not disregard past morphological and ultrastructural studies suggesting a close relationship between *Astrephomene* and *Gonium* [50, 60, 61]. These taxa differ from the Volvocaceae in that each cell, rather than the entire colony, is surrounded by a tripartite boundary [62]. This feature distinguishes their mode of colony formation from all other colonial algae within the Volvocaceae; our results suggest that it is ancestral to the Goniaceae + Volvocaceae and lost in the Volvocaceae.

2.5 Volvox section Volvox is sister to the remaining Volvocaceae.

Our data indicate that *Volvox* section *Volvox* is not a subclade within either the *Pandorina* + *Volvulina* + *Colemanosphaera* (PVC) or *Eudorina* + *Volvox* + *Pleodorina* (EVP) subclades. Older studies based on the *rbcL* chloroplast gene [49], ITS-1 and ITS-2 sequences [39], and morphology [50] suggest that section *Volvox*belongs to a clade that encompasses *Eudorina*, *Pleodorina*, and other *Volvox*species. More recent studies of the volvocine algae based on 5 chloroplast genes, or based on multiple datasets that include 1 chloroplast gene [37], suggest that section *Volvox* belongs to a clade that includes *Pandorina*, *Volvulina*, and *Platydorina* [14, 31], and (in the studies where it was included) *Colemanosphaera* [34, 40]. By contrast, all of our analyses indicate that section *Volvox* is monophyletic and sister to the remaining Volvocaceae (MLBS = 83, BPP = 1.0, CPP = 0.73). AU tests rejected the monophyly of section *Volvox* + *Colemanosphaera* + *Platydorina* (p-AU = 4.64e− 88) and the monophyly of section *Volvox* + the PVC clade (p-AU = 0.0332) (Additional file 1: Fig. S1c). These results bolster our finding that section *Volvox* is sister to the remaining Volvocaceae (Figs. 2 and 3).

2.6 Cellular differentiation independently arose at least four times in the volvocine algae.

The last major difference between our results and earlier phylogenies concerns the number of independent origins of cellular differentiation. Prior literature suggests that cellular differentiation independently evolved at least three times: once in *Astrephomene*, once in section *Volvox*, and at least once in the EVP clade [14, 36]. By contrast, our results show a *minimum* of four independent origins of cellular differentiation: one in *Astrephomene*, one in

section *Volvox*, and at least two in the EVP clade (Fig. 5a). We cannot exclude the possibility of two additional independent origins in the branches leading to *Pleodorina starrii* and *Volvox gigas*(Fig. 5a). In *Astrephomene*, section *Volvox*, *Pleodorina*, and *Volvox dissipatrix*, differentiated cells carry out the function of motility, whereas undifferentiated cells participate in both motility and reproduction [15]. The remaining *Volvox* species within the EVP clade have all evolved specialized germ cells for reproduction and somatic cells for motility [25, 30].

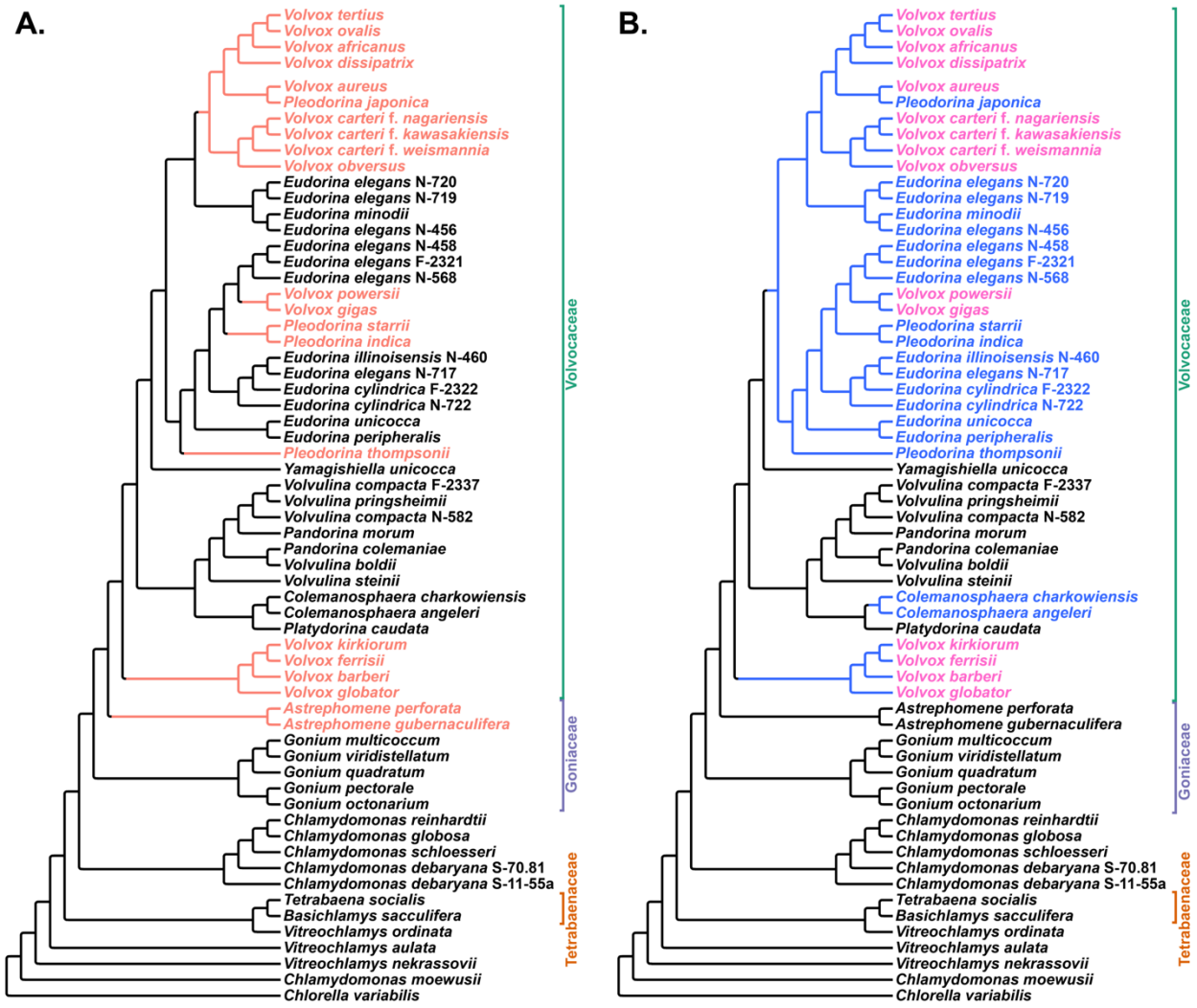


Figure 5. (A) Phylogeny of the volvocine algae highlighting the lineages in which soma differentiation has evolved (peach). This tree indicates a minimum of four and maximum of six independent origins of cellular differentiation **(B)** Phylogeny of the volvocine algae highlighting

the lineages that are isogamous (black), anisogamous (blue), and oogamous (names in pink only). Both phylogenies were inferred using maximum likelihood.

2.7 Isogamy is the ancestral mode of sexual reproduction.

Consistent with past studies, our results suggest that isogamy, the production of similar sized, motile gametes, is the ancestral mode of sexual reproduction among the volvocine algae (Fig. 5b and Additional file 1: Table S2). Isogamy is present in the unicellular genera *Chlamydomonas* and *Vitreochlamys* and is retained within the multicellular genera *Astrephomene*, *Basichlamys*, *Gonium*, *Pandorina*, *Platydorina*, *Tetrabaena*, *Volvulina*, and *Yamagishiella*. *Colemanosphaera*, *Eudorina*, *Pleodorina*, and *Volvox* have all evolved either anisogamy or oogamy [34, 63,64,65]. Anisogamy appears to have independently evolved at least three times from an isogamous ancestor: in section *Volvox* and in both *Colemanosphaera* and EVP. Conventional anisogamy, which consists of two motile gamete types of unequal size, appears in *Colemanosphaera*, *Eudorina*, and *Pleodorina*. This finding differs from those of Hanschen et al. [29], who reported that anisogamy independently evolved twice among the volvocine algae from isogamous ancestors. Oogamy, a specialized form of anisogamy where the female gamete is immotile and significantly larger than the motile, male gamete, is inferred to have independently evolved at least three times in lineages leading to section *Volvox*, *V. gigas* + *V. powersii*, and in the clade containing *V. africanus*, *V. aureus*, *V. carteri*, *V. dissipatrix*, *V. obversus*, *V. ovalis*, and *V. tertius* [34, 63, 64]. This last finding confirms results from Hanschen et al. [29], who also reported at least three independent origins of oogamy among the volvocine algae.

2.8 Platydorina caudata is sister to Colemanosphaera, and Pandorina is paraphyletic with respect to Volvulina.

Within the PVC clade, our results add further support to the view that *Pandorina* is paraphyletic with respect to *Volvulina* (Figs. 2 and 3) [14, 25, 29, 33, 34, 39, 66]. Also, consistent with other multi-gene analyses *Colemanosphaera* appears to be monophyletic with high support (MLBS = 100, BPP = 1.0, CPP = 1.0) and sister to *Platydorina* (MLBS = 100, BPP = 1.0, CPP = 1.0) (Figs. 2 and 3) [31, 34, 35].

2.9 The genera Eudorina, Volvox and Pleodorina are polyphyletic.

Yamagishiella unicocca is sister to the *Eudorina+Volvox+Pleodorina* (EVP) clade, which encompasses two large subclades (MLBS=99, BPP=1.0, CPP=0.88) (Figs. 2and 3). Our results support prior work suggesting that the genera *Volvox*, *Eudorina*and *Pleodorina* are not monophyletic [14, 25, 29,30,31, 33,34,35,36, 39, 67, 68]. The genus *Volvox* appears to be polyphyletic, with members represented across the two EVP subclades and the section *Volvox* clade. Members of both the *Pleodorina*and *Eudorina* genera are inferred to be polyphyletic across the two EVP subclades.

Historically, the genus *Volvox* has been divided into 4 sections – *Copelandosphaera*, *Janetosphaera*, *Merrillosphaera*, and *Volvox* – based on morphological [69] and molecular data [67]. A recent section-level revision of the genus *Volvox* [35] resulted in the creation and deletion of sections *Besseyosphaera* and *Copelandosphaera*, respectively. Hereafter, we will only refer to the revised taxonomic sections proposed by Nozaki et al. [35], with which our maximum likelihood, Bayesian inference, and coalescent-based results are in agreement (Additional file 1: Fig. S3) [35]. Our coalescent-based analysis suggests that each of the four sections is monophyletic, and that none encompass novel taxa not listed by Nozaki et al. [35] (Fig. 3). The branching order of our ML and BI analyses, however, suggests that

section *Merrillosphaera* is not monophyletic (Additional file 1: Fig. S3) [35]. Our ML and BI analyses indicate that *V. africanus*, *V. dissipatrix*, *V. ovalis*, and *V. tertius* form a clade with *V. aureus* and *P. japonica* that is separate from the other *Merrillosphaera* taxa (MLBS=65, BPP=0.99) (Additional file 1: Fig. S3) [35]. In contrast, our CB analysis provides strong support (CPP=0.99) for the inference that the *Merrillosphaera* species are monophyletic (Fig. 3). Heeding our support values rather than only the branching order, we propose that the taxonomic system of the genus *Volvox* as outlined by Nozaki and colleagues [35] be retained.

2.10 Unicellular taxa are nested within the clade containing the colonial volvocine algae.

Of the unicellular taxa, *Chlamydomonas debaryana*, *C. globosa*, *C. reinhardtii*, *C. schloesseri*, and *Vitreochlamys ordinata* are nested within the clade containing the colonial volvocine algae. Our results confirm prior studies showing the genus *Vitreochlamys* to be polyphyletic [38, 48]. The closest unicellular relative to the clade that contains the colonial algae + *C. reinhardtii* is suggested to be *V. aulata* (Figs. 2 and 3). This suggests that at least some members of *Vitreochlamys* are very closely related to the colonial volvocine algae. This relationship had been previously suggested by other studies [38, 70] including Nakazawa et al. [48], whose ultrastructural studies uncovered striking similarities in how these taxa formed pyrenoids and eyespot apparati (stigma), and established their tripartite cell walls.

Chlamydomonas is a polyphyletic genus [20, 38, 71, 72] composed of at least 500 species [72]. Although we sampled only a handful of *Chlamydomonas* species, our data support this view and broadly agree with the *Chlamydomonas* relationships inferred by Pröschold et al. [37], who used a combination of molecular phylogenetic analyses, sporangium wall lysis tests, and ultrastructural analyses. Our data strongly support *C. schloesseri* being sister to *C. reinhardtii* + *C. globosa* (MLBS = 100, BPP = 1.0, CPP = 1.0) and

designating *C. schloesseri* as a "true" *Chlamydomonas* species, as suggested by Pröschold et al. [37]. Our study is also in agreement with a recent study by Craig et al. [20] that shows *C. schloesseri* being sister to *C. reinhardtii* + *C. globosa*. Also, like Pröschold et al. [37], our analyses indicate that *C. debaryana* SAG 70.81 is sister to *Chlamydomonas schloesseri* and its relatives (MLBS = 100, BPP = 1.0, CPP = 1.0). However, unlike the Pröschold et al. [37] study, which proposed that strain *C. debaryana*/*Edaphochlamys debaryana* (SAG 11-55a) is sister to the Tetrabaenaceae, our analyses support the view that *C. debaryana*/*Edaphochlamys debaryana* is more closely related to *C. reinhardtii* (MLBS = 91, BPP = 1.0, CPP = 0.81) than to the colonial algae. Our finding is further supported by Craig et al. [20] who inferred that *C. debaryana*/*Edaphochlamys debaryana* + *Chlamydomonas sphaeroides* is sister to the clade containing *C. schloesseri* + *C. reinhardtii* + *C. globosa*. Our placement of *C. debaryana* (SAG 11-55a) could be a result of limited (N = 6) sampling within the *Chlamydomonas* genus, which was more extensively sampled by Pröschold et al. [37] (N > 30). Consistent with a prior study, *C. moewusii* appears to be more distantly related to the colonial volvocines than is *Vitreochlamys nekrassovii* [14].

2.11 Conclusions

Using a 40-protein dataset, we have shown that the Tetrabaenaceae and the Goniaceae + Volvocaceae likely represent two independent origins of multicellularity and that cellular differentiation has independently evolved at least four, and possibly six times within the volvocine algae. The separate origin of multicellularity within the Tetrabaenaceae highlights the need for certain volvocine genomes, such as *Vitreochlamys ordinata*, to be sequenced, assembled and annotated. Because *Vitreochlamys ordinata* is the unicellular sister taxon to the multicellular

Tetrabaenaceae, detailed analysis of its genome could give future researchers insight into how the simple form of multicellularity observed among the Tetrabaenaceae might have evolved.

Our results suggest that both multicellularity and cellular differentiation are evolutionarily labile traits within the volvocine algae. We have established a robust phylogeny of this group, which we hope will assist future efforts aimed at re-evaluating ancestral character states and understanding the origins of multicellularity and cellular differentiation in the volvocine green algae. The fruit of such efforts could then be used to carry out ancestral-state reconstruction of traits related to cellularity, differentiation, and gamete size as well as to discern the evolutionary history of gene families across the volvocine algae as a whole and within its major clades.

CHAPTER 3:

METHODOLOGY

3.1 Strains and Culture Conditions

Algal strains used in this study were obtained from the National Institute for Environmental Studies (NIES, Japan), the Culture Collection of Algae at the University of Göttingen (SAG, Germany), and the Culture Collection of Algae at the University of Texas at Austin (UTEX, USA). Strain provenance and culture collection ID numbers are shown in Table 1, with previously published data designated with an asterisk. All cultures were grown at 20–26 °C under cool-white LED lamps (4300K) with an intensity of 2500–2700 lux under a 14-h light/10-h dark cycle. A detailed description of each strain's morphology, degree of cellular differentiation, and gamete size, as well as the medium used to culture each strain is provided in Additional file 1: Tables S2 and S3 [73, 74, 75, 76, 77], respectively.

3.2 RNA Extraction Procedures

Two protocols were used to isolate total RNA: a modified version of the TRizol RNeasy method described by Matt and Umen [78] and a slightly modified QIAGEN RNeasy Plant Mini Kit protocol. For a detailed description of each, please see Additional file 2: RNA extraction procedures. Information on the protocol used for each strain is provided in Table 1.

3.3 Library Preparation and Sequencing

Before generating a sequencing library, RNA quality and quantity were assessed by Nanodrop and Qubit (Thermo Fisher Scientific, Waltham, MA 02451 USA). RNA integrity was

evaluated using an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA 95051, USA). mRNA was isolated using poly T beads, whereafter Illumina libraries were prepared using the NEBNext Ultra II Directional RNA Library Prep Kit. Library concentrations were determined fluorometrically; sequencing was carried out on the Illumina NovaSeq 6000 platform (Illumina, Inc., San Diego, CA 92122 USA) to generate 151 bp paired-end reads.

3.4 Quality Control of Reads

Raw read quality was assessed through FastQC v.0.11.8 with an additional FastQC assessment post-trimming. Quality control of the raw reads was completed with Trimmomatic v.0.39 [79] where the bases at the 5′ and 3′ end of each read are trimmed if found to be below a quality score of 3. A 4-base sliding window approach was used to trim the rest of the read once average quality fell below a score of 15; reads that were below a minimum length of 36 bases were discarded (LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36). If adapter content was detected by FastQC the additional ILLUMINACLIP step was used with the "TruSeq3-PE-2.fa" file provided by the Trimmomatic developers. If performed, the following ILLUMINACLIP parameters were used: 2:30:10 at the beginning of each command line. This allows for 2 "seed" mismatches where the seed is a short segment of the adapter that is being aligned in every section of the read. If more than 2 mismatches occurred, no trimming of the read occurred. Additionally, there had to be at least 30 matched bases in the paired-end palindrome read alignment and at least 10 matched bases between an adapter sequence and read.

3.5 De Novo Assembly

SOAPdenovo-Trans v1.0.4 [80] was used to assemble de novo transcriptomes from the quality filtered, paired-end reads using a k-mer size of 25 (SOAPdenovo-Trans-31mer all -s <config input file> -o <outfile> -K 25). GapCloser from the SOAPdenovo package was utilized to close gaps in each transcriptome using the same configuration file, which contains read-specific information and file paths, from the previous step (-b <config file> -a <.scafSeq file output by SOAPdenovo-Trans> -o <outfile> -l <max read length, int value> -t <thread number>). Default parameters were used for CD-HIT v4.8.1 [81] to reduce redundant transcripts from our de novo transcriptomes.

3.6 Orthologous Gene Identification for Phylotranscriptomic Analysis

The evolutionary history of the volvocine algae dates back at least 200 million years [25]. Over this timescale nucleotide sequences become saturated with substitutions, diminishing their phylogenetic utility [82]. Amino acid sequences were therefore chosen for our alignments, as they are known to be more reliable for ascertaining distant evolutionary relationships [83]. De Clerck and colleagues identified 58 nuclear protein-coding, single-copy genes that were members of highly conserved gene families across the green algae (*Chlorophyceae*, *Prasinophytes*, and *Trebouxiophyceae*) and land plants (*Streptophyta*) [84]. Their amino acid alignment of the 58 nuclear protein-coding genes that includes *Chlamydomonas reinhardtii* CC-503, *Chlorella variabilis* NC64A, *Gonium pectorale* NIES-2863, and *Volvox carteri* HK10 was kindly shared with our research team. Out of the 58 genes shared, we used 40 for our gene alignments. In order to identify those specific genes in the de novo transcriptomes of our taxa, a Basic Local Alignment Search Tool (BLAST)

server was established in our lab, and a unique BLAST database for each taxon was created following the instructions in the BLAST manual. A BLASTP search using the *C. reinhardtii* CC-503, *G. pectorale* NIES-2863, and *V. carteri* HK10 genes from De Clerck et al. [84] as our query sequences enabled us to identify the orthologous genes for each of our taxa.

3.7 Gene Sequence Alignments and Phylotranscriptomic Analysis

The BLASTP results were used to identify the scaffold and open read frame where each gene was located in a strain's transcriptome. Using a custom Python script (Additional file 3), each scaffold was extracted from its transcriptome and translated in the appropriate reading frame; then, the translated scaffold was added to an alignment file. For consistency, we generated de novo transcriptomes since we lacked a reference genome for most of our sequenced strains. At times, a gene was found to be incomplete for a given taxon due to assembler or sequencing error after manual examination. When this was determined to be the case, the gene was manually stitched together. This was done in a highly conservative manner: if we could not ascertain whether or not a gene was incomplete due to assembler or sequencing error, then it was excluded from the alignment for the given species. We treated the data from previously published studies in the same fashion as data generated in our lab by filtering the raw reads through quality trimming, then assembling de novo transcriptomes using the same programs and parameters (Table 1).

Amino acid sequences were aligned using MUSCLE v3.8.31 [85]. Alignments were also subjected to manual alignment in Aliview v1.26 [86]; extraneous data were trimmed, leaving only the aligned genes. Ambiguously aligned regions were eliminated from each alignment leaving only conserved and reliably aligned regions for phylogenetic analysis using the following

parameters in Gblocks v0.91b [87]: -t=p -b3=8 -b4=2 -b5=h -b6=y. Phyutility v2.7.1 [88] was used to concatenate all gene alignment files.

Single-gene alignments were subjected to ML and BI analyses in order to infer single-gene phylogenies. Single-gene phylogenies were then further analyzed using a coalescent-based approach. The concatenated multi-gene alignment was partitioned so that the appropriate model of protein substitution was applied to each gene for the supermatrix phylogenetic approach under ML and BI.

The ML and BI analyses of the concatenated dataset used a partitioning strategy where the best evolutionary model for each gene was predicted by ProtTest v3.4.2 under the Akaike Information Criterion (AIC). For information regarding each predicted evolutionary model, please refer to Additional file 1: Table S4 [84, 89,90,91]. The ML analysis was conducted using IQtree v1.6.12 [92] under partition models [93]. Support values reported for the IQtree ML analysis were estimated through the bootstrap technique where 1000 ultrafast bootstrap replicates were generated [94]. The BI analysis was performed with MrBayes 3.2.7a [95] with 3 heated and 1 cold Markov chains, where trees were sampled every 1000 generations for a total of 1,000,000 generations with 1000 trees discarded at the beginning of each chain (ngen = 100000000, samplefreq = 1000, burnin = 1000, nruns = 4, nchains = 4, starttree = random).

ASTRAL [96] was used to perform the coalescent-based analysis where all 40 single-gene phylogenies produced by IQtree were used as the input after collapsing branches with low bootstrap support (< 10) using Newick Utilities v1.6 [97]. Posterior probabilities were assessed for the Bayesian and coalescent-based analyses in MrBayes and ASTRAL, respectively. Lastly, approximately unbiased (AU) tests with 100,000 RELL re-samplings were conducted to test certain key topologies and hypotheses using IQtree (-zw 100000 -au) (Additional file 1: Fig. S1).

APPENDIX A: SUPPLEMENTARY INFORMATION 1

Table S1. Accession numbers of previously published RNA-seq data. Under Collection ID, "FACHB" refers to Culture Collection of Freshwater Algae at the Institute of Hydrobiology, Chinese Academy of Sciences (FACHB, China), and "NIES" refers to National Institute for Environmental Studies (NIES, Japan). All FACHB strains used in this study were sequenced by Hu *et* al. [47] under BioProject PRJNA532307. *Tetrabaena socialis* NIES-571 was sequenced by Featherston *et al.* [21] under BioProject PRJNA393411.

RNA-seq Data Used	Collection ID	Accession #
Colemanosphaera angeleri	FACHB 2363	SRX5666821
C. charkowiensis	FACHB 2326	SRX5666822
Eudorina cylindrica	FACHB 2322	SRX5666825
E. elegans	FACHB 2321	SRX5666826
Pandorina colemaniae	FACHB 2361	SRX5666819
P. morum	FACHB 2362	SRX5666823
Tetrabaena socialis	NIES-571	SRX3367144
Volvulina compacta	FACHB 2337	SRX5666820
Yamagishiella unicocca	FACHB 2364	SRX5666824

Table S2. Information on sampled genera regarding cellularity, typical cell number, differentiation, and gamete size.

	Typical Cell #	Differentiation	Gamete Size
Unicellular Genera			
Chlorella	1	-	Isogamy
Chlamydomonas	1	-	Isogamy
Vitreochlamys	1	-	Isogamy
Multicellular Genera			
Astrephomene	32-128	Somatic	Isogamy
Basichlamys	4	-	Isogamy
Colemanosphaera	16-32	-	Anisogamy
Eudorina	16-32	-	Anisogamy
Gonium	8-32	-	Isogamy
Pandorina	16-32	-	Isogamy
Platydorina	16-32	-	Isogamy
Pleodorina	32-128	Somatic	Anisogamy
Tetrabaena	4	-	Isogamy
Volvox	500-50,000	Somatic and Germ*	Oogamy
Volvulina	8-16	-	Isogamy
Yamagishiella	16-32	-	Isogamy

* *Volvox* lineages that have evolved specialized somatic and germ cell lines are: *Volvox africanus*, *V. carteri*, *V. gigas*, and *V. obversus*. All other *Volvox* species have specialized somatic cells where undifferentiated cells partake in motility and reproductive functions.

Table S3. Medium used to culture each sequenced strain. Under Collection ID, "NIES" refers to National Institute for Environmental Studies (NIES, Japan), "SAG" refers to Culture Collection of Algae at the University of Göttingen (SAG, Germany), and "UTEX" refers to The Culture Collection of Algae at the University of Texas at Austin (UTEX, USA). AF-6 [73], AF-6/2 [73], CA [74], MG [75], and VT [76] were prepared according to media recipes supplied by NIES. HS media [77] was prepared according to media recipe supplied by Chlamydomonas Resource Center. Soilwater: GR+ media was purchased from UTEX.

Taxa	Collection ID	Culture Medium
Astrephomene gubernaculifera	NIES-418	VT
A. perforata	NIES-564	VT
Basichlamys sacculifera	NIES-566	AF-6
Colemanosphaera charkowiensis	NIES-3383	AF-6
Chlamydomonas debaryana	SAG 11-55a	HS
	SAG 70.81	HS
C. globosa	SAG 81.72	HS
C. moewusii	SAG 11-16f	HS
C. schloesseri	SAG 2486	HS
Eudorina cylindrica	NIES-722	AF-6
E. elegans	NIES-456	VT
	NIES-458	VT
	NIES-568	CA
	NIES-717	AF-6
	NIES-719	VT
	NIES-720	VT
E. illinoisensis	NIES-460	VT
E. minodii	NIES-856	AF-6
E. peripheralis	NIES-725	AF-6
E. unicocca	SAG 24-1c	AF-6
Gonium multicoccum	NIES-737	VT
G. octonarium	NIES-851	AF-6
G. quadratum	NIES-653	AF-6
G. viridistellatum	NIES-654	VT
Pandorina colemaniae	NIES-572	AF-6
P. morum	NIES-890	AF-6
Platydorina caudata	NIES-728	MG
Pleodorina indica	NIES-736	AF-6
P. japonica	UTEX 2523	Soilwater: GR+
P. starrii	NIES-1362	AF-6
	NIES-1363	AF-6
P. thompsonii	NIES-4126	AF-6
Vitreochlamys aulata	NIES-878	AF-6
	SAG 80.81	AF-6

V. nekrassovii	SAG 11-10	AF-6
V. ordinata	NIES-882	AF-6
Volvox africanus	NIES-863	AF-6/2
V. aureus	NIES-541	VT
V. barberi	NIES-730	AF-6
V. carteri f. *kawasakiensis*	NIES-732	VT
V. carteri f. *nagariensis*	NIES-865	MG
V. carteri f. *weismannia*	NIES-866	VT
V. dissipatrix	NIES-4128	AF-6
V. ferrisii	NIES-3986	AF-6
V. gigas	NIES-867	MG
V. globator	SAG 199.80	AF-6
V. kirkiorum	NIES-543	VT
V. obversus	NIES-868	AF-6
V. ovalis	NIES-2569	AF-6
V. powersii	NIES-4127	AF-6
V. tertius	NIES-544	AF-6
Volvulina boldii	NIES-893	MG
V. compacta	NIES-582	VT
V. pringsheimii	NIES-895	MG
V. steinii	SAG 90-1	AF-6

Table S4. 40 genes with best predicted evolutionary model under the Akaike information criterion (AIC). The range of National Center for Biotechnology Institute (NCBI) accession numbers for the assembled contigs have been provided for each gene. Gene names from De Clerck *et* al. [84] were used and unchanged. Under Evolutionary model, "I" refers to invariable sites, "G" refers to gamma-distributed rates, "F" refers to empirical frequency estimation, "LG" refers to the Le & Gascuel substitution model [89], "JTT" refers to the Jones, Taylor, and Thornton substitution model [90], and "DAYHOFF" refers to the Dayhoff substitution model [91].

Gene Name	Evolutionary model (AIC)	NCBI Accession Numbers
HOM04ULVA001181	LG+I+G+F	MZ444701 - MZ444757
HOM04ULVA001650	LG+I+G+F	MZ444758 - MZ444816
HOM04ULVA001976	JTT+I+G	MZ444817 - MZ444879
HOM04ULVA002177	JTT+I+G	MZ444880 - MZ444934
HOM04ULVA002242	DAYHOFF+G	MZ444935 - MZ444989
HOM04ULVA002252	LG+I+G	MZ444990 - MZ445052
HOM04ULVA002273	JTT+I+G	MZ445053 - MZ445114
HOM04ULVA002313	LG+I+G	MZ445115 - MZ445171
HOM04ULVA002345	JTT+I+G+F	MZ445172 - MZ445232
HOM04ULVA002356	LG+G+F	MZ445233 - MZ445294
HOM04ULVA002369	LG+I+G	MZ445295 - MZ445356
HOM04ULVA002371	LG+I+G+F	MZ445357 - MZ445418
HOM04ULVA002373	LG+I+G	MZ445419 - MZ445480
HOM04ULVA002381	LG+I+G	MZ445481 - MZ445543
HOM04ULVA002396	LG+I+G	MZ445544 - MZ445606
HOM04ULVA002447	LG+I+G	MZ445607 - MZ445668
HOM04ULVA002537	LG+G	MZ445669 - MZ445722
HOM04ULVA002538	JTT+G+F	MZ447029 - MZ447071
HOM04ULVA002544	LG+I+G+F	MZ446968 - MZ447028
HOM04ULVA002581	LG+G+F	MZ446909 - MZ446967
HOM04ULVA002624	JTT+I+G	MZ446849 - MZ446908
HOM04ULVA002640	LG+I+G	MZ446789 - MZ446848
HOM04ULVA002681	JTT+I+G	MZ446731 - MZ446788
HOM04ULVA002762	JTT+G	MZ446669 - MZ446730
HOM04ULVA002763	JTT+G+F	MZ445844 - MZ445899
HOM04ULVA002810	LG+I+G	MZ445786 - MZ445843
HOM04ULVA002814	LG+I+G+F	MZ446609 - MZ446668
HOM04ULVA002888	JTT+I+G	MZ446559 - MZ446608
HOM04ULVA002962	JTT+I+G	MZ446499 - MZ446558
HOM04ULVA002986	LG+I+G+F	MZ445900 - MZ445961
HOM04ULVA003057	JTT+I+G	MZ446025 - MZ446081
HOM04ULVA003095	LG+G	MZ445962 - MZ446024
HOM04ULVA003157	JTT+I+G	MZ446436 - MZ446498
HOM04ULVA003199	JTT+I+G	MZ446374 - MZ446435

HOM04ULVA003203	JTT+I+G	MZ446315 - MZ446373
HOM04ULVA003211	LG+G+F	MZ446256 - MZ446314
HOM04ULVA003266	JTT+G	MZ446205 - MZ446255
HOM04ULVA003755	LG+I+G+F	MZ446143 - MZ446204
HOM04ULVA003786	LG+I+G+F	MZ446082 - MZ446142
HOM04ULVA004274	LG+I+G	MZ445723 - MZ445785

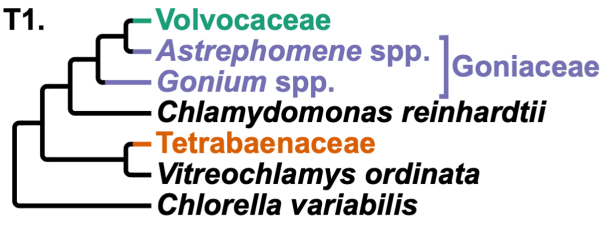
Table S5. National Center for Biotechnology Institute (NCBI) accession numbers for BioProject PRJNA701495. "NIES" refers to National Institute for Environmental Studies (NIES, Japan), "SAG" refers to Culture Collection of Algae at the University of Göttingen (SAG, Germany), and "UTEX" refers to The Culture Collection of Algae at the University of Texas at Austin (UTEX, USA).

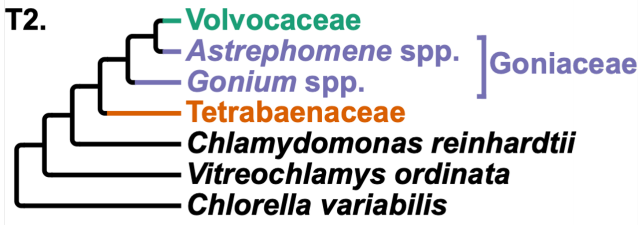
Taxa	Collection ID	NCBI Accession Number
Astrephomene gubernaculifera	NIES-418	SRR13719267
A. perforata	NIES-564	SRR13719264
Basichlamys sacculifera	NIES-566	SRR13719253
Colemanosphaera. charkowiensis	NIES-3383	SRR13719279
Chlamydomonas debaryana	SAG 11-55a	SRR13719240
	SAG 70.81	SRR13719274
C. globosa	SAG 81.72	SRR13719238
C. moewusii	SAG 11-16f	SRR13719241
C. schloesseri	SAG 2486	SRR13719237
Eudorina cylindrica	NIES-722	SRR13719256
E. elegans	NIES-456	SRR13719287
	NIES-458	SRR13719286
	NIES-568	SRR13719262
	NIES-717	SRR13719259
	NIES-719	SRR13719258
	NIES-720	SRR13719257
E. illinoisensis	NIES-460	SRR13719266
E. minodii	NIES-856	SRR13719250
E. peripheralis	NIES-725	SRR13719255
E. unicocca	SAG 24-1c	SRR13719239
Gonium multicoccum	NIES-737	SRR13719234
G. octonarium	NIES-851	SRR13719233
G. quadratum	NIES-653	SRR13719260
G. viridistellatum	NIES-654	SRR13719236
Pandorina colemaniae	NIES-572	SRR13719242
P. morum	NIES-890	SRR13719246
Platydorina caudata	NIES-728	SRR13719254
Pleodorina indica	NIES-736	SRR13719251
P. japonica	UTEX 2523	SRR13719268
P. starrii	NIES-1362	SRR13719282
	NIES-1363	SRR13719281
P. thompsonii	NIES-4126	SRR13719277
Vitreochlamys aulata	NIES-878	SRR13719285
	SAG 80.81	SRR13719273
V. nekrassovii	SAG 11-10	SRR13719243
V. ordinata	NIES-882	SRR13719284
Volvox africanus	NIES-863	SRR13719270
V. aureus	NIES-541	SRR13719265
V. barberi	NIES-730	SRR13719252

V. carteri f. *kawasakiensis*	NIES-732	SRR13719235
V. carteri f. *nagariensis*	NIES-865	SRR13719249
V. carteri f. *weismannia*	NIES-866	SRR13719248
V. dissipatrix	NIES-4128	SRR13719244
V. ferrisii	NIES-3986	SRR13719278
V. gigas	NIES-867	SRR13719247
V. globator	SAG 199.80	SRR13719271
V. kirkiorum	NIES-543	SRR13719263
V. obversus	NIES-868	SRR13719269
V. ovalis	NIES-2569	SRR13719280
V. powersii	NIES-4127	SRR13719276
V. tertius	NIES-544	SRR13719275
Volvulina boldii	NIES-893	SRR13719245
V. compacta	NIES-582	SRR13719261
V. pringsheimii	NIES-895	SRR13719283
V. steinii	SAG 90-1	SRR13719272

Figure S1. Approximately Unbiased (AU) tests comparing key hypotheses for our 40-protein concatenated dataset. Each AU test used all 68 taxa from our study, but trees shown have been truncated so as to be easily viewed/read. T1 always represents the branching order of the unconstrained tree, and all other trees are constrained trees. Hypotheses tested: (A) Monophyly of the colonial volvocine algae, (B) Monophyly of Goniaceae, (C) Section *Volvox* sister to *Colemanosphaera* or PVC clade.

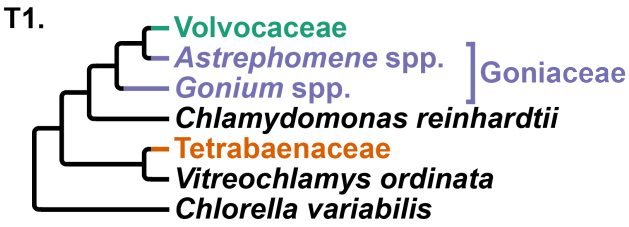
A. **Monophyly of the colonial volvocine algae**

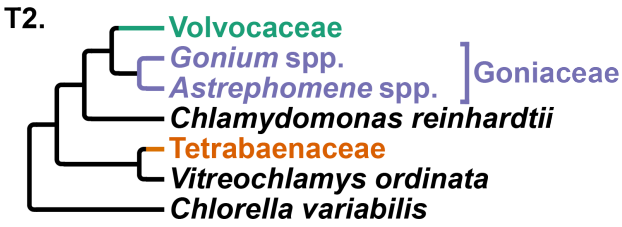




Tree	logL	deltaL	p-AU
1	-308397.5847	0	1.0 +
2	-308653.5707	255.99	2.82e-38 -

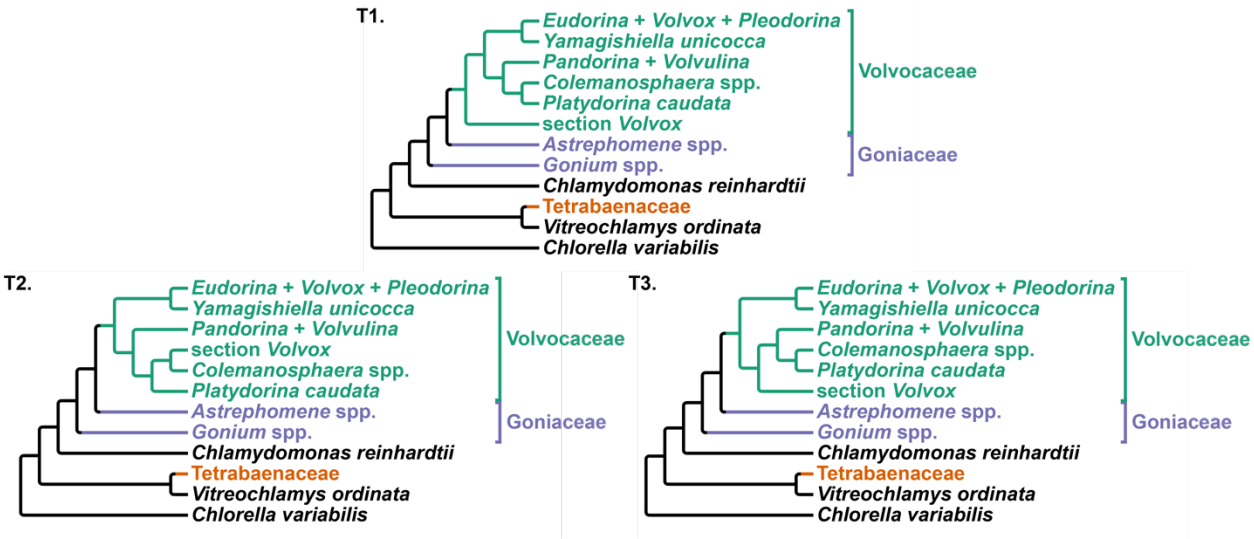
B. **Monophyly of Goniaceae**





Tree	logL	deltaL	p-AU
1	-308397.5785	0	0.955 +
2	-308429.6318	32.053	0.0446 -

C. **Section *Volvox* sister to *Colemanosphaera* or PVC clade**



Tree	logL	deltaL	p-AU
1	-308397.5847	0	0.967 +
2	-309045.8233	648.24	4.64e-88 -
3	-308434.3051	36.72	0.0332 -

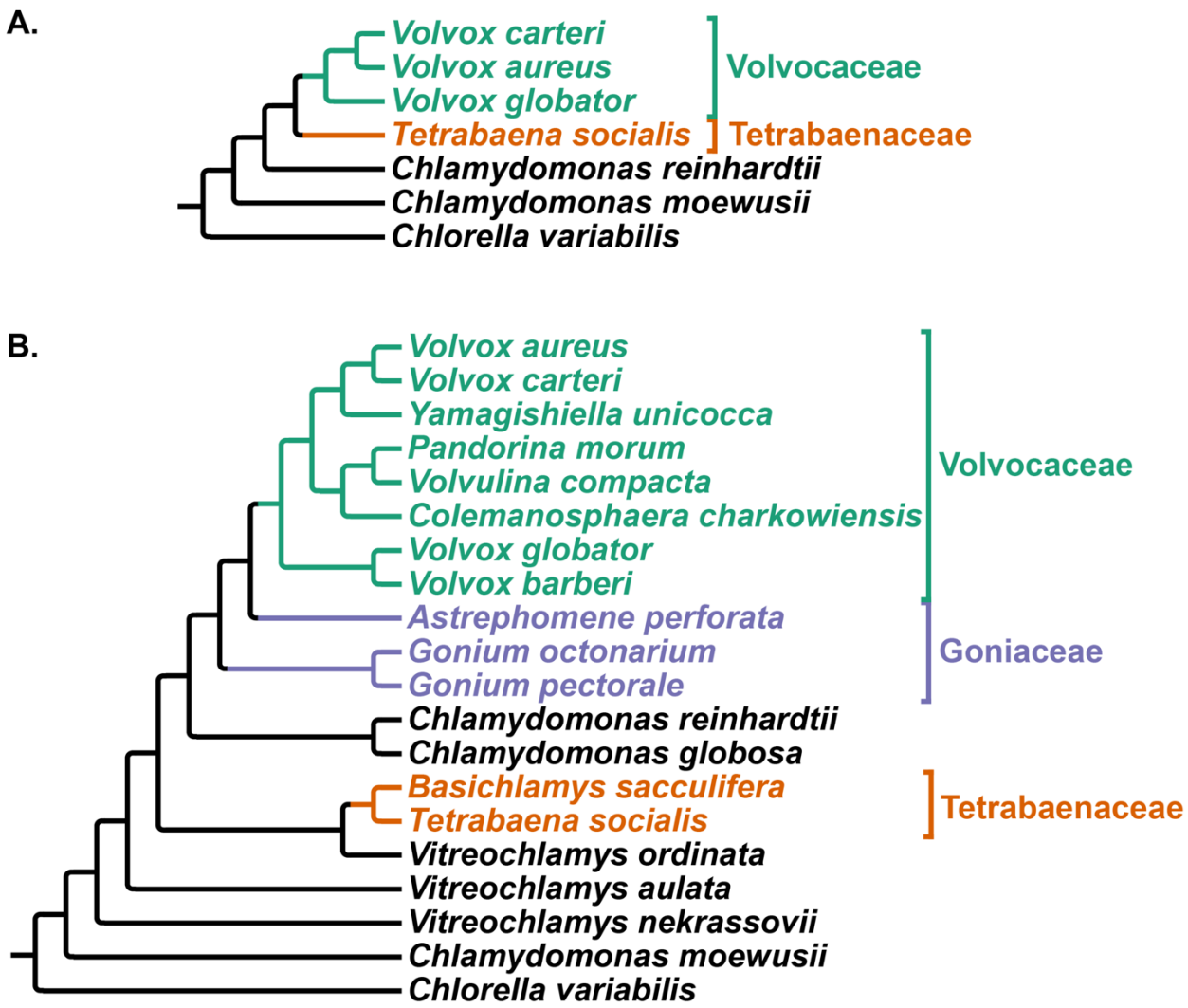


Figure S2. (A) Phylogeny that represents a replication of the Zhang *et al.* [51] results as they relate to the volvocine algae. The phylogenetic tree shown is based on 12,650 aligned amino acid positions among 7 taxa, and it was inferred using the maximum likelihood method in IQtree. **(B)** Phylogeny of the volvocine algae that represents a change in the branching order once more volvocine taxa are sampled for phylogenetic inference. This phylogeny shows the non-monophyly of the colonial algae and the Tetrabaenaceae sister to *Vitreochlamys ordinata*. The phylogenetic tree shown is based on 12,650 aligned amino acid positions among 20 taxa, and it was inferred using the maximum likelihood method in IQtree.

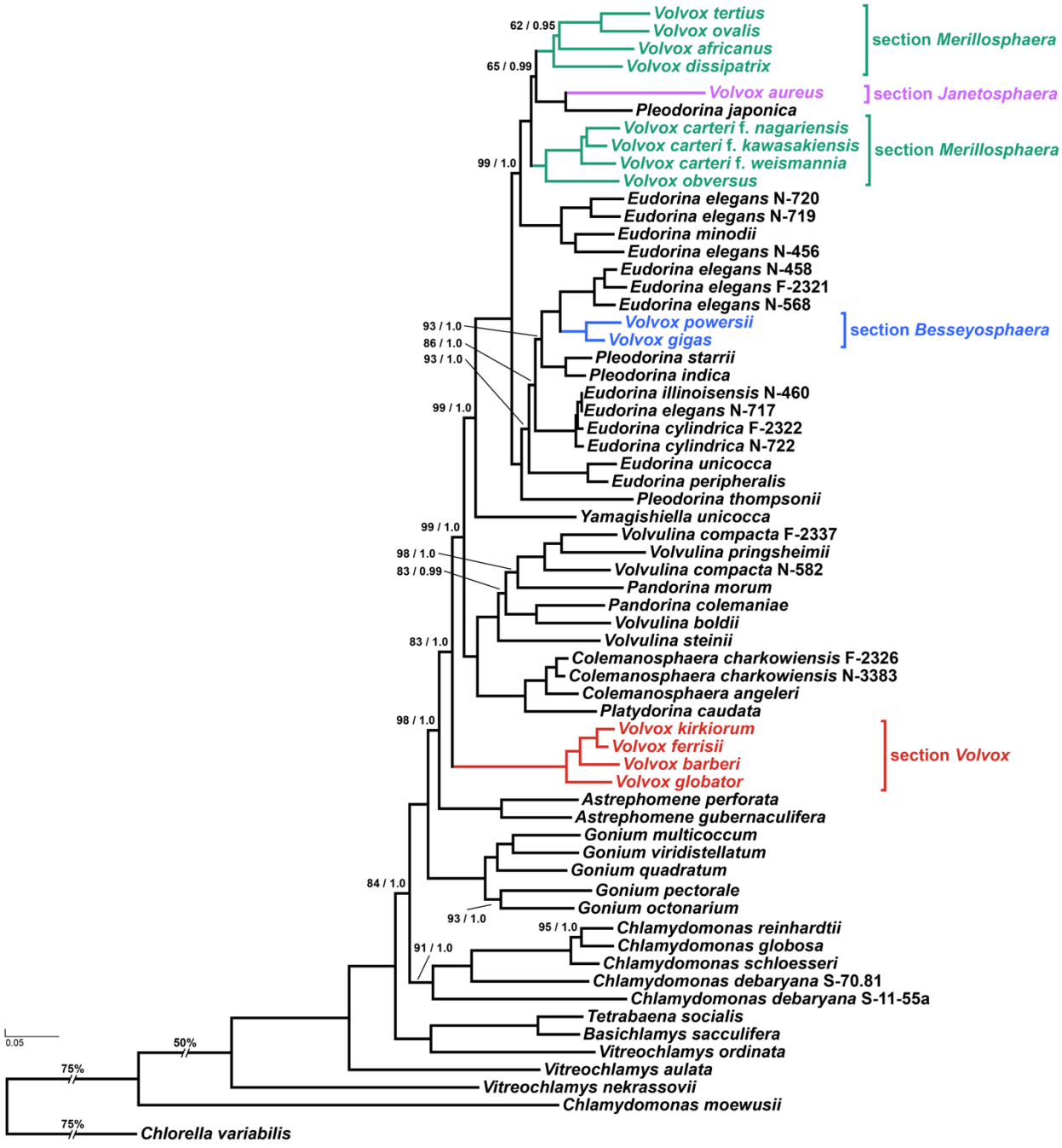


Figure S3. Phylogeny of the volvocine algae which highlights the four sections of genus *Volvox* as recognized by Nozaki *et al.* [35]: *Besseyosphaera* (blue), *Janetosphaera* (purple), *Merrillosphaera* (green), and *Volvox* (red). The phylogenetic tree shown is based onv12,650 aligned amino acid positions of 68 taxa inferred using the maximum likelihood method in IQtree. Numbers on branches represent bootstrap values and Bayesian posterior probabilities, respectively. Branch lengths correspond to genetic divergence, as indicated by the scale bar. The

Tetrabaenaceae, Goniaceae, and Volvocaceae are indicated by the orange, purple, and green highlight, respectively.

APPENDIX B: SUPPLEMENTARY INFORMATION 2

Appendix B.1 RNA extraction procedures

Two RNA isolation protocols were followed for total RNA extraction: TRizol RNeasy and QIAGEN RNeasy Plant Mini Kit.

TRIZOL RNEASY PROTOCOL Cells/Colonies were pelleted and then resuspended in 3 mL 2% SDS in tris-buffered saline and 3 mL TRizol (guanidinium thiocyanate and acid phenol) in a 15 mL conical centrifuge tube and vortexed for approximately 30 seconds. Cells/Colonies were then flash frozen in liquid N_2. After sample was thawed, a phase separation was performed by centrifugation at 3000 *x g* for 10 minutes at room temperature. The aqueous phase was transferred to a 15 mL QIAGEN MaXtract High Density conical tube (prior to transferring the aqueous phase, the gel in the MaXtract tube was pelleted by centrifugation at 1500 *x g* for 1 minute). 0.6 mL chloroform was added to the MaXtract tube which was subsequently capped and shaken for approximately 15 seconds. After shaking, the MaXtract tube was left in a fume hood at room temperature for 3 minutes and then centrifuged at 1500 x g for 5 minutes at 4^{o}C to perform another phase separation. The aqueous phase was transferred to a 15 mL conical tube and 1.5 mL of isopropanol was added to the aqueous phase. The sample was left to incubate for 10 minutes at room temperature, and then another centrifugation step at 3000 *x g* for 10 minutes was performed. After centrifugation the supernatant was discarded, and the pellet was washed with 1 mL 75% ethanol. The ethanol was carefully pipetted off and the pellet was allowed to airdry for 5-10 minutes. The RNeasy mini spin column was centrifuged at 10,000 rpm or 10,976 x g for 15 seconds, followed by DNase digestion. Flow-through from the last step was discarded, and 350 µL of Buffer RW1 was added to the RNeasy Mini spin column and centrifuged at

10,000 rpm or 10,976 x g for 15 seconds. Flow-through was discarded. To a 70 µL aliquot of Buffer RDD was added 10 µL of QIAGEN RNase-free DNase. The resulting 80 µL Buffer RDD-DNase mixture was pipetted directly to the membrane of the RNeasy mini spin column and allowed to incubate for 15 minutes before proceeding with the protocol. After the incubation period, 350 µL of Buffer RW1 was added to the RNeasy spin column and centrifuged at 10,000 rpm or 10,976 x g for 15 seconds. Flow-through was discarded. 500 µL Buffer RPE was added to the RNeasy spin column and centrifuged at 10,000 rpm or 10,976 x g for 15 seconds. Flow-through was discarded. A second wash with 500 µL Buffer RPE was performed, and the RNeasy spin column was centrifuged at 10,000 rpm or 10,976 x g for 2 minutes. Flow-through was discarded, and the RNeasy spin column was placed inside a 1.5 mL collection tube supplied by QIAGEN. 30 µL of sterile RNase-free water was pipetted directly on the membrane and allowed to incubate for 1 minute, after which the RNeasy spin column was centrifuged at 10,000 rpm or 10,976 x g for 1 minute. RNA concentration and A_{260}/A_{280} ratios were measured via Nanodrop (Thermo Fisher Scientific, Waltham, MA 02451 USA), and if RNA concentrations were >50 ng/µl then another elution step with 30 µL of RNase-free water was performed.

QIAGEN PLANT MINI KIT PROTOCOL Cells/Colonies were pelleted by centrifugation, resuspended in 500 µL of RLT Buffer RLT (β-mercaptoethanol was added to the Buffer RLT solution prior to starting) then vortexed for ca. 30 seconds. Resuspended cells/colonies were flash frozen in liquid N_2, then thawed and centrifuged at 10,000 rpm or 10,976 x g for 2 minutes in a QIAshredder spin column. The resulting lysate and any precipitate were transferred to a collection tube provided in the QIAGEN kit, whereafter 200 µL of 100% ethanol was added to the lysate and immediately mixed and transferred to an RNeasy Mini spin column. The RNeasy mini spin column was centrifuged at 10,000 rpm or 10,976 x g for 15 seconds, after which the

optional DNase digestion step was performed. The flow-through from the last step was discarded, and 350 μL of Buffer RW1 was added to the RNeasy Mini spin column and centrifuged at 10,000 rpm or 10,976 x g for 15 seconds. Flow-through was discarded. A 70 μL aliquot of Buffer RDD was prepared and 10 μL of QIAGEN RNase-free DNase was added to it. The 80 μL Buffer RDD-DNase mixture was pipetted directly to the membrane of the RNeasy mini spin column and allowed to incubate for 15 minutes before the protocol was continued. Following incubation. 350 μL of Buffer RW1 was added to the RNeasy spin column, which was centrifuged at 10,000 rpm or 10,976 x g for 15 seconds. Flow-through was discarded. Then, 500 μL Buffer RPE was added to the RNeasy spin column and centrifuged at 10,000 rpm or 10,976 x g for 15 seconds. Flow-through was discarded. A second wash with 500 μL RPE buffer was performed, and the RNeasy spin column was centrifuged at 10,000 rpm or 10,976 x g for 2 minutes. Flow-through was discarded, and the RNeasy spin column was placed inside a 1.5 mL collection tube supplied by QIAGEN. 30 μL of RNase-free water was pipetted directly on the membrane and allowed to incubate for 1 minute. The RNeasy spin column was centrifuged at 10,000 rpm or 10,976 x g for 1 minute. RNA concentration and A_{260}/A_{280} ratios were measured via Nanodrop. If RNA concentrations were >50 ng/μl, then another elution step with 30 μL of RNase-free water was performed.

Appendix B.2 Confirming the conspecificity of Tetrabaena socialis N-571 and N-691

To test whether *Tetrabaena socialis* N-571 and N-691 are conspecific, we examined morphological, sequence, and phylogenetic data. Nozaki and Ohtani [52] examined *T. socialis* N-691 and found that N-691 was morphologically identical to *T. socialis* N-571 except for forming slightly larger colonies. Next, we compared ITS-1, 5.8S rRNA, and ITS-2 sequences from Mai and Coleman [53] between the two *T. socialis* strains; we found that they were 92% identical. Then, we performed a BLAST search of the aforementioned sequences: the first ten BLAST hits all matched the same sequences from *T. socialis* strains with 100% query coverage. Using the BLAST software, neighbor joining and fast minimum evolution distance trees were inferred from the BLAST results; the trees produced show *T. socialis* strains N-571 and N-691 to be reciprocally monophyletic.

APPENDIX C: SUPPLEMENTARY INFORMATION 3

```
#!/usr/bin/env python3

# The following script is used for a single taxon.
# Purpose of script: Read BLAST XML output files to get specific hit information (e.g. information regarding first BLAST hit, second BLAST hit, etc.), use that BLAST information to locate the scaffold of a gene we are wanting to find, translate the scaffold in all reading frames, and then add the entire scaffold in the appropriate reading frame to an alignment file.
# Steps of the program:
# Step 1: Parse BLAST XML output files to get specific hit information. This program obtains three pieces of information from BLAST output: 1) "Query ID" which corresponds to the name of the query sequence 2) "Hit ID" which corresponds to the specific scaffold name in the transcriptome file where gene is located 3) "Hsp hit hseq" which corresponds to the sequence that best matches the query sequence.
# Step 2: Use the first hit id from BLAST XML file to locate the scaffold in the transcriptome file.
# Step 3: Translate the ENTIRE scaffold sequence in all three forward and all three reverse reading frames.
# Step 4: Locate the "Hsp hit hseq" in one of the 6 reading frames.
# Step 5: Write the original query name + query seq to a file with the hit organism name + hit sequence in a uniquely named directory.
# print commands (commented out) have been given throughout so that the process of the script can be seen in terminal.

import os
import xml.etree.ElementTree as ET  # Imports xml.etree.ElementTree which allows you to easily parse an XML file.
#  Regular expressions cannot (correctly) parse an XML file though I have seen it done. See famous post on StackOverflow for more information:
https://stackoverflow.com/questions/1732348/regex-match-open-tags-except-xhtml-self-contained-tags/1732454#1732454
#  XML docs have sections called 'elements' which are defined by a beginning '<' and ending '>' tag.
#  The information contained within the tags '<info>' are called 'child elements'.
#  'root' is the main header at the top. It is called 'root' because the rest of the tree is rooted to it.
#  Subelements of the tree are called 'children'.
#  Sometimes tags will only contain text and you have to call it in your script as such.
#  If you want to navigate to a specific section of the XML file just type out the path.
from Bio.Seq import Seq  # Be sure to have biopython. Seq object allows you use Python strings with biological methods.
from Bio.Alphabet import generic_dna
```

```
path_to_vca_de_clerck = ""  # Path to Volvox carteri amino acid seqs that were shared with us by the De Clerck team.
vca_seqs_de_clerck = os.listdir(path_to_vca_de_clerck)

path_to_allxmltblastn_results = ""  # Path to the tblastn XML files that contain all BLAST hits for taxon.
xml_dir = os.listdir(path_to_allxmltblastn_results)

path_to_transcriptome = "" # PATH to appropriate transcriptome where scaffold ID will be located.
transcriptomefile = open(path_to_transcriptome, 'r')

output_dir = ""  # This creates the main output directory where the vca seqs and scaffolds translated in the correct reading frame will be output to.
if not os.path.exists(output_dir):  # This is checking to see if the path already exists. If it does not...
   os.makedirs(output_dir)  # ... then the directory that I named and assigned to a variable will be created.

# Read in tblastn results & identify top hit in tblastn results for all xml files (40) per species.
Query_Seq_id_list = []  # This list will be used later to create a dictionary: key= query_seq_id, value= first_hit_id
First_Hit_id_list = []  # This list will be used during the step where we search the transcriptome.
First_Hsp_hseq_list =[]  # This list will be used later when locating the correct reading frame translation.
for file in xml_dir:
   #print(file)
   if file.endswith(".xml"):
      xml_file_path = os.path.join(path_to_allxmltblastn_results, file)
      #print(xml_file_path)
      xml_files = open(xml_file_path, 'r')
      for lines in xml_files:
         #print(lines)
         tree = ET.parse(xml_files)  # ElementTree is now parsing all of the XML tblastn files.
         root = tree.getroot()  # This defines the root for each of the XML files. Since all the files have an identical format the root of each file is 'BlastOutput' (for my data, at least).
         #print(root.tag)

         for query_seq_id in root.findall(""):  # This for loop is capturing the query seq id that is associated to the De Clerck seq. In the XML file this corresponds to "<Iteration_query-def>" Under "<Iteration>".
            #print("query_seq_id =", query_seq_id.text)
```

```
            Query_Seq_id = query_seq_id.text
            Query_Seq_id_list.append(Query_Seq_id)

        for first_hit_id in root.findall(""):  # This for loop is capturing the Hit_id in the XML path of the first "Hit". This is specified by the "Hit[1] in the path. Example: "./BlastOutput_iterations/Iteration/Iteration_hits/Hit[1]/Hit_id"
            fill = first_hit_id.text  # Due to the format of the XML file the information contained within the tags are all 'text'. You must specify for the parser that the format of the info is 'text'.
            #fill2 = fill.replace('dbj', '')  # This and the replace below were added due to the Hit_id of one of the names having these characters added when they were not present in the normal tblastn results.
            #fill3 = fill2.replace('|', '')
            #First_Hit_id = fill3
            #print('First_Hit_id =', First_Hit_id)
            First_Hit_id_list.append(First_Hit_id)  # The for loop above where 'First_Hit_id' is captured only stores the final value when the program finishes its last loop. This step was necessary because we needed to store those values in a list in order to iterate over them.

        for first_hsp_hseq in root.findall(""):  # This for loop is capturing the first Hsp_seq in the path of the first Hit and first Hsp (this needs to be explicit because there can be multiple Hsp_hseqs in a BLAST hit). Example: "./BlastOutput_iterations/Iteration/Iteration_hits/Hit[1]/Hit_hsps/Hsp[1]/Hsp_hseq".
            filler = first_hsp_hseq.text
            replace_dashes = filler.replace('-', '')
            First_Hsp_hseq = replace_dashes
            #print('First_Hsp_hseq=', First_Hsp_hseq)
            First_Hsp_hseq_list.append(First_Hsp_hseq)

#print('first_hseq_list =', First_Hsp_hseq_list)

# Grab top hit scaffold ID from tblastn results and trace it back to the scaffold in the transcriptome.
frameF1_trans_list = []
frameF2_trans_list = []
frameF3_trans_list = []
frameR1_trans_list = []
frameR2_trans_list = []
frameR3_trans_list = []

for scaffold in transcriptomefile:
   if any(item in scaffold for item in First_Hit_id_list):
      contig_id = scaffold
      contig_nucl_seq = next(transcriptomefile)
      #print(contig_id)
      #print(contig_nucl_seq)
```

```
# Translate the top hit scaffold in all 6 reading frames.
# 5'-3' ORFs.
scaffold_seq = Seq(contig_nucl_seq, generic_dna)
#print("scaffold_seq =", scaffold_seq)
frameF1 = scaffold_seq
frameF2 = scaffold_seq[1:]
frameF3 = scaffold_seq[2:]
# print(frameF1)
# print(frameF2)
# print(frameF3)
frameF1_trans = frameF1.rstrip().translate()
frameF2_trans = frameF2.rstrip().translate()
frameF3_trans = frameF3.rstrip().translate()

frameF1_trans_list.append(frameF1_trans)
frameF2_trans_list.append(frameF2_trans)
frameF3_trans_list.append(frameF3_trans)

#print("frameF1_trans =", frameF1_trans)
#print("frameF2_trans =", frameF2_trans)
#print("frameF3_trans =", frameF3_trans)
# Reverse complement ORFs.
complement_seq = scaffold_seq.complement()
# print(complement_seq)
rev_compliment = scaffold_seq.reverse_complement()
# print("rev_compliment =", rev_compliment)

frameR1 = rev_compliment[1:]
frameR2 = rev_compliment[2:]
frameR3 = rev_compliment[3:]

# print("frameR1 =", frameR1)
# print("frameR2 =", frameR2)
# print("frameR3 =", frameR3)
frameR1_trans = frameR1.rstrip().translate()
frameR2_trans = frameR2.rstrip().translate()
frameR3_trans = frameR3.rstrip().translate()

frameR1_trans_list.append(frameR1_trans)
frameR2_trans_list.append(frameR2_trans)
frameR3_trans_list.append(frameR3_trans)

#print("frameR1_trans =", frameR1_trans)
#print("frameR2_trans =", frameR2_trans)
#print("frameR3_trans =", frameR3_trans)
```

```
    # Use top tblastn hit to ID correct translation out of the 6 reading frames.

    six_frame_translation_list = [frameF1_trans_list, frameF2_trans_list,
frameF3_trans_list, frameR1_trans_list, frameR2_trans_list, frameR3_trans_list]

correct_frame_translation_list = []
for frame in six_frame_translation_list:
  for item in frame:
    if any(sbs in item for sbs in First_Hsp_hseq_list):  # I used an arbitrary 'sbs' for
'substring' during this 'if statement' because I was finding a substring within a string (i.e. a
partial sequence within a full sequence).
      #print('item =', item)
      correct_frame_translation_list.append(item)

#print('correct_frame_translation_list =', correct_frame_translation_list)
dictionary_Query_seq_and_Hit_id = dict(zip(Query_Seq_id_list, First_Hit_id_list))
dictionary_Hit_id_and_Hsp_hseq = dict(zip(First_Hit_id_list, First_Hsp_hseq_list))  #
Dictionary contains keys (First_Hit_id) and its associated values (First_Hsp_hseq).
#print('dictionary =', dictionary_Hit_id_and_Hsp_hseq)

# This section takes the dictionary above (dictionary_Hit_id_and_Hsp_hseq) and replaces
the incomplete First_Hsp_hseqs with the complete translated contig sequence.
for key, value in dictionary_Hit_id_and_Hsp_hseq.items():
  #print('key =', key)
  #print('value =', value)
  for trans_seq in correct_frame_translation_list:
    if value in trans_seq:
      dictionary_Hit_id_and_Hsp_hseq[key] = trans_seq
      #print('new_dic =', dictionary_Hit_id_and_Hsp_hseq)
      dictionary_Hit_id_and_correct_frame_translation =
dictionary_Hit_id_and_Hsp_hseq

#print('dict_w_query_and_hit_ids =', dictionary_Query_seq_and_Hit_id)
#print('dict_w_correct_frame_translations =',
dictionary_Hit_id_and_correct_frame_translation)

# Read into all the Volvox carteri seqs from De Clerck and append them to outfiles.
for file in vca_seqs_de_clerck:
  #print(file)
  if file.endswith(".fasta"):
    vca_seqs_path = os.path.join(path_to_vca_de_clerck, file)
    vca_seqs = open(vca_seqs_path, 'r')
    with open(os.path.join(output_dir, file), 'a') as output:
      for lines in vca_seqs:
        output.write(lines)
```

```
# Next two main For Loops are appending the correct scaffold name and its translated
sequence in the correct reading frame to the same file with the De Clerck seq.
for filenames in os.listdir(output_dir):
    path_output = os.path.join(output_dir, filenames)
    #print(path_output)
    files_in_ouput = open(path_output, 'r')
    with open(path_output, 'a') as output2:
        for info in files_in_ouput:
            #print(info)
            for k, v in dictionary_Query_seq_and_Hit_id.items():
                if k in info:
                    output2.write(''+v+'\n')

for filenames in os.listdir(output_dir):
    path_output = os.path.join(output_dir, filenames)
    #print(path_output)
    files_in_ouput = open(path_output, 'r')
    with open(path_output, 'a') as output2:
        for info in files_in_ouput:
            for scaffold_id, translated_seq in
dictionary_Hit_id_and_correct_frame_translation.items():
                if scaffold_id in info:
                    #print(str(translated_seq))
                    output2.write(str(translated_seq))

# Final step is to rename all files in output dir so that each retains their gene identifier but
also states which organism each file belongs to.
for file_names in os.listdir(output_dir):
    new_filenames = file_names.replace('', '') # All files have '_aln' in name. This is finding it
and replacing it with the species identifier and what the file is.
    os.rename(os.path.join(output_dir, file_names), os.path.join(output_dir, new_filenames))
```

REFERENCES

1. Grosberg RK, Strathmann RR (2007) The Evolution of Multicellularity: A Minor Major Transition? Annu Rev Ecol Evol Syst 38:621–654

2. Maynard Smith J, Szathmáry E (1995) *The major transitions in evolution*. Oxford University Press

3. Szathmáry E, Maynard Smith J (1995) The major evolutionary transitions. Nature 374:227–232

4. West SA, Fisher RM, Gardner A, Kiers ET (2015) Major evolutionary transitions in individuality. Proc Natl Acad Sci 112:10112–10119

5. Hallmann A (2011) Evolution of reproductive development in the volvocine algae. Sex Plant Reprod 24:97–112

6. Wahl ME, Murray AW (2016) Multicellularity makes somatic differentiation evolutionarily stable. Proc Natl Acad Sci 113:8362–8367

7. Chen L, Xiao S, Pang K, Zhou C, Yuan X (2014) Cell differentiation and germ–soma separation in Ediacaran animal embryo-like fossils. Nature 516:238–241

8. Han T, Runnegar B (1992) Megascopic eukaryotic algae from the 2.1-billion-year-old negaunee iron-formation, Michigan. Science 257:232–235

9. Schirrmeister BE, de Vos JM, Antonelli A, Bagheri HC (2013) Evolution of multicellularity coincided with increased diversification of cyanobacteria and the Great Oxidation Event. Proc Natl Acad Sci 110:1791–1796

10. Herron MD, Borin JM, Boswell JC, Walker J, Chen I-CK, Knox CA, Boyd M, Rosenzweig F, Ratcliff WC (2019) De novo origins of multicellularity in response to predation. Sci Rep 9:2328

11. Quintero-Galvis JF, Paleo-López R, Solano-Iguaran JJ, Poupin MJ, Ledger T, Gaitan-Espitia JD, Antoł A, Travisano M, Nespolo RF (2018) Exploring the evolution of multicellularity in *Saccharomyces cerevisiae* under bacteria environment: An experimental phylogenetics approach. Ecol Evol 8:4619–4630

12. Ratcliff WC, Denison RF, Borrello M, Travisano M (2012) Experimental evolution of multicellularity. Proc Natl Acad Sci 109:1595–1600

13. Ratcliff WC, Herron MD, Howell K, Pentz JT, Rosenzweig F, Travisano M (2013) Experimental evolution of an alternating uni- and multicellular life cycle in *Chlamydomonas reinhardtii*. Nat Commun 4:1–7

14. Herron MD, Michod RE (2008) Evolution of complexity in the volvocine algae: transitions in individuality through Darwin's eye. Evolution 62:436–451

15. Kirk DL (2005) A twelve-step program for evolving multicellularity and a division of labor. BioEssays 27:299–310

16. Koyanagi KO (2015) Inferring Cell Differentiation Processes Based on Phylogenetic Analysis of Genome-Wide Epigenetic Information: Hematopoiesis as a Model Case. Genome Biol Evol 7:699–705

17. van Leeuwenhoek A (1700) Part of a letter from Mr Antony van Leeuwenhoek, concerning the worms in sheeps livers, gnats, and animalcula in the excrements of frogs. Philos Trans R Soc Lond 22:509–518

18. Kirk DL (2001) Germ–Soma Differentiation in *Volvox*. Dev Biol 238:213–223

19. Schmitt R (2003) Differentiation of germinal and somatic cells in *Volvox carteri*. Curr Opin Microbiol 6:608–613

20. Craig RJ, Hasan AR, Ness RW, Keightley PD (2021) Comparative genomics of *Chlamydomonas*. Plant Cell. https://doi.org/10.1093/plcell/koab026

21. Featherston J, Arakaki Y, Hanschen ER, Ferris PJ, Michod RE, Olson BJSC, Nozaki H, Durand PM (2018) The 4-Celled *Tetrabaena socialis* Nuclear Genome Reveals the Essential Components for Genetic Control of Cell Number at the Origin of Multicellularity in the Volvocine Lineage. Mol Biol Evol 35:855–870

22. Hanschen ER, Marriage TN, Ferris PJ, et al (2016) The Gonium pectorale genome demonstrates co-option of cell cycle regulation during the evolution of multicellularity. Nat Commun 7:1–10

23. Merchant SS, Prochnik SE, Vallon O, et al (2007) The *Chlamydomonas* Genome Reveals the Evolution of Key Animal and Plant Functions. Science 318:245–250

24. Prochnik SE, Umen J, Nedelcu AM, et al (2010) Genomic Analysis of Organismal Complexity in the Multicellular Green Alga *Volvox carteri*. Science 329:223–226

25. Herron MD, Hackett JD, Aylward FO, Michod RE (2009) Triassic origin and early radiation of multicellular volvocine algae. Proc Natl Acad Sci 106:3254–3258

26. Brunet T, King N (2017) The Origin of Animal Multicellularity and Cell Differentiation. Dev Cell 43:124–140

27. Lang NJ (1963) Electron Microscopy of the Volvocaceae and Astrephomenaceae. Am J Bot 50:280–300

28. Larson A, Kirk MM, Kirk DL (1992) Molecular Phylogeny of the Volvocine Flagellates. Mol Biol Evol 9:85–105

29. Hanschen ER, Herron MD, Wiens JJ, Nozaki H, Michod RE (2018) Multicellularity Drives the Evolution of Sexual Traits. Am Nat 192:E93–E105

30. Herron MD, Desnitskiy AG, Michod RE (2010) Evolution of developmental programs in *Volvox* (Chlorophyta). J Phycol 46:316–324

31. Hanschen ER, Herron MD, Wiens JJ, Nozaki H, Michod RE (2018) Repeated evolution and reversibility of self-fertilization in the volvocine green algae*. Evolution 72:386–398

32. Nakada T, Tomita M, Nozaki H (2010) *Volvulina compacta* (Volvocaceae, Chlorophyceae), new to Japan, and its phylogenetic position. J Jpn Bot 85:364–369

33. Nozaki H, Misawa K, Kajita T, Kato M, Nohara S, Watanabe MM (2000) Origin and Evolution of the Colonial Volvocales (Chlorophyceae) as Inferred from Multiple, Chloroplast Gene Sequences. Mol Phylogenet Evol 17:256–268

34. Nozaki H, Yamada TK, Takahashi F, Matsuzaki R, Nakada T (2014) New "missing link" genus of the colonial volvocine green algae gives insights into the evolution of oogamy. BMC Evol Biol 14:1–11

35. Nozaki H, Matsuzaki R, Yamamoto K, Kawachi M, Takahashi F (2015) Delineating a New Heterothallic Species of *Volvox* (Volvocaceae, Chlorophyceae) Using New Strains of "*Volvox africanus*." PLOS ONE 10:e0142632

36. Grochau-Wright ZI, Hanschen ER, Ferris PJ, Hamaji T, Nozaki H, Olson BJSC, Michod RE (2017) Genetic basis for soma is present in undifferentiated volvocine green algae. J Evol Biol 30:1205–1218

37. Pröschold T, Darienko T, Krienitz L, Coleman AW (2018) *Chlamydomonas schloesseri* sp. nov. (Chlamydophyceae, Chlorophyta) revealed by morphology, autolysin cross experiments, and multiple gene analyses. Phytotaxa 362:21–38

38. Nakada T, Tsuchida Y, Tomita M (2019) Improved taxon sampling and multigene phylogeny of unicellular chlamydomonads closely related to the colonial volvocalean lineage Tetrabaenaceae-Goniaceae-Volvocaceae (Volvocales, Chlorophyceae). Mol Phylogenet Evol 130:1–8

39. Coleman AW (1999) Phylogenetic analysis of "Volvocacae" for comparative genetic studies. Proc Natl Acad Sci 96:13892–13897

40. Hu Y, Xing W, Song H, Zhu H, Liu G, Hu Z (2019) Evolutionary Analysis of Unicellular Species in Chlamydomonadales Through Chloroplast Genome Comparison With the Colonial Volvocine Algae. Front Microbiol 10:1351

41. Dong W, Liu J, Yu J, Wang L, Zhou S (2012) Highly Variable Chloroplast Markers for Evaluating Plant Phylogeny at Low Taxonomic Levels and for DNA Barcoding. PLoS ONE 7:e35071

42. Jakob SS, Blattner FR (2006) A Chloroplast Genealogy of Hordeum (Poaceae): Long-Term Persisting Haplotypes, Incomplete Lineage Sorting, Regional Extinction, and the Consequences for Phylogenetic Inference. Mol Biol Evol 23:1602–1612

43. Xu B, Wu N, Gao X-F, Zhang L-B (2012) Analysis of DNA sequences of six chloroplast and nuclear genes suggests incongruence, introgression, and incomplete lineage sorting in the evolution of *Lespedeza* (Fabaceae). Mol Phylogenet Evol 62:346–358

44. Rose J, Toledo CA, Lemmon EM, Lemmon AR, Sytsma K (2020) Out of Sight, Out of Mind: Widespread Nuclear and Plastid-Nuclear Discordance in the Flowering Plant Genus *Polemonium* (Polemoniaceae) Suggests Widespread Historical Gene Flow Despite Limited Nuclear Signal. Syst Biol 70:162–180

45. Soltis DE, Kuzoff RK (1995) Discordance between nuclear and chloroplast phylogenies in the *Heuchera* group (Saxifragaceae). Evolution 49:727–742

46. Yu W-B, Huang P-H, Li D-Z, Wang H (2013) Incongruence between Nuclear and Chloroplast DNA Phylogenies in *Pedicularis* Section *Cyathophora* (Orobanchaceae). PLoS ONE 8:e74828

47. Hu Y, Xing W, Song H, Hu Z, Liu G (2020) Comparison of colonial volvocine algae based on phylotranscriptomic analysis of gene family evolution and natural selection. Eur J Phycol 55:100–112

48. Nakazawa A, Krienitz L, Nozaki H (2001) Taxonomy of the unicellular green algal genus *Vitreochlamys* (Volvocales), based on comparative morphology of cultured material. Eur J Phycol 36:113–128

49. Nozaki H (1996) Morphology and evolution of sexual reproduction in the Volvocaceae (Chlorophyta). J Plant Res 109:353–361

50. Nozaki H, Itoh M (1994) Phylogenetic relationships within the colonial volvocales (chlorophyta) inferred from cladistic analysis based on morphological data. J Phycol 30:353–365

51. Zhang Z, Qu C, Yao R, Nie Y, Xu C, Miao J, Zhong B (2019) The Parallel Molecular Adaptations to the Antarctic Cold Environment in Two Psychrophilic Green Algae. Genome Biol Evol 11:1897–1908

52. Nozaki H, Ohtani S (1992) *Gonium sociale* (Volvocales, Chlorophyta) from Antarctica. Jpn J Phycol 40:267–271

53. Mai JC, Coleman AW (1997) The Internal Transcribed Spacer 2 Exhibits a Common Secondary Structure in Green Algae and Flowering Plants. J Mol Evol 44:258–271

54. Hedtke SM, Townsend TM, Hillis DM (2006) Resolution of Phylogenetic Conflict in Large Data Sets by Increased Taxon Sampling. Syst Biol 55:522–529

55. Pollock DD, Zwickl DJ, McGuire JA, Hillis DM (2002) Increased Taxon Sampling Is Advantageous for Phylogenetic Inference. Syst Biol 51:664–671

56. Zwickl DJ, Hillis DM (2002) Increased Taxon Sampling Greatly Reduces Phylogenetic Error. Syst Biol 51:588–598

57. Nozaki H, Itoh M, Watanabe MM, Kuroiwa T (1996) Ultrastructure of the vegetative colonies and systematic position of *Basichlamys* (Volvocales, Chlorophyta). Eur J Phycol 31:67–72

58. Nozaki H, Ito M, Uchida H, Watanabe MM, Takahashi H, Kuroiwa T (1997) Phylogenetic analysis of *Yamagishlella* and *Platydorina* (volvocaceae, chlorophyta) based on rbcl gene sequences. J Phycol 33:272–278

59. Nozaki H, Itoh M, Sano R, Uchida H, Watanabe MM, Kuroiwa T (1995) Phylogenetic relationships within the colonial volvocales (chlorophyta)inferred from rbcl gene sequence data. J Phycol 31:970–979

60. Nozaki H (1990) Ultrastructure of the extracellular matrix of *Gonium* (Volvocales, Chlorophyta). Phycologia 29:1–8

61. Nozaki H, Kuroiwa T (1992) Ultrastructure of the extracellular matrix and taxonomy of *Eudorina*, *Pleodorina* and *Yamagishiella* gen. nov.(Volvocaceae, Chlorophyta). Phycologia 31:529–541

62. Kirk DL, Birchem R, King N (1986) The extracellular matrix of *Volvox*: a comparative study and proposed system of nomenclature. J Cell Sci 80:207–231

63. Herron MD (2016) Origins of multicellular complexity: *Volvox* and the volvocine algae. Mol Ecol 25:1213–1223

64. da Silva J, Drysdale VL (2018) Isogamy in large and complex volvocine algae is consistent with the gamete competition theory of the evolution of anisogamy. Proc R Soc B Biol Sci 285:20181954

65. Umen JG (2011) Evolution of sex and mating loci: An expanded view from Volvocine algae. Curr Opin Microbiol 14:634–641

66. Coleman AW (2001) Biogeography and speciation in the *Pandorina*/*Volvulina* (chlorophyta) superclade. J Phycol 37:836–851

67. Nozaki H (2003) Origin and evolution if the genera *Pleodorina* and *Volvox* (Volvocales). Biologia (Bratisl) 58:425–431

68. Nozaki H, Ott FD, Coleman AW (2006) Morphology, molecular phylogeny and taxonomy of two new species of *Pleodorina* (Volvoceae, Chlorophyceae). J Phycol 42:1072–1080

69. Smith GM (1944) A Comparative Study of the Species of *Volvox*. Trans Am Microsc Soc 63:265–310

70. Nozaki H, Ohta N, Takano H, Watanabe MM (1999) Reexamination of phylogenetic relationships within the colonial Volvocales (Chlorophyta): an analysis of *atp*B and *rbc*L gene sequences. J Phycol 35:104–112

71. Nakada T, Misawa K, Nozaki H (2008) Molecular systematics of Volvocales (Chlorophyceae, Chlorophyta) based on exhaustive 18S rRNA phylogenetic analyses. Mol Phylogenet Evol 48:281–291

72. Pröschold T, Marin B, Schlösser UG, Melkonian M (2001) Molecular Phylogeny and Taxonomic Revision of *Chlamydomonas* (Chlorophyta). I. Emendation of *Chlamydomonas* Ehrenberg and *Chloromonas* Gobi, and Description of *Oogamochlamys* gen. nov. and *Lobochlamys* gen. nov. Protist 152:36

73. Kato S (1982) Laboratory culture and morphology of *Colacium vesiculosum* Ehrb. (Euglenophyceae). Jpn J Phycol 30:63–67

74. Ichimura T, Watanabe M (1974) The *Closterium calosporum* complex from the Ryukyu Islands - Variation and taxonomical problems. Mem Natn Sci Mus Tokyo 7:

75. Ichimura T (1973) The life cycle and its control in some species of *Closterium*, with special reference to the biological species problems. Dr. Diss.

76. Starr RC (1973) Special methods-dry soil samples, Ed. by Stein, J.R. Cambridge University Press, Cambridge

77. Sueoka N (1960) Mitotic replication of deoxyribonucleic acid in *Chlamydomonas reinhardtii*. Natl Acad Sci U S Am 46:83–91

78. Matt GY, Umen JG (2018) Cell-Type Transcriptomes of the Multicellular Green Alga *Volvox carteri* Yield Insights into the Evolutionary Origins of Germ and Somatic Differentiation Programs. G3amp58 GenesGenomesGenetics 8:531–550

79. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120

80. Xie Y, Wu G, Tang J, et al (2014) SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics 30:1660–1666

81. Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150–3152

82. Hasegawa M, Hashimoto T (1993) Ribosomal RNA trees misleading? Nature 361:23–23

83. Loomis WF, Smith DW (1990) Molecular phylogeny of *Dictyostelium discoideum* by protein sequence comparison. Proc Natl Acad Sci 87:9093–9097

84. De Clerck O, Kao S-M, Bogaert KA, et al (2018) Insights into the Evolution of Multicellularity from the Sea Lettuce Genome. Curr Biol 28:2921-2933.e5

85. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797

86. Larsson A (2014) AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics 30:3276–3278

87. Castresana J (2000) Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. Mol Biol Evol 17:540–552

88. Smith SA, Dunn CW (2008) Phyutility: a phyloinformatics tool for trees, alignments and molecular data. Bioinformatics 24:715–716

89. Le SQ, Gascuel O (2008) An Improved General Amino Acid Replacement Matrix. Mol Biol Evol 25:1307–1320

90. Jones DavidT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Bioinformatics 8:275–282

91. Dayhoff MO, Schwartz RM, Orcutt BC (1978) A Model of Evolutionary Change in Proteins. Atlas Protein Seq Struct 5:345–352

92. 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 32:268–274

93. Chernomor O, von Haeseler A, Minh BQ (2016) Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. Syst Biol 65:997–1008

94. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS (2018) UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol Biol Evol 35:518–522

95. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP (2012) MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. Syst Biol 61:539–542

96. Zhang C, Rabiee M, Sayyari E, Mirarab S (2018) ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics 19:153

97. Junier T, Zdobnov EM (2010) The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. Bioinformatics 26:1669–1670

98. Lindsey C, Rosenzweig F, Herron MD Phylotranscriptomics points to multiple independent origins of multicellularity and cellular differentiation in the volvocine algae. NCBI BioProjects, https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA701495

99. Lindsey C, Rosenzweig F, Herron MD Phylotranscriptomics points to multiple independent origins of multicellularity and cellular differentiation in the volvocine algae. Dryad, doi:10.5061/dryad.v9s4mw6w5

100. Hu Y, Xing W, Song H, Hu Z, Liu G Comparison of colonial volvocine algae based on phylotranscriptomic analysis of gene family evolution and natural selection. NCBI BioProjects, https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA532307

101. Featherston J, Arakaki Y, Hanschen ER, Ferris PJ, Michod RE, Olson BJSC, Nozaki H, Durand PM The 4-Celled *Tetrabaena socialis* Nuclear Genome Reveals the Essential Components for Genetic Control of Cell Number at the Origin of Multicellularity in the Volvocine Lineage. NCBI BioProjects, https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA393411