Placing Environmental Next-Generation Sequencing Amplicons from Microbial Eukaryotes into a Phylogenetic Context

Micah Dunthorn,*[,1] Johannes Otto,[1] Simon A. Berger,[2] Alexandros Stamatakis,[2,3] Frédéric Mahé,[1] Sarah Romac,[4,5] Colomban de Vargas,[4,5] Stéphane Audic,[4,5] BioMarKs Consortium,[†] Alexandra Stock,[1] Frank Kauff,[6] and Thorsten Stoeck[1]

[1]Department of Ecology, University of Kaiserslautern, Kaiserslautern, Germany

[2]Scientific Computing Group, Heidelberg Institute for Theoretical Studies, Heidelberg, Germany

[3]Institute of Theoretical Informatics, Department of Informatics, Karlsruhe Institute of Technology, Karlsruhe, Germany

[4]CNRS, UMR 7144, Adaptation et Diversité en Milieu Marin, Station Biologique de Roscoff, Roscoff, France

[5]UPMC Université de Paris 6, UMR 7144, Station Biologique de Roscoff, Roscoff, France

[6]Department of Molecular Phylogenetics, University of Kaiserslautern, Kaiserslautern, Germany

[†]The BioMarks Consortium includes: Bente Edvardsen (Department of Biology, University of Oslo, Oslo, Norway); Ramon Massana (Department of Marine Biology and Oceanography, Institute de Ciències del Mar, Barcelona, Catalonia, Spain); Fabrice Not and Nathalie Simon (CNRS, Université Pierre et Marie Curie [Paris 06], UMR 7144, Station Biologique de Roscoff, Roscoff, France); and Adriana Zingone (Stazione Zoologica Anton Dohrn, Villa Comunale, Naples, Italy)

***Corresponding author:** E-mail: dunthorn@rhrk.uni-kl.de.

Associate editor: Andrew Roger

Abstract

Nucleotide positions in the hypervariable V4 and V9 regions of the small subunit (SSU)-rDNA locus are normally difficult to align and are usually removed before standard phylogenetic analyses. Yet, with next-generation sequencing data, amplicons of these regions are all that are available to answer ecological and evolutionary questions that rely on phylogenetic inferences. With ciliates, we asked how inclusion of the V4 or V9 regions, regardless of alignment quality, affects tree topologies using distinct phylogenetic methods (including PairDist that is introduced here). Results show that the best approach is to place V4 amplicons into an alignment of full-length Sanger SSU-rDNA sequences and to infer the phylogenetic tree with RAxML. A sliding window algorithm as implemented in RAxML shows, though, that not all nucleotide positions in the V4 region are better than V9 at inferring the ciliate tree. With this approach and an ancestral-state reconstruction, we use V4 amplicons from European nearshore sampling sites to infer that rather than being primarily terrestrial and freshwater, colpodean ciliates may have repeatedly transitioned from terrestrial/freshwater to marine environments.

Key words: **Ciliophora, Colpodea, marine–freshwater transitions, phylogenetics, SSU-rDNA.**

Introduction

High-throughput next-generation sequencing (NGS) technologies are now being implemented in studies aimed at elucidating the patterns and processes of environmental microbial eukaryotic diversity. With its change in scale in the amount of data, initial NGS results have revealed more complex microbial eukaryotic communities composed of substantially more molecular operational taxonomic units (MOTUs) than previously determined using culture-dependent and Sanger sequencing methodologies (Amaral-Zettler et al. 2009; Stoeck et al. 2009; Nolte et al. 2010; Pawlowski et al. 2011; Logares et al. 2012; Bittner et al. 2013). This change in the scale of available data has not only allowed for finding amplicons of previously known lineages but also led to the discovery of new amplicons of yet unknown lineages that were largely overlooked because of their rarity and low abundance (e.g., Bråte et al. 2010; Lecroq et al. 2011; Orsi et al. 2011; Berney et al. 2013).

One way to analyze NGS data of amplicons from known and unknown lineages is to place them in a phylogenetic context. For example, Lecroq et al. (2011) used Illumina sequencing to find numerous, new MOTUs that phylogenetically nested within the basal grade of soft-walled, single-chambered Foraminifera. Another example is from Bråte et al. (2010), who used 454 pyrosequencing to uncover new Perkinsea MOTUs and applied the resulting phylogenetic data to evaluate marine–freshwater transitions.

A potential problem in using NGS data in phylogenetic analyses is that the data are composed of relatively short amplicons (Huber et al. 2009), and thus may not contain sufficient signal for an accurate phylogenetic placement. Earlier environmental Sanger sequencing studies generated an entire length of the small subunit ribosomal DNA (SSU-rDNA) locus or at least a large fraction of it (e.g., Richards et al. 2005; Doherty et al. 2007; Massana and Pedrós-Alió 2008; Not et al. 2009; Behnke et al. 2010; Scheckenbach et al. 2010).

Article

In current NGS studies though, only short fragments of SSU-rDNA can be generated, such as amplicons of the hypervariable V4 and/or V9 regions (e.g., Stoeck et al. 2010).

Large sections of the V4 and V9 regions are normally hard to align because of rapid rates of evolution and variation in indel lengths and are thus removed/masked before conducting standard phylogenetic analyses. But with NGS data, amplicons of these hypervariable regions are all that we have. If we are going to ask phylogenetically aware questions using short NGS amplicons, we first need to show that short NGS amplicons are indeed useful for answering these phylogenetic questions. Using ciliates as model microbial eukaryotes, we therefore investigate to which extent the inclusion of the V4 or V9 region, irrespective of their respective alignment quality, affects topological inferences using distinct alignments and phylogenetic inference methods. We show that V4 amplicons placed into an alignment of full SSU-rDNA sequences and analyzed under maximum likelihood with RAxML represents the best approach. We subsequently use this result to disentangle possible terrestrial/freshwater–marine transitions in colpodean ciliates using nearshore marine V4 amplicon data.

Results

Clade Placement

To infer relationships among the ciliates, 308 full-length ciliate SSU-rDNA Sanger sequences from all major ciliate clades, plus two outgroups, were extracted from GenBank (table 1 and supplementary table S1, Supplementary Material online). Five alignments were then assembled: "SSU," with ambiguously aligned positions conservatively removed, including in the hypervariable regions; "SSU-V4," with the entire V4 region included no matter how badly aligned; "SSU-V9," with the entire V9 region included no matter how badly aligned; "V4," with only the V4 region and all other positions removed; "V9," with only the V9 region and all other positions removed. Phylogenetic trees from these five alignments were inferred using four methods: Neighbor Joining (NJ) using a likelihood matrix of distances with PairDist (Appendix A), maximum likelihood using RAxML (Stamatakis 2006), and Bayesian inference with both MrBayes (Ronquist and Huelsenbeck 2003) and PhyloBayes (Lartillot and Philippe 2004; Lartillot et al. 2009). Here we present the RAxML tree with the best-known maximum likelihood score from the SSU alignment (fig. 1). All trees, including all species names and all bipartition support values (bootstraps and posterior probabilities), can be found in the supplement (supplementary file S1 and fig. S1, Supplementary Material online).

To assess the ability of these methods to recover the taxonomy of ciliates (table 1) and to more easily compare the trees, the "clade placement" was measured (table 2*a*). The clade placement value indicates the number of species inferred in a taxon's largest monophyletic group divided by the total number of species originally sampled in that group. To take into account that sampling was uneven, the "weighted-clade placement" was calculated (table 2*a* and fig. 2). The weighted-clade placement is the mean of the clade placements that is weighted to reflect the number of the sample species within each taxon. This measure was deployed only for the 12 major ciliate taxa (Armophorea, *Cariacothrix*, Colpodea, Heterotrichea, Karyorelictea, Litostomatea, Oligohymenophorea, Nassophorea, Phyllopharyngea, Plagiopylea, Prostomatea, and Spirotrichea); the values for the higher taxa are not independent from these lower taxon values. Bipartition support values for the taxon's largest monophyletic group were also calculated (table 2*b*).

Table 1. Number of Full-Length SSU-rDNA GenBank Accessions Sampled from the 12 Major Ciliate Clades to Infer the Ciliate Tree of Life.

Taxon	Number of Sequences
Postciliodesmatophora	
Heterotrichea	**18**
Karyorelictea	**8**
Intramacronucleata	
***Cariacothrix* + Lamellicorticata + Spirotrichea**	
***Cariacothrix* + Spirotrichea**	
Cariacothrix	**2**
Spirotrichea	**90**
Lamellicorticata	
Armophorea	**11**
Litostomatea	**47**
CONthreeP	
Colpodea	**24**
Oligohymenophorea	**64**
Nassophorea	**10**
Phyllopharyngea	**17**
Plagiopylea	**7**
Prostomatea	**10**
Outgroup	**2**
Total	***310***

NOTE.—Six larger taxa or grouping, as well as two outgroups, are also shown. Taxonomy follows Adl et al. (2012) and is unranked. Species names and GenBank numbers can be found in supplementary table S1, Supplementary Material online.

For the SSU alignment, the clade placements are similar for all four phylogenetic methods. Five taxa show values of 100% with all phylogenetic methods: Karyorelictea, *Cariacothrix*, CONthreeP, Phyllopharyngea, and Plagiopylea. Most of the remaining taxa have a clade placement that ranges between 90% and 100%. Values for Armophorea, Nassophorea, and Prostomatea are less than 90% for all four methods. The Nassophorea have the lowest clade placement at 40%; nonmonophyly of this taxon has been shown elsewhere (Gong et al. 2009). Both RAxML and MrBayes produced a weighted-clade placement of 93%, while PairDist is 85% and PhyloBayes 91%. Bipartition support values for the taxon's largest monophyletic group per taxon are largely similar across phylogenetic methods.

Caricothrix shows a clade placement of 100% for all four phylogenetic methods (although only two GenBank accessions of this small taxon are sampled here). The phylogenetic placement of *Caricothrix* varies among methods, and all are with unsupported bipartitions: sister to *Caenomorpha* (Armophorea) with PairDist; sister to *Caryotricha* and *Kiitricha* (Spirotrichea) with RAxML; sister to the clade

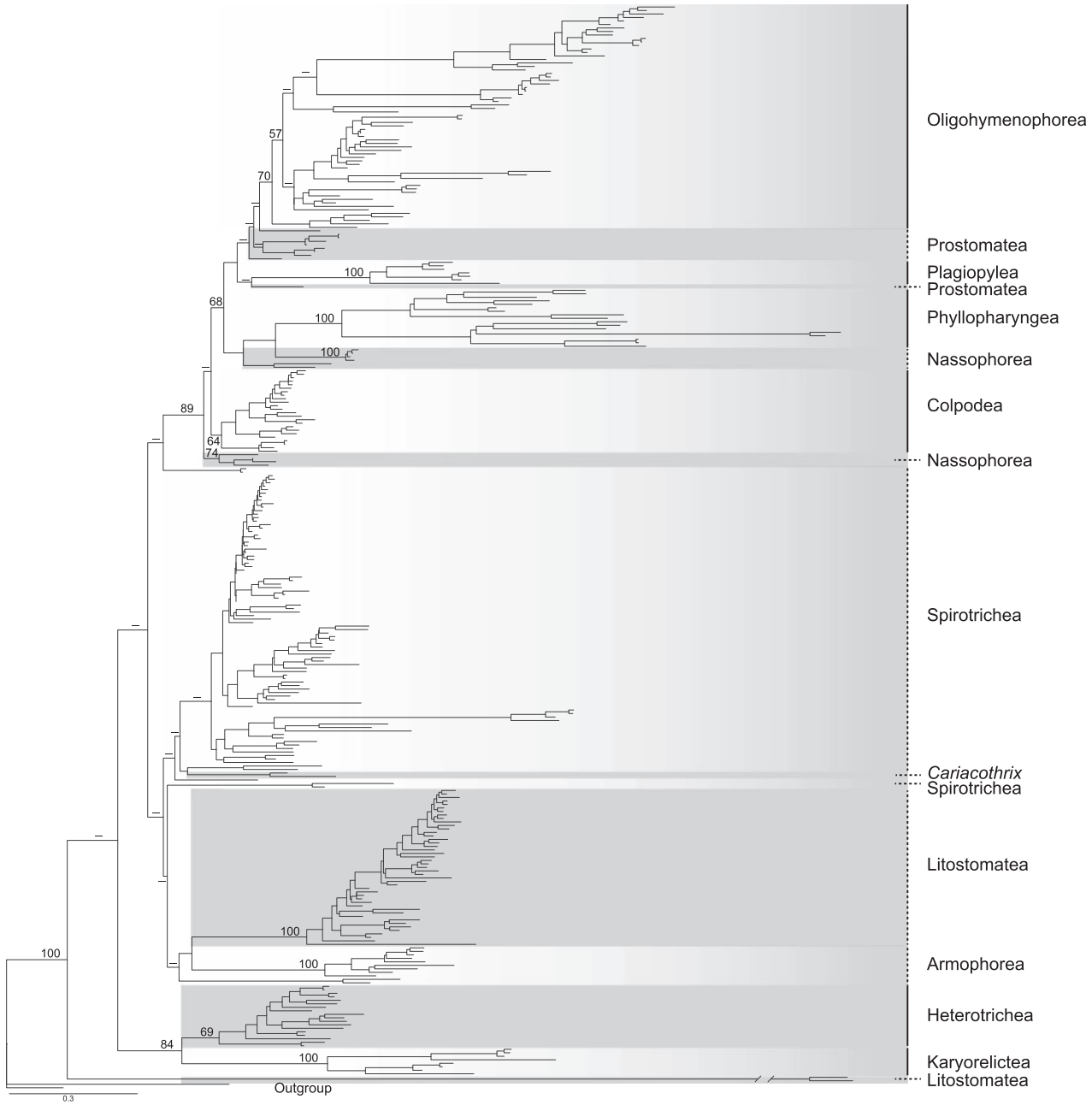

Fig. 1. Phylogeny of the ciliates inferred from the SSU alignment. The RAxML tree with the best-known maximum likelihood score is shown. The 12 major ciliate clades (table 1) are labeled. Bipartition supports from 1,000 bootstraps for the deepest nodes are labeled, which correspond to table 2*b*; values ≤50% are shown as dashes. See supplementary figure S1, Supplementary Material online for all species names. For all trees and bipartition supports, from all alignments and phylogenetic methods, see supplementary file 1, Supplementary Material online. Solid line, monophyletic group; dotted line, nonmonophyletic group.

formed by *Caryotricha*, *Kiitricha*, and *Phacodinium* (Spirotrichea) with MrBayes; and sister to the Litostomatea with PhyloBayes. Lamellicorticata has a clade placement of 96% with RAxML and MrBayes. The problem is due to *Mesodinium pulex* and *M. rubrum* (= *Myrionecta rubra*), which are located on a very long branch at the base of the ciliate tree; the long-branch problem of these two species has been noted elsewhere (Johnson et al. 2004). When *M. pulex* and *M. rubrum* are removed, we obtain a clade placement value of 100% from both RAxML and MrBayes, although bipartition support values are unsupported. The Colpodea has a 100% clade placement for PairDist, RAxML, and MrBayes; with both PairDist and MrBayes, this bipartition is supported while it is not with RAxML.

Overall, the topologies from the SSU-V4 and SSU-V9 alignments are similar for well-supported bipartitions, with most of the taxa being inferred to be fully, or at least mostly, monophyletic. For both SSU-V4 and SSU-V9, seven taxa obtain clade placements of 100% from all phylogenetic methods: Heterotrichea, Karyorelictea, *Cariacothrix*,

Table 2. Ability of Different Alignments (SSU, SSU-V4, SSU-V9, V4, and V9) and Phylogenetic Methods (PairDist, RAxML, MrBayes, and PhyloBayes) to Infer Major Ciliate Taxa.

Taxon	Alignment and phylogenetic method																			
	SSU				SSU–V4				SSU–V9				V4				V9			
	PairDist	RAxML	MrBayes	PhyloBayes	PairDist	RAxML	MrBayes	PhyloBayes	PairDist	RAxML	MrBayes	PhyloBayes	PairDist	RAxML	MrBayes	PhyloBayes	PairDist	RAxML	MrBayes	PhyloBayes
a) Clade placement																				
Postciliodesmatophora	100	100	100	61	100	100	100	100	69	100	100	69	69	69	65	69	53	26	61	15
Heterotrichea	100	100	100	88	100	100	100	100	100	100	100	100	100	100	94	100	77	38	88	22
Karyorelictea	100	100	100	100	100	100	100	100	100	100	100	100	75	100	100	100	50	50	50	50
Intramacronucleata	99	99	99	99	99	99	51	100	99	46	99	99	93	40	15	15	21	51	15	5
Cariacothrix + Spirotrichea + Lamellicorticata	61	97	97	97	98	97	97	98	68	60	60	97	59	30	30	30	38	85	30	6
Cariacothrix + Spirotrichea	95	95	95	90	91	95	97	41	97	40	97	90	96	15	18	17	43	19	4	4
Cariacothrix	100	100	100	100	100	100	100	100	100	100	100	100	0	0	0	0	100	100	100	100
Spirotrichea	97	92	92	92	93	95	92	42	100	41	92	92	86	15	18	17	44	20	4	4
Lamellicorticata	77	96	96	81	77	81	93	77	77	96	93	77	96	77	77	77	75	77	77	10
Armophorea	81	81	81	81	81	81	81	81	81	81	81	81	81	81	81	81	54	54	45	45
Litostomatea	95	95	95	95	95	95	95	95	95	95	95	95	95	95	95	95	93	95	95	12
CONthreeP	100	100	100	100	100	100	100	100	100	100	100	100	74	83	18	19	24	14	11	11
Colpodea	100	100	100	83	100	100	100	100	100	100	100	100	83	83	50	25	50	16	29	29
Oligohymenophorea	51	100	100	100	57	100	100	100	35	100	100	100	37	51	28	29	23	23	23	23
Nassophorea	40	40	40	40	60	40	40	40	60	40	40	40	40	60	100	40	40	40	40	40
Phyllopharyngea	100	100	100	100	100	100	100	100	100	100	100	100	52	52	52	52	23	46	17	47
Plagiopylea	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	85	100	42	57
Prostomatea	70	70	70	70	100	40	40	40	80	70	70	70	70	40	40	30	30	33	20	20
Weighted average	85	93	93	91	86	93	92	78	83	78	93	93	73	56	50	46	50	40	35	21
b) Bipartition support																				
Postciliodesmatophora	83	84	90	63	94	82	100	100	100	86	93	100	90	76	100	87	–	–	84	88
Heterotrichea	100	69	95	63	100	67	100	100	100	71	100	100	90	76	100	87	–	–	84	59
Karyorelictea	100	100	100	98	100	100	100	100	100	100	100	93	–	72	96	56	–	73	92	88
Intramacronucleata	62	–	100	100	83	–	50	100	94	88	54	100	–	–	100	100	–	–	96	62
Cariacothrix + Spirotrichea + Lamellicorticata	–	–	78	98	91	–	50	86	–	–	69	97	–	99	100	100	–	–	96	65
Cariacothrix + Spirotrichea	–	–	99	97	–	–	80	95	–	–	69	100	–	–	60	83	–	–	98	98
Cariacothrix	100	100	100	100	100	100	100	100	100	100	100	100	–	–	–	–	73	100	98	94
Spirotrichea	–	–	100	97	–	–	100	95	–	–	100	100	–	–	60	83	–	–	98	98
Lamellicorticata	100	–	76	78	100	–	66	100	100	–	54	100	–	99	100	100	–	–	96	77
Armophorea	100	100	100	100	100	100	100	100	100	100	100	100	99	97	100	90	–	–	91	79
Litostomatea	100	100	100	100	100	100	100	100	100	100	100	100	98	99	100	100	–	–	96	77
CONthreeP	66	89	100	100	82	88	100	100	91	88	100	100	–	–	63	56	–	–	88	62
Colpodea	86	64	97	53	98	75	100	84	99	71	99	83	52	58	77	92	–	–	99	97
Oligohymenophorea	–	70	100	100	61	85	100	100	–	73	100	100	–	–	100	98	73	92	88	62
Nassophorea	100	100	100	81	–	100	100	100	–	83	100	100	99	–	52	100	88	95	100	97
Phyllopharyngea	100	100	100	100	100	100	100	100	100	100	100	91	97	89	100	98	–	–	93	75
Plagiopylea	100	100	100	100	100	100	100	100	100	100	100	100	100	99	100	100	–	–	100	64
Prostomatea	–	–	99	63	54	99	100	100	–	–	100	77	–	65	73	98	–	–	87	75

Note.—*a*) Clade placement values, shown in %, measure the number of species inferred to be in a taxon's largest monophyletic group divided by the total number of sampled species in that taxon. The higher the value, the better the alignments and methods were at in inferring the taxonomy. Cells that are darker gray have a higher clade placement value. The weighted-clade placement, which is the weighted mean of the clade placements, takes into account uneven taxon sampling (see also fig. 2). *b*) Bipartition support values, shown in %, for the largest monophyletic group inferred in a taxon. For PairDist and RAxML, splits are considered supported if they are ≥70%; for the Bayesian posterior probabilities, splits are considered supported if they are ≥95. Well-supported clades are dark gray; unsupported clades are light gray. Bipartitions ≤50 are shown as dashes.

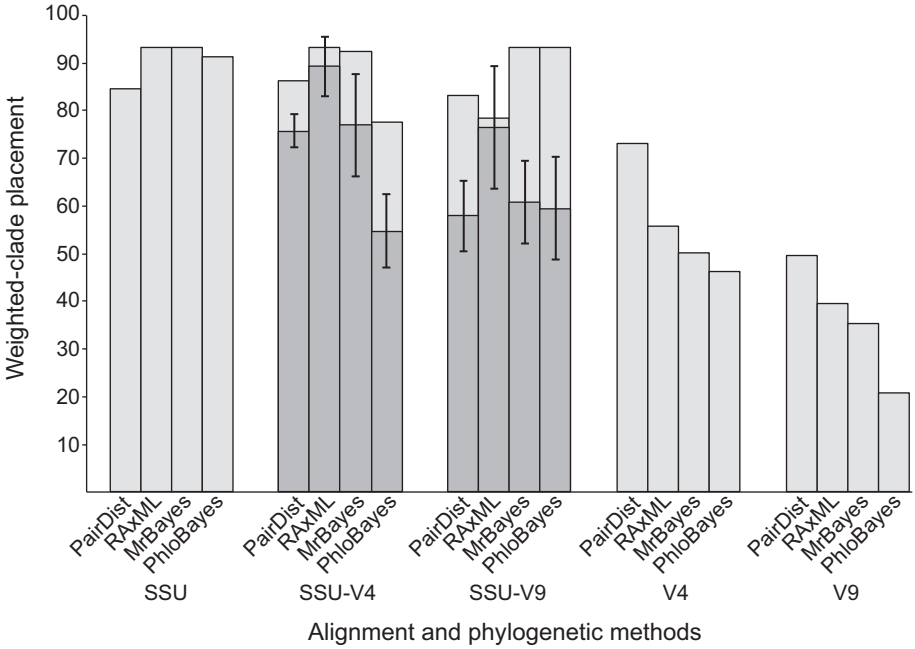

Fig. 2. Weighted-clade placement values, shown as %, which is a mean of clade placements that takes into account uneven sampling among taxa. Light gray bars are from table 2*a*, and are from alignments that either have full-length sequences (SSU, SSU-V4, SSU-V9) or sequencing of just the hypervariable regions (V4, V9). Dark gray bars are the averages of ten replicates of randomly truncating 50 full SSU-rDNA sequences to just the V4 or just the V9 region in the two alignments that included all of the V4 (SSU-V4) or all of the V9 (SSU-V9) regions regardless of alignment quality; this simulates a mixed alignment of full-length Sanger sequences and short NGS amplicons. Black lines are standard deviations.

Conthreep, Phyllopharyngea, and Plagiopylea. Like the results above from the SSU alignment, most of the other taxa have clade placements ranging between 90% and 100%. Values for Armophorea, Nassophorea, and Prostomatea were less than 90% for all four methods. With the SSU-V4 alignment, RAxML has a weighted-clade placement of 93%, MrBayes 92%, PairDist 86%, and PhyloBayes 78%. With the SSU-V9 alignment, MrBayes and PhyloBayes have a weighted-clade placement of 93%, RAxML 78%, and PairDist 83%. Bipartition supports for these taxa are also similar among the SSU, SSU-V4, and SSU-V9 alignments.

Fewer taxa are fully, or at least mostly, monophyletic in the topologies inferred from the V4 and V9 alignments. With the V4 alignment, only Plagiopylea receives a clade placement of 100% from all phylogenetic methods; Heterotrichea, Karyorelictea, and Litostomatea have clade placements ranging between 90% and 100%. With the V9 alignment, only *Cariacothrix* has a clade placement of 100% from all phylogenetic methods; for all other taxa, the clade placements range between 4% and 95%. For both the V4 and V9 alignments, the weighted-clade placements are substantially lower than the other alignments, with PairDist being the highest at 73% for the V4 alignment and 50% for the V9 alignment. This overall decrease in clade placement and weighted-clade placement values in the V4 and V9 alignments is mirrored by the overall decrease in bipartition support values for the largest monophyletic group inferred for these taxa (table 2*b*).

Randomly Truncating Sequences to V4 or V9

To measure the general effect on phylogenetic inference of including short sequences, such as environmental NGS amplicons, into alignments with full-length Sanger sequences, 50 randomly chosen sequences from the SSU-V4 and SSU-V9 alignments were truncated to just the V4, or just the V9, region. This random truncation was repeated ten times. The resulting "mixed" alignments were then analyzed with all four phylogenetic programs, and the weighted-clade placement value was measured (fig. 2). For the SSU-V4 alignments (with 50 sequences randomly truncated, ten times, to just the V4 region), the means of the weighted-clade placements are PairDist = 76%, RAxML = 89%, MrBayes = 77%, and PhyloBayes = 55%. For the SSU-V9 alignments (with 50 sequences randomly truncated, ten times, to just the V9 region), the means of the weighted-clade placements are PairDist = 58%, RAxML = 76%, MrBayes = 61%, and PhyloBayes = 59%. For both the V4 and V9, the highest mean of the weighted-clade placement is obtained with RAxML.

For each phylogenetic method, all weighted-clade placement values for these mixed alignments (with both truncated and full-length sequences) are lower than the values for the alignments containing just full-length sequences. Only with RAxML, the weighted-clade placement value for both mixed alignments is higher than the values for the V4 alignment analyzed with PairDist. There is also a smallest decrease in this value for RAxML for both mixed alignments.

Average Placement Error of V4 and V9

To further compare the hypervariable V4 and V9 regions of SSU-rDNA in a phylogenetic context, we asked how congruent the alignment sites within these regions are with the RAxML tree inferred from the SSU alignment. To do this,

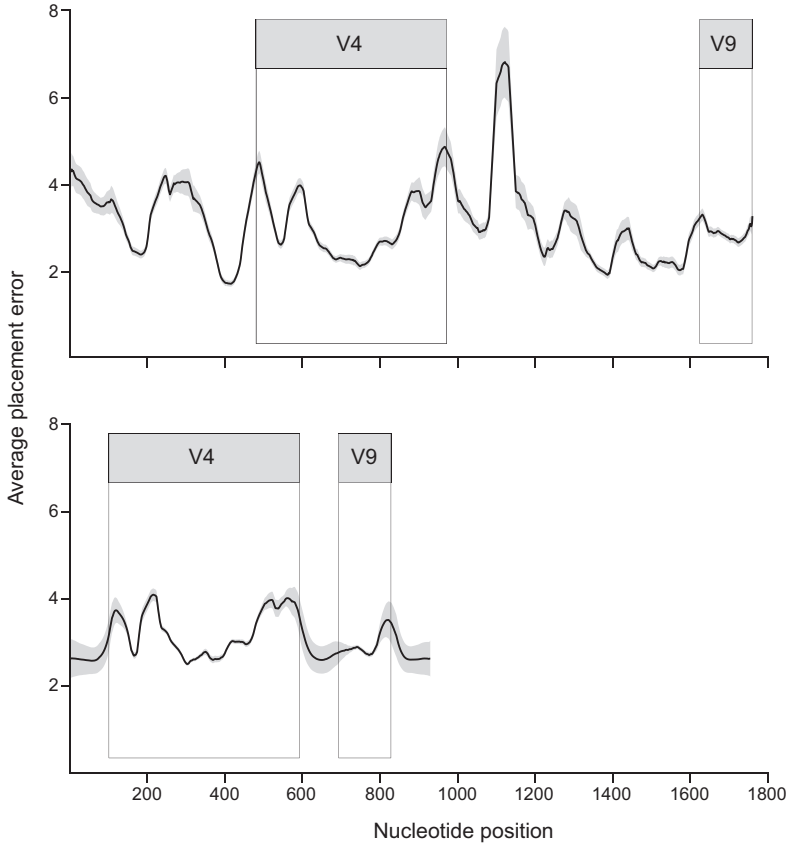

FIG. 3. Average placement error from a 50-nt sliding window as implemented in RAxML. The lower-than average placement value is, the more congruent the nucleotide site is with the input tree. (*a*) Average placement error from the SSU alignment in which all of the V4 and V9 regions were included irrespective of alignment quality. Black line, average placement error calculated with the input tree being the RAxML tree with the best-known maximum likelihood score inferred from the SSU alignment; gray shading, standard deviation of the average placement errors calculated with 200 input trees that were the bootstraps trees from the SSU alignment. (*b*) Average placement error from 200 alignments that included all of the V4 and V9 regions as well as 100 random nucleotides in-between, and on either side, of each hypervariable region. The random nucleotides varied across all 200 alignments. The input tree was the RAxML tree with the best-known maximum likelihood score inferred from the SSU alignment. Black line, mean of the average placement errors from the 200 alignments; gray shading, standard error of the average placement errors calculated from the 200 input alignments.

we used a sliding window algorithm that is implemented in RAxML. The sliding window size was set to 50 bp. The resulting "average placement error" for each nucleotide position (averaged over the 50-bp-wide sliding window) provides a measure for its congruence with the tree. The lower the average placement error, the more congruent the site is with the tree.

The first analysis with the 50-bp sliding window used the SSU alignment in which all parts of both the V4 and the V9 regions were included regardless of alignment quality (fig. 3*a*). The input tree was the RAxML tree with the best-known maximum likelihood score from the SSU alignment (i.e., fig. 1). Within the V4, the mean of the average (over all 50 sliding windows that entail a site) placement error being 3.1, the minimum is 2.1 and the maximum is 4.9. The values for V9 were less variable, with mean 2.9, minimum 2.7, and maximum at 3.3. To assess the variability of average placement errors as a function of the input tree, we also applied the sliding window approach to 200 bootstrap trees. The standard deviation from the average placement errors of the 200 bootstrap trees (from the SSU alignment) was then computed (fig. 3*a*). Overall, the standard deviations were close to the average placement errors, showing that slight variations in the input tree have only slight effects on the output of the sliding window algorithm. In other words, the sliding window algorithm is not sensitive to topologically distinct, yet reasonable (i.e., nonrandom), input trees.

Because of the 50-bp sliding window, the average placement errors for the first and last 49 nt positions within the V4 and V9 regions are affected by flanking nucleotide positions on either side of them. Furthermore, as the V9 region stretches to the 3′ end of the alignment, the last 49 nt of the V9 region has a window that uses fewer and fewer positions as it moves toward the 3′ end of the alignment. Thus, the second analysis with the 50-nt long sliding window used 200 resampled alignments to show not only variability caused by the input alignments but also to have nucleotide positions beyond the 3′ end of the V9 region (fig. 3*b*). Each alignment included the V4 and V9 regions as well as 300 randomly chosen sites from the rest of the alignment. These random sites were placed in-between the V4 and V9 regions and at the 5′ and 3′ ends of the alignment; that is, 100 random

Table 3. Number and Origin of Marine Colpodean MOTUs from Five European Nearshore Marine Sampling Sites.

MOTU Name	Number of Amplicons in MOTU	Sample Location	Sampling Depth	Nucleotide Source		
				DNA	RNA	DNA and RNA
Marine Colpodea MOTU 1	2	Blanes	Subsurface		X	
Marine Colpodea MOTU 2	1	Varna	Anoxic		X	
Marine Colpodea MOTU 3	4	Blanes, Naples	Subsurface			X
Marine Colpodea MOTU 4	3	Blanes	Subsurface		X	
Marine Colpodea MOTU 5	1	Blanes	Subsurface	X		
Marine Colpodea MOTU 6	4	Naples	DCM		X	
Marine Colpodea MOTU 7	1	Blanes	Subsurface		X	
Marine Colpodea MOTU 8	2	Blanes, Varna	Subsurface/anoxic		X	
Marine Colpodea MOTU 9	4	Varna	Anoxic	X		
Marine Colpodea MOTU 10	1	Blanes	Subsurface	X		
Marine Colpodea MOTU 11	2	Blanes	Subsurface		X	
Marine Colpodea MOTU 12	3	Naples	Sediment		X	
Marine Colpodea MOTU 13	24	Naples, Oslo	Sediment			X
Marine Colpodea MOTU 14	4	Naples	Sediment		X	
Marine Colpodea MOTU 15	1	Naples	Subsurface		X	
Marine Colpodea MOTU 16	1	Naples	Sediment		X	
Marine Colpodea MOTU 17	3	Naples	DCM		X	
Marine Colpodea MOTU 18	447	Blanes, Gijon, Naples, Roscoff	Subsurface/DCM			X
Marine Colpodea MOTU 19	12	Naples	Subsurface/DCM		X	
Marine Colpodea MOTU 20	1	Gijon	Subsurface		X	
Marine Colpodea MOTU 21	1	Blanes	Subsurface	X		
Marine Colpodea MOTU 22	1	Blanes	Subsurface		X	

NOTE.—Amplicons derived from 454 pyrosequencing of hypervariable V4 region of SSU-rDNA were grouped at 98% similarity. Sampling of DNA and RNA occurred at four depths (Varna is the only site with an anoxic layer). DCM, deep chlorophyll maximum.

sites-V4 region-100 random sites-V9 region-100 random sites. These alignments were then used to run the sliding window algorithm using, as above, the RAxML tree with the best-known maximum likelihood score from the SSU alignment.

From these 200 alignments, the individual average placement errors were used to calculate mean and standard deviations. Within the V4, the mean of the average placement errors is 3.3, while the minimum is 2.5 and the maximum is 4.1. As above, the values for V9 were less variable, with a mean average placement error being 3.0, minimum is 2.7 and maximum is 3.5. Within the flanking regions that varied across all 200 alignments, the standard deviations are higher (fig. 3*b*) than in the analysis above (fig. 3*a*). Within the V4 and v9, there was also considerable variability in the first and last 49 positions. Some variation exists in the center of the V4 and V9 regions (although they are the same across all 200 alignments), which derives from the slight errors induced when RAxML is estimating likelihoods. Overall, by randomizing the flanking regions and making sure there is a flanking region on either side of the alignment region of interest, the mean of the average placement errors increases only slightly for the V4 and V9 regions.

Marine Environmental V4 Amplicons

A total of 87,724 marine ciliate V4 amplicons were obtained from DNA and RNA samples collected from nearshore marine sites off the cost of Blanes, Spain; Gijon, Spain; Naples, Italy; Oslo, Norway; Roscoff, France; and Varna, Bulgaria. Of these, 523 blasted to the Colpodea—one of 12 major ciliate clades sensu Adl et al. (2012) and Foissner et al. (2011). These amplicons came from both plankton and sediment. Grouping these 523 colpodean amplicons at 98% similarity resulted in 22 MOTUs (table 3). One of these MOTUs, the Marine Colpodea MOTU 18, contains 447 of the amplicons (or 85%). On the other hand, 19 MOTUs contain less than four amplicons; these 19 MOTUs can be considered as being a part of Sogin et al.'s (2006) "rare biosphere." Eighteen of the MOTUs are found at just one out of five sampling sites. Marine Colpodea MOTU 18—the MOTU composed of 85% of the amplicons—is found at four of the five sampling sites in the subsurface and deep chlorophyll maximum: Barcelona, Gijon, Naples, and Roscoff.

Four MOTUs, the Marine Colpodea MOTUs 5, 9, 10, and 21, are composed of amplicons derived exclusively from DNA; the organisms from which these amplicons derive could have been dead or in a cyst stage, rather than being metabolically active (Stoeck et al. 2007; Not et al. 2009). The remaining MOTUs are composed of amplicons derived exclusively from RNA samples or from both RNA and DNA samples. Marine Colpodea MOTU 9 is found in DNA samples from the anoxic layer in Varna; given that it was found with just DNA, there is no evidence that it was metabolically active at the time of collection. Marine Colpodea MOTU 8 is also found in the anoxic layer in Varna with RNA; however, given that it was also found in the oxic layer in Barcelona,

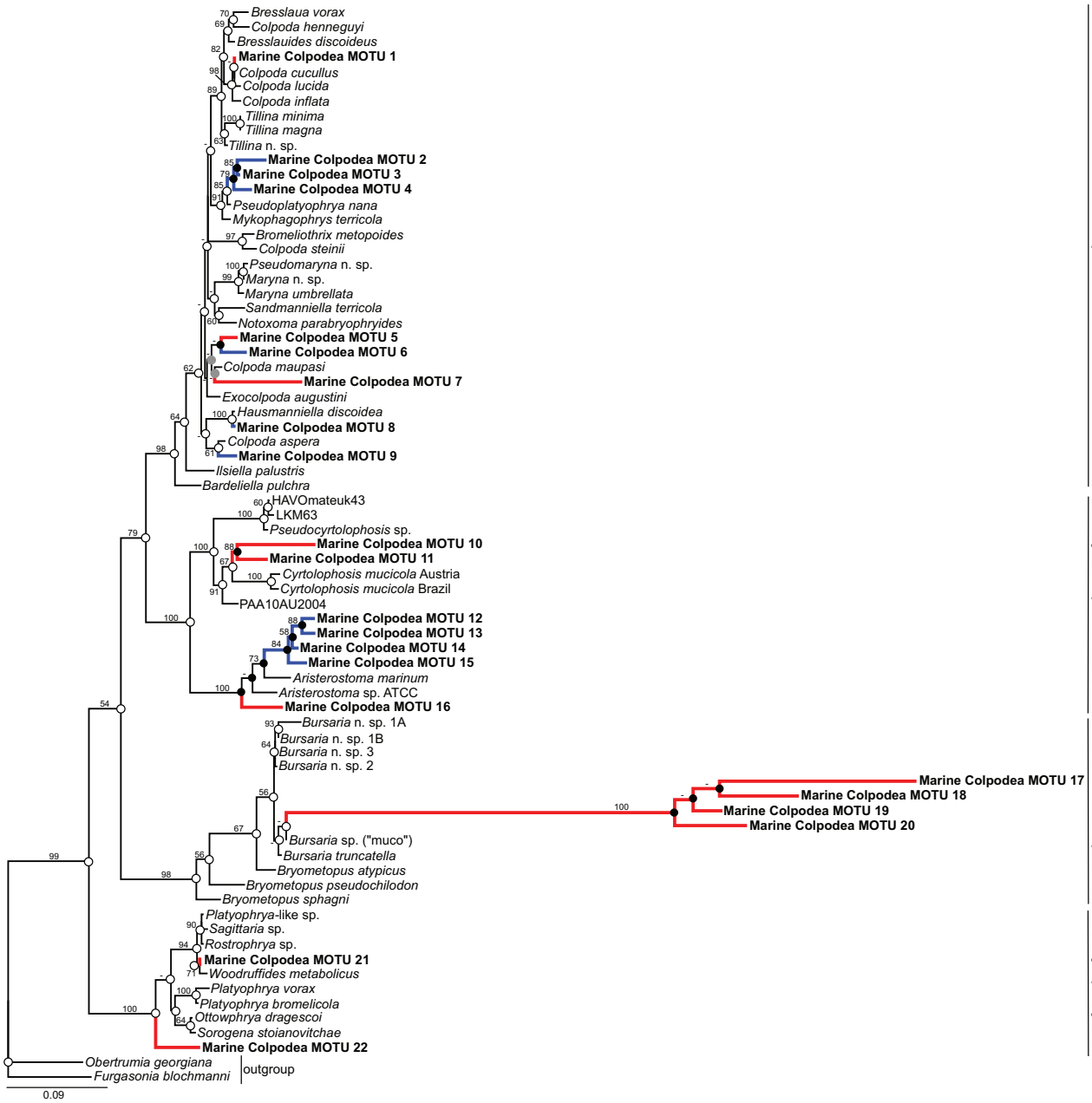

FIG. 4. Phylogenetic placement of 22 MOTU representatives (in bold), from European marine V4 amplicons grouped at 98% similarity, into the Colpodea. The RAxML tree with the best-known maximum likelihood score is shown, with bipartition support from 1,000 bootstraps; values ≤50% are shown as dashes. *Aristerostoma* (in the Cyrtolophosidida) is the only clade previously known to be marine; all other colpodean are previously known only from freshwater and terrestrial environments. Additional support for phylogenetic placements of the MOTUs was estimated using the Evolutionary Placement Algorithm (EPA): blue lines, high confidence, with likelihood weight ratio ≥95%; red lines, low confidence with likelihood weight ratio <95%, where alternative placements are near-equally valid. Ancestor states were parsimoniously inferred: white circle, freshwater/terrestrial; black circle, marine; and gray circle, freshwater/terrestrial or marine.

and that oxic to anoxic transitions in ciliates are rare (Forster et al. 2012), the RNA found in the Varna sample could have come from a cell that recently descended from the above oxic layer, or the pyrosequencing was able to recover what little RNA was inside their cysts. If future morphological studies do show that the organisms from which Marine Colpodea MOTUs 8 and 9 derive are metabolically active in these deep marine zones, it will be the first time that a colpodean is shown to be capable of anaerobic metabolism.

Phylogenetic Placement of Marine V4 MOTUs

One representative from each of the aforementioned 22 MOTUs was placed into the alignment of 51 full-length SSU-rDNA Sanger sequences from Foissner et al. (2011) and the tree inferred with RAxML (fig. 4). Marine Colpodea MOTUs 12, 13, 14, 15, and 16 form a clade with *Aristerostoma*, which is the one previously known marine colpodean, in the subclade Cyrtolophosidida. The remaining 18 MOTUs fall throughout the tree within all four major colpodean subclades sensu Foissner et al. (2011): Bursariomorphida,

Colpodida, Cyrtolophosidida, and Platyophryida. Marine Colpodea MOTUs 2, 3, and 4 nest within the clade formed by *Mykophagophrys* and *Pseudoplatyophrya* (Grossglockneriida). The Grossglockneriida have modified oral structures that allow them to feed on fungal hyphae and yeast cells (Foissner 1993, 1999); the organisms from which the three MOTUs derive may likewise be mycophagous.

The clade formed by Marine Colpodea MOTUs 17, 18, 19, and 20 is situated at a long branch relative to the rest of the tree. Therefore, the Evolutionary Placement Algorithm (EPA; Berger et al. 2011) was used to infer alternative phylogenetic placements of the MOTUs. There is high support for the original RAxML placement of ten of the MOTUs (shown in blue in fig. 4). There is little support for the placement of the remaining MOTUs (shown in red in fig. 4). Most of the alternative placements are within the same four major colpodean subclades (i.e., within the Bursariomorphida, Colpodida, Cyrtolophosidida, or Platyophryida; data not shown). For marine Colpodea MOTUs 17, 18, 19, and 20—which are situated on the long branch—EPA suggests alternative placements not only throughout the colpodean tree but also within the outgroups (data not shown); molecular evidence that these four marine MOTUs belong to the Colpodea is therefore lacking.

Reconstructing Ancestral States of Colpodean Ciliates

Mesquite (Maddison WP and Maddison DR 2011) was used to parsimoniously reconstruct the ancestral state of the Colpodea using the RAxML tree above (from the alignment containing marine colpodean MOTU representatives of V4 amplicons and full-length Sanger sequences) and ignoring the potential alternative EPA placements. The 22 marine colpodean MOTUs and the two *Aristerostoma* were coded as marine; the remaining species were coded as freshwater/terrestrial. The ancestral states for the deepest nodes within the tree are inferred to be freshwater/terrestrial (fig. 4), with the coded state of the outgroup not affecting this result (data not shown). Given this ancestral reconstruction, there have been at least ten transitions to the marine environment within the colpodean ciliates. No transitions from marine to freshwater/terrestrial were reconstructed.

Discussion

Biogeographical research in microbial eukaryotes has traditionally focused on evaluating levels of globally dispersed species versus those that are locally endemic (Finlay 2002; Katz et al. 2005; Richards et al. 2005; Fenchel and Finlay 2006; Foissner et al. 2008). This research has greatly expanded over the past few years to address additional questions such as how chemical gradients structure communities (Behnke et al. 2010; Orsi, Song, et al. 2012), how dynamic are communities over time (Doherty et al. 2007; Lara et al. 2011), the extent of the "rare biosphere" (Stoeck et al. 2010; Bittner et al. 2013; Egge et al. 2013), and whether there are freshwater–marine transitions (Logares et al. 2009; Forster et al. 2012). Research on frequency and polarity of freshwater–marine transitions, in particular, has at its core phylogenetically aware questions.

This study was designed to look at the efficacy of using short amplicons, derived from environmental NGS studies, in asking questions that rely on phylogenies such as the frequency and occurrence of freshwater–marine transitions. To provide a comparative background, we first inferred a broadly sampled ciliate tree using full-length Sanger SSU-rDNA sequences. We then asked how our ability to infer the ciliate tree is affected by including all of the hypervariable V4 or V9 regions or just using either the V4 or V9 regions. We put the result into an ecologically usable context by inferring the number and direction of salinity transitions in one group of ciliates.

A Broadly Sampled Ciliate Tree

Molecular phylogenetic inferences in ciliates have primarily focused on sequencing the SSU-rDNA of only one or two major clades (Strüder-Kypke et al. 2006; Agatha and Strüder-Kypke 2007; Schmidt et al. 2007; Utz and Eizirik 2007; Dunthorn et al. 2008; Gong et al. 2009; Yi et al. 2010; Zhan et al. 2013). For those few analyses that have sampled all major ciliate clades, taxon sampling was low and/or key taxa were missing (e.g., Riley and Katz 2001; Lynn 2003; Gong et al. 2009; Phadke and Zufall 2009; Vďačný et al. 2010). We therefore lack a ciliate tree inferred from a broad sampling of all taxa that can reveal which morphological hypotheses are supported and which ones require further scrutiny using improved taxon and/or character sampling.

We fill in this gap here by inferring the ciliate tree based on a broad taxon sampling that includes representative sequences from all major ciliate clades from Adl et al. (2012). This alignment of 308 ciliate sequences and two outgroups—the SSU alignment—was masked to remove ambiguously aligned nucleotide positions (e.g., those in the hypervariable V4 and V9 regions) and analyzed with four distinct phylogenetic methods (PairDist, RAxML, MrBayes, and PhyloBayes). Overall, the inferred trees uncover and confirm many clades found in previous studies; where there are disagreements between this and previous studies, the bipartition support values are low (table 2, fig. 1, supplementary fig. S1 and file S1, Supplementary Material online). The clade placements for each taxon are similar across phylogenetic methods, and many of the taxa analyzed here are inferred to be monophyletic (table 2).

Only three relationships inferred in the trees from the SSU alignment will be discussed in detail here: *Cariacothrix*, Lamellicorticata, and Colpodea. A close phylogenetic relationship between *Cariacothrix* and Spirotrichea was recognized when Stoeck et al. (2003) first uncovered this taxon in a molecular environmental diversity survey of the anoxic Cariaco Basin of the coast of Venezuela. In a later phylogenetic analysis, Orsi, Edgcomb, et al. (2012) inferred that the *Cariacothrix* are sister to the Spirotrichea with high node support; however, in this analysis, some key spirotrichean taxa were not included (e.g., *Caryotricha* and *Kiitricha*). Here, *Cariacothrix* either nests within the Spirotrichea (with

RAxML and MrBayes), forms a clade with some Armophorea (with PairDist), or are sister to the Litostomatea (with PhyloBayes); none of these relationships, though, are well supported. Whether or not the *Cariacothrix* are one of the 12 major clades (sensu Adl et al. 2012), or one of the 12 major classes (sensu Orsi, Edgcomb, et al. 2012), of the ciliates is therefore currently not answered by the SSU-rDNA data.

Previously published articles are ambiguous about Lamellicorticata, a taxon that unites the largely anaerobic Armophorea with the free-living or anaerobic/symbiotic Litostomatea. Vďačný et al. (2010) inferred Lamellicorticata to be monophyletic when they excluded *Caenomorpha* (Armophorea) and many Spirotrichea, while Zhang et al. (2012) did not recover it, and Miao et al. (2009) found it to be monophyletic but with no support. With the inclusion of all major groups in the Armophorea and the exclusion of the *M. pulex* and *M. rubrum* (which are situated on long branches), Lamellicorticata are here inferred to be monophyletic with RAxML and MrBayes, but with low bipartition support.

Although the Colpodea are recognized as a taxon based on the presence of the LKm (left kinetodesmal) fiber in the somatic ciliature (Lynn 1976; Small and Lynn 1981; Foissner 1993), molecular support for monophyly from nuclear SSU-rDNA data was lacking in an earlier study when all potentially closely related outgroups were included (Dunthorn et al. 2008). Later nuclear and mictochondrial SSU-rDNA analyses did not sample sufficient outgroups to allow for a meaningful test of monophyly (Lynn et al. 1999; Lasek-Nesselquist and Katz 2001; Dunthorn et al. 2008, 2009; Foissner and Stoeck 2009; Bourland et al. 2011; Dunthorn et al. 2011; Foissner et al. 2011; Quintela-Alonso et al. 2011; Bourland et al. 2012; Dunthorn, Katz, et al. 2012; Bourland et al. 2013; Foissner et al. 2013). With the increased taxon sampling here, the Colpodea are inferred to be monophyletic with PairDist, RAxML, and MrBayes; we obtain high bipartition support for this relationship only from PairDist and MrBayes.

Efficacy of Short NGS Amplicons in Phylogenetic Inference

Little work has been done in justifying which hypervariable region should be amplified and sequenced in environmental surveys using NGS technologies. Pawlowski and Lecroq (2010), for example, found that Foraminifera amplicons of the Helix 37 can be used to distinguish among identified species. Also, Dunthorn, Klier, et al. (2012) and Pernice et al. (2013) found that genetic distances from the V4 region more closely resemble those obtained from the full-length SSU-rDNA than genetic distances from V9. A further criterion to choose among hypervariable regions is their phylogenetic signals.

One problem with hypervariable regions is that their fast-evolving nucleotide sites make it difficult to align them unambiguously. Removal of ambiguously aligned regions in multiple sequence alignments is a standard practice in phylogenetic analyses (Swofford et al. 1996; Castresana 2000; Löytynoja and Milinkovitch 2001), and there are programs that can automatically remove these positions (Talavera and Castresana 2007; Penn, Privman, Ashkenazy, et al. 2010; Penn, Privman, Landan, et al. 2010). In ciliates in specific, efforts have been made to remove ambiguous positions, such as in the hypervariable V4 and V9 regions (e.g., Dunthorn et al. 2011; Zhan et al. 2013).

The exclusion of parts of the V4 and V9 regions that produced ambiguously aligned positions (SSU alignment) versus inclusion of all of these regions (SSU-V4 and SSU-V9 alignments) had little effect on clade and weighted-clade placements among the different phylogenetic methods (table 2*a* and fig. 2). Also, there was also little difference in well-supported bipartitions from the trees inferred using these alignments (table 2*a*, fig. 1, and supplementary fig. S1, Supplementary Material online). Regardless of the small effect on phylogenetic outcomes that resulted from including all of the hypervariable regions in alignments, however, the problem of assessing positional homology within the V4 and V9 regions remains.

In contrast to the relative lack of impact on the phylogenies inferred from including all of the V4 and V9 regions, regardless of alignment quality, into SSU-rDNA alignments, there is a dramatic difference among phylogenetic relationships in the trees inferred from alignments consisting of just the V4 and V9 regions (table 2*a* and fig. 2). This simulates the probable result of using only NGS amplicons for phylogenetic inferences. The clade placements, weighted-clade placements, and bipartition support values from the V4 alignment are better than those from V9 alignment, but both are substantially lower than the values from the SSU, SSU-V4, and SSU-V9 alignments. With the methods used here, PairDist is the best when using alignments of just the hypervariable regions: for example, when estimating phylogenetic distances among archaeal, bacterial, or microbial eukaryotic communities using UniFrac (Lozupone et al. 2006).

The short amplicons produced by NGS technologies mean that sequencing the entire SSU-rDNA locus in environmental studies intended to address ecological or evolutionary questions is impractical at present. Furthermore, as discussed above, using amplicons of just the hypervariable V4 or V9 regions to answer phylogenetically aware questions in ciliates represents a suboptimal strategy. Weighted-clade placements (fig. 2) support a strategy of including short V4 or V9 amplicons in "mixed" alignments (short V4 or V9 amplicons combined with full-length Sanger SSU-rDNA sequences), as the results are more similar to those from alignments of full-length sequences alone than what would be obtained by relying on using sequences of just the V4 or V9 regions. Finally, RAxML was the most accurate method among the four tested for inferring phylogenies from mixed alignments of short amplicons and full-length sequences.

Including V4 amplicons in alignments of full-length Sanger sequences may thus represent the most appropriate approach for using NGS data to answer phylogenetic questions about ciliates, but this does not mean that all nucleotide positions in the V4 region yield better inferences than those in the V9 region. In analyses using a 50-nt sliding window in RAxML (fig. 3), the mean of average placement errors is lower

for the V9 region, but the V4 region still has individual positions whose average placement errors are lower than any position in V9.

Freshwater/Terrestrial-to-Marine Transitions in Colpodean Ciliates

Transitions between freshwater and marine environments in microbial eukaryotes are thought to be infrequent given the physicochemical barrier of salinity gradients and the ecological barrier of colonization in the face of already established competitors (Logares et al. 2009). A few studies have nevertheless found evidence for at least some, presumably ancient, transitions between these environments in the Cryptophyceae (von der Heyden et al. 2004), Foraminifera (Holzmann et al. 2003), Haptophyta (Simon et al. 2013), and Perkinsea (Bråte et al. 2010). There is evidence for more recent transitions in ciliates (Finlay et al. 2006; Bachy et al. 2012; Forster et al. 2012).

Given what was shown above about the efficacy of short NGS sequences in ciliate phylogenetic inferences, we used V4 amplicon data derived from nearshore European marine environments to ask whether there were similar freshwater/terrestrial to marine transitions in the ciliate clade Colpodea. There are major disagreements about this group of approximately 200 known species: is this group monophyletic?, what relationships exist among subclades?, and are they secretively sexual? (Foissner 1993; Dunthorn et al. 2008; Dunthorn and Katz 2010; Foissner et al. 2011). What is not a source of contention is that the colpodeans are primarily freshwater and terrestrial (Foissner 1993; Lynn 2008). They are easily found in soils, mosses, ponds, and plant-held waters (Foissner 1987, 1993; Foissner et al. 2002; Kreutz and Foissner 2006; Lara et al. 2007; Dunthorn, Stoeck, et al. 2012). The one well-documented marine clade is that of the two to three species in *Aristerostoma* (Foissner 1993; Dunthorn et al. 2009). There are reports of *Rhyposophrya aplanata* also being marine (Kahl 1933; Kiesselbach 1936), but reliable morphological data to identify *R. aplanata* as a colpodean are lacking (Foissner 1993).

From six offshore marine sampling sites around Europe, 523 colpodean amplicons were recovered and grouped into 22 MOTUs (table 3). One of these MOTUs (MOTU 18) contains 85% of the sequences and was found in all but one of the sampling sites. Many of the other MOTUs are rare. In a phylogenetic inference of the short amplicons using RAxML, in combination with full-length Sanger sequences, some MOTUs formed a clade with the already-known marine *Aristerostoma* (fig. 4). The other MOTUs fell throughout the tree, pointing toward a potentially large diversity of currently unknown marine colpodean ciliates. There was variable support for the exact placement of these MOTUs with EPA (fig. 4); in particular, EPA placed MOTUs 17, 18, 19, and 20 (which are on a long branch) not only throughout the Colpodea but also in the outgroup. Future studies targeting the full-length sequence from the organisms from which these four MOTUs derive are needed to more accurately place them either within the ciliates or in other eukaryotic taxa.

Bachy et al. (2012) used a parsimony-based ancestral state reconstruction method to show that, in the globally distributed and ecologically important loricate Choreotrichia ciliates, there have been at least three transitions from an ancestral marine habitat to freshwater. Here, we also used a parsimony reconstruction to infer the direction of these transitions in the colpodean ciliates (fig. 4). The reconstructed ancestral state is freshwater/terrestrial, with multiple, and recent, transitions to the marine environment. While also performing a likelihood reconstruction is tempting, providing transition rates for these transitions would be extremely hard to justify.

Rather than primarily being a freshwater and terrestrial clade as was traditionally thought (Foissner 1993; Lynn 2008), these data from the environmental amplicons point to the colpodeans as being equally marine. What these marine species look like, though, awaits future morphological research. Given this radical marine perspective on the Colpodea, we offer four reasons why this diversity has previously been undetected. First, given the rarity of most of these MOTUs, they would have easily been missed in previous morphological and Sanger sequencing studies that would only have picked up only the most common ciliates. Second, they may be visually unremarkable and easily missed in morphological studies that focused on more charismatic ciliates such as in the Karyorelictea, Oligohymenophorea, and Spirotrichea. Third, those researchers studying the colpodeans primarily focus on terrestrial environments (e.g., Foissner 1993; Dunthorn et al. 2008; Quintela-Alonso et al. 2011; Bourland et al. 2012). Fourth, the colpodeans amplified and pyrosequenced from the marine samples may not actually be normally metabolically active in the sampling sites and were merely from recent continental runoff waters. This last option can easily explain those few MOTUs that are exclusively represented by just DNA. Those MOTUs represented by RNA could be explained by such deep pyrosequencing that the little RNA that was present in dividing and resting cysts was picked up by the sampling and pyrosequencing methodologies. A detailed morphological study targeting colpodeans in marine environments is needed to evaluate these alternatives.

Materials and Methods

Taxonomy and Taxon Sampling

The latest ciliate classification from Adl et al. (2012) was largely followed. What differs is that the circumscription of the Nasspohorea follows Lynn (2008); that is, the Synhymenia are included in the Nassophorea and not in the Phyllopharyngea. We also use Vďačný et al.'s (2010) Lamellicorticata. Following Garcia-Cuetos et al. (2012), we use *M. rubrum* instead of *M. rubra*.

For the broad ciliate tree, GenBank SSU-rDNA sequences were downloaded from 308 ciliate morphospecies representing all 12 major clades (table 1 and supplementary file S1, Supplementary Material online), which are labeled as classes

in Lynn (2008): Armophorea, *Cariacothrix*, Colpodea, Heterotrichea, Karyorelictea, Litostomatea, Nassophorea, Oligohymenophorea, Phyllopharyngea, Plagiopylea, Prostomatea, and Spirotrichea. A dinoflagellate and apicomplexan were used as outgroups. For the Colpodea tree, the SSU-rDNA sequences and alignment from Foissner et al. (2011) was used. The hypervariable V4 and V9 regions of the SSU-rDNA locus, as defined here, are those that are amplified by the primers from Stoeck et al. (2010).

Alignments

The 308 ciliate plus two outgroup sequences were aligned using Hmmer v2.3.2 (Eddy 1998) with default settings. The training alignment for model building comprised available ciliate SSU-rDNA sequences downloaded from the European Ribosomal Database (Wuyts et al. 2004) and aligned according to their secondary structure. The alignment was manually curated, and ambiguously aligned positions were conservatively removed with MacClade v4.08 (Maddison DR and Maddison WP 2003). The final alignment includes 1,494 nt positions, of which 916 are parsimony-informative.

For comparison, ambiguously aligned positions were also removed using Gblocks v0.91 b (Castresana 2000; Talavera and Castresana 2007), mostly using default parameters, with the exception that smaller final blocks, gap positions within the final blocks, and less strict flanking positions were allowed. The resulting Gblocks alignment contains 1,491 characters, of which 935 are parsimony-informative (data not shown). Because of the close similarity in length between the two masking methods, we only used the manually masked alignment in our analyses.

Four final alignments were then constructed for further phylogenetic analyses: "SSU," "SSU-V4," "SSU-V9," "V4," and "V9." The SSU alignment is the 1,494 nt alignment computed with Hmmer, with ambiguously aligned positions conservatively removed in MacClade, including the hypervariable regions. The SSU-V4 alignment is the complete SSU alignment with the entire V4 region included no matter how badly aligned. The SSU-V9 alignment is the complete SSU alignment with the entire V9 region included no matter how badly aligned. V4 is only, and all, of the V4 region with all other positions removed. V9 includes only, and all, the V9 region with all other positions removed.

To test whether inclusion of short amplicons affects topological inferences, 50 full-length sequences from the SSU-V4 and SSU-V9 alignments were randomly truncated to just the V4 or just the V9 region, respectively. This random truncation was performed ten times. Resulting alignments were then analyzed with the same phylogenetic methods as the full-length sequence alignments.

Phylogenetic Analyses

The GTR-I-Γ evolutionary model was the best-fit model selected by the Akaike information criterion in MrModeltest v2 (Nylander 2004). All alignments were analyzed with four methods. PairDist (Appendix A), using the GTR-I-Γ model, with node support from 1,000 multiparametric bootstrap replicates. RaxML v7.2.5 (Stamatakis 2006), using the GTR-I-Γ model, with support values obtained from a 50% majority rule consensus tree of 1,000 nonparametric bootstrap replicates. MrBayes v3.2.1 (Ronquist and Huelsenbeck 2003) was run, using the GTR-I-Γ model, with four chains running 10 million generations sampling every 1,000 generations. And, to account to the possibility of model and rate variation, PhyloBayes v3.2 e (Lartillot and Philippe 2004; Lartillot et al. 2009), using the "Q-matrix mixture" (QMM) model, running at least 1.5 million generations and sampling every cycle. For both Bayesian methods, the first 25% of sampled trees were considered as burn-in trees and were discarded before constructing the majority-rule consensus trees on the remaining 75% of the sampled trees. Trees were visualized with FigTree v1.3.1 (Rambaut 2006). For PairDist and RAxML, splits were considered to be supported if they were supported by $\geq$70% of the trees (Hillis and Bull 1993); for the Bayesian posterior probabilities, splits were considered to be supported if they were $\geq$95 (Alfaro et al. 2003).

To evaluate the ability of the phylogenetic programs to infer the ciliate classification used here, we calculated the "clade placement" for each alignment version (SSU, SSU-V4, etc.). The clade placement, presented in % values, measures the number of species inferred to be a taxon's largest monophyletic group in relation to the total number of expected (according to the assumed classification) species/sequences in that taxon:

$$\text{Clade placement} = \frac{\text{Number of species in taxon's largest monophyletic group}}{\text{Number of species sampled in taxon}}.$$

The total number of expected species in that taxon is the number of taxa sampled from GenBank (table 1). If, for example, all Colpodea formed a monophyletic group, the corresponding clade placement would be 24/24 = 100%. On the other hand, if the two sampled *Cariacothrix* did not form a clade, its clade placement would be 0/2 = 0%.

To further quantify the ability of phylogeny programs to recover the ciliate classification for each alignment version, the weighted-clade placement was calculated. The weighted-clade placement, also presented in % values, was calculated by taking into account that some clades consist of many sequences, while others have few:

$$\text{Weighted-clade placement} = \frac{\sum_{1}^{12} \frac{\text{Number of species in taxon} \times \text{taxon's percent clade placement}}{308}}{12}.$$

The total number of species here is 308; that is, all sampled ciliates (table 1). Average clade placement was then calculated from just the 12 major, and independent, ciliate clades (labeled as classes in Lynn [2008]).

To evaluate the ability of individual nucleotide positions in the hypervariable V4 and V9 regions to infer the ciliate tree prior to the phylogenetic analyses, we used the sliding window algorithm (with a window size of 50 nt) as implemented in RAxML. This algorithm allows for calculating the

congruence of individual sites of a given alignment with a given tree topology inferred on that alignment. The algorithm can be invoked using the following command:

raxmlHPC-SSE3 -f S -s <alignment.file> -W 50 -t <tree.file> -m GTRIGAMMA -n sliding.

The resulting "average placement error" that is calculated for each nucleotide position (averaged over the 50-nt sliding window) provides an estimate for the congruence of the phylogenetic signal of each site with the given tree (in this case, the best-known ML tree for the SSU alignment inferred with RAxML). The lower the average placement error, the more congruent the signal at a nucleotide site will be with the tree. Higher values indicate incongruence. The input tree was always the best-known RAxML tree from the SSU alignment. The first alignment used with the sliding window algorithm was the SSU alignment in which all parts of both the V4 and the V9 regions were included no matter how badly aligned. To assess the variability of the sliding window algorithm results as a function of the input tree, we also executed sliding window analyses on 200 RAxML bootstrap trees inferred on the SSU alignment. To assess variability as a function of the input alignment, we generated 200 distinct alignments that always included the unaltered V4 and V9 regions, as well as 300 randomly chosen sites from the SSU alignment excluding the V4 and V9 regions. These 300 sites were inserted between V4 and V9 regions and on either side of the V4 and V9 regions; that is, 100 random sites/V4 region/100 random sites/V9 region/100 random sites. The placement errors inferred via the sliding window algorithm of these 200 alignments were then compared with the placement errors inferred on the original full-length SSU alignment.

Pyrosequencing and Analyses of Environmental OTUs

Coastal marine ciliate V4 454 pyrosequencing amplicons from the BioMarKs consortium (www.biomarks.eu, last accessed February 12, 2014) were obtained from samples taken at the following sampling sites and on the following dates: Blanes, Spain (2010); Gijon, Spain (2010); Naples, Italy (2009 and 2010); Oslo, Norway (2009 and 2010); Roscoff, France (2010); and Varna, Bulgaria (2010). Collection, amplification, 454 pyrosequencing, and data cleaning methods followed Logares et al. (2012). Amplicons were grouped into MOTUs in JAguc (Nebel, Wild, et al. 2011), using a 98% similarity value following Nebel, Pfabel, et al. (2011). There are many ways to handle rare sequences and MOTUs composed of one to few amplicons (Gobet et al. 2010; Kunin et al. 2010; Behnke et al. 2011); however, given the low number of resulting amplicons that Blast to the Colpodea and because the amplicons are similar to GenBank accessions of morphologically identified and Sanger sequenced Colpodea, we kept all single singletons (i.e., those MOTUs with just one amplicon). A MOTU representative (the most abundant) was then blasted, using JAguc, to the reference taxonomic database of 308 ciliates from each of the 12 major ciliate clades (supplementary table S1, Supplementary Material online). The colpodean amplicons are deposited at the European Nucleotide Archive (accession number PRJEB5048).

One representative from each of the 22 MOTUs was then aligned to its respective best Blast hit to the alignment from Foissner et al. (2011), with the entire V4 region included. Placements of the amplicons used the pairwise alignment option in MacClade, and the results were checked and modified by eye. A RAxML tree was inferred using the GTR-Γ model (with the GTR-I-Γ model resulting in the same topology—data not shown). Alternative phylogenetic placements of the MOTUs were calculated with the EPA from Berger et al. (2011), using the reference tree from Foissner et al. (2011). Support for alternative phylogenetic placements of the MOTUs was estimated by the EPA using likelihood weights: placements with ≥95% were considered to be of high confidence, while <95% was considered to be of low confidence.

Using the tree inferred above with the environmental MOTUS, a parsimony-based ancestral trait reconstruction was performed with Mesquite v2.75 (Maddison WP and Maddison DR 2011) and default parameters. Terrestrial and freshwater was coded as one character state; given that ciliates can only be metabolically active when there is water, there is not much difference in water, soil, and pond. Marine was coded as the second state.

Supplementary Material

Supplementary table S1, file S1, and figure S1 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank the editors and two anonymous reviewers for helpful comments and suggestions. This work was supported by the Deutsche Forschungsgemeinschaft (DFG, grant DU1319/1-1) to M.D.; DFG (grant STA/860-3) to A.Sta.; European Funding Agencies from the ERAnet program BiodivERsA under the BioMarKs project (grant 2008-6530) to C.dV, T.S., and the BioMarKS Consortium; Carl-Zeiss postdoc grant to A.Sto.; and DFG (grants STO414/3-1 and STO414/3-2) to T.S.

Appendix A. PairDist

PairDist's Logic

The classical NJ approach operates on a matrix of pairwise distances calculated from a multiple sequence alignment. In such an alignment, sequences are aligned, usually by a software package, in such a way so that the overall mismatches among all sequences are minimized according to an optimization criterion. From a computational point of view, multiple sequence alignments are an *n*-complete problem for which an optimal solution cannot be achieved in a reasonable time frame. Many software packages with a variety of alignment strategies, optimization criteria, and numerous user-definable parameters exist.

A key problem in alignments is the question of positional homology—which nucleotide or amino acid positions are homologous to each other, and thus suitable for comparison,

for example, when calculating a maximum likelihood distance of two sequences. For regions in the alignment with a very high level of substitutions–deletions, such as seen in hypervariable gene regions, statements of positional homology can be highly speculative, and many equal or nearly equal solutions may exist. As a consequence, such regions are often excluded from analyses, because wrong assumptions about positional homology can have deteriorating effects on both precision and accuracy of the results. On the other hand, excluding data loss information as an inability to generate a multiple sequence alignment for certain regions does not necessarily mean that those regions do not contain valuable information.

PairDist, a program developed by Frank Kauff, is an attempt to overcome this problem for nucleotide sequences that were previously problematic by reducing the data to smaller taxonomic sets so they are more easily alignable. Pairdist.py is a python script that connects the commands *clustalw2* from the ClustalW package (Thompson et al. 1994) with *dnadist* and *neighbor* from PHYLIP (Felsenstein 2005). Rather than calculating sequence distance from a full multiple sequence alignment, each sequence pair is aligned independently (with *clustalw2*) before the calculation of the Maximum Likelihood distance with *dnadist*. From the resulting pairwise distances, a matrix is generated, which serves as an input for *neighbor*. Neighbor finally calculates an NJ tree. A bootstrap option is available, where the two-sequence alignment is bootstrapped before distance calculation. Modifying this script to handle protein sequences will occur in a later release.

Requirements

In order to run PairDist, other software needs to be installed on your system. First, the python module Biopython (www.biopython.org, last accessed February 12, 2014, Cock et al. 2009) is required for pairdist.py. Pairdist has been tested with version 1.61; newer are likely to work as well. Second, from the PHYLIP software package (http://www.phylippackage.org, last accessed February 12, 2014), the commands *dnadist* and *neighbor* are needed. Third, from the ClustalW package (http://www.clustal.org/clustal2, last accessed February 12, 2014), the command *clustalw2* is highly recommended. An alternative for *clustalw2* is available in the Biopython package and integrated in pairdist.py, but execution time is greatly decreased when *clustalw2* is not available.

After a standard installation of Biopython, the PHYLIP package, and ClustalW2, pairdist.py should run without changes, assuming that *clustalw2*, *dnadist*, and *neighbor* are in your path and are available systemwide. For installation details of the prerequisite software packages, please consult their respective manuals.

Installation and Use

The program pairdist.py is written in the python programming language (www.python.org, last accessed February 12, 2014) and available for download at: https://github.com/frederic-mahe/pairdist (last accessed February 12, 2014). Unzip the package and copy the executable pairdist.py either into the folder where your data lives or in any location of your system path, e.g. /bin, /usr/bin, or /usr/local/bin or in most Linux or Mac systems. The details may vary according to the specific setup of your computer.

Pairdist.py is a simple command line tool. Given an input file in FASTA format, the program is called as

```
python pairdist.py <sequences.fas>
```

where <sequences.fas> is to be replaced by the file name of your input FASTA file.

Among various intermediate files produced by the script, the resulting tree in Newick format is written to a file that has the same name as your input file with the suffix ".tree" added. It can be read as input and displayed by most applications for visualization of phylogenetic trees, for example, FigTree (Rambaut 2006).

With the options -b and -n, a bootstrap run is performed with a number of replicates specified with -n, for example,

```
python pairdist.py <sequences.fas> -b -n 100
```

will, in addition to the NJ tree, also calculate 100 bootstrap replicates, written to the file pairdist_bootstrap.trees. Bootstrap trees and NJ tree are then merged into a single output file, named as above. This file contains the NJ tree with branch lengths together with the bootstrap frequencies and can be displayed using FigTree and other programs.

References

Adl SM, Simpson AG, Lane CE, Lukes J, Bass D, Bowser SS, Brown M, Burki F, Dunthorn M, Hampl V, et al. 2012. Revised classification of the protists. *J Eukaryot Microbiol.* 59:429–493.

Agatha S, Strüder-Kypke M. 2007. Phylogeny of the order Choreotrichida (Ciliophora, Spriotricha, Oligotrichea) as inferred from morphology, ultrastructure, ontogenesis, and SSrRNA sequences. *Eur J Protistol.* 43:37–63.

Alfaro ME, Zoller S, Lutzoni F. 2003. Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and boostrapping in assessing phylogenetic confidence. *Mol Biol Evol.* 20:255–266.

Amaral-Zettler LA, McCliment EA, Ducklow HW, Huse SM. 2009. A method for studying protistian diversity using massively parallel sequencing of V9 hypervariable region of small-subunit ribosomal RNA genes. *PLoS One* 4:7.

Bachy C, Gómez F, López-García P, Dolan JR, Moreira D. 2012. Molecular phylogeny of tintinnid ciliates. *Protist* 163:873–887.

Behnke A, Barger KJ, Bunge J, Stoeck T. 2010. Spatio-temporal variations in protistan communities along an O_2/H_2S gradient in the anoxic Framvaren Fjord (Norway). *FEMS Microbiol Ecol.* 72:89–102.

Behnke A, Engel M, Christen R, Nebel M, Kleln RR, Stoeck T. 2011. Depicting more accurate pictures of protistan community complexity using pyrosequencing of hypervariable SSU rRNA gene regions. *Environ Microbiol.* 13:340–349.

Berger SA, Krompass D, Stamatakis A. 2011. Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood. *Syst Biol.* 60:291–302.

Berney C, Romac S, Mahé F, Santini S, Siano R, Bass D. 2013. Vampires in the oceans: predatory cercozoan amoebae in marine habitats. *ISME J.* 7:2387–2399.

Bittner L, Gobet A, Audic S, Romac S, Egge E, Santini S, Ogata H, Probert I, Edvardsen B, de Vargas C. 2013. Diversity patterns of uncultured

haptophytes unravelled by pyrosequencing in Naples Bay. *Mol Ecol.* 22:87–101.
Bourland WA, Hampikian G, Vd'ačný P. 2012. Morphology and phylogeny of a new woodruffiid ciliate, *Etoschophrya inornata* sp.n. (Ciliophora, Colpodea, Platyophryida), with an account on evolution of platyophryids. *Zool Scripta.* 41:400–416.
Bourland WA, Vd'ačný P, Davis MC, Hampikian G. 2011. Morphology, morphometrics and molecular characterization of *Bryophrya gemmea* n. sp. (Ciliophora, Colpodea): implications for the phylogeny and evolutionary scenario for the formation of oral ciliature in order Colpodida. *J Eukaryot Microbiol.* 58: 22–36.
Bourland WA, Wendell L, Hampikian G, Vd'ačný P. 2013. Morphology and phylogeny of *Bryophryoides ocellatus* n. g., n. sp. (Ciliophora, Colpodea) from in situ soil percolates of Idaho, U.S.A. *Eur J Protistol.* 50:47–67.
Bråte J, Logares R, Berney C, Ree DK, Klaveness D, Jakobsen KS, Shalchian-Tabrizi K. 2010. Freshwater Perkinsea and marine-freshwater colonizations revealed by pyrosequencing and phylogeny of environmental rDNA. *ISME J.* 4:1144–1153.
Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol.* 17: 540–552.
Cock PJA, Antao T, Chang AT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25:1422–1423.
Doherty M, Costas B, McManus GB, Katz LA. 2007. Culture-independent assessment of planktonic ciliate diversity in coastal northwest Atlantic water. *Aquat Microb Ecol.* 48:141–154.
Dunthorn M, Eppinger M, Schwarz MVJ, Schweikert M, Boenigk J, Katz LA, Stoeck T. 2009. Phylogenetic placement of the Cyrtolophosididae Stokes, 1888 (Ciliophora; Colpodea) and neotypification of *Aristerostoma marinum* Kahl, 1931. *Int J Syst Evol Microbiol.* 59:167–180.
Dunthorn M, Foissner W, Katz LA. 2008. Molecular phylogenetic analysis of class Colpodea (phylum Ciliophora) using broad taxon sampling. *Mol Phylogenet Evol.* 48:316–327.
Dunthorn M, Foissner W, Katz LA. 2011. Expanding character sampling in ciliate phylogenetic inference using mitochondrial SSU-rDNA as a molecular marker. *Protist* 162:85–99.
Dunthorn M, Katz LA. 2010. Secretive ciliates and putative asexuality in microbial eukaryotes. *Trends Microbiol.* 18:183–188.
Dunthorn M, Katz LA, Stoeck T, Foissner W. 2012. Congruence and indifference between two molecular markers for understanding oral evolution in the Marynidae *sensu lato* (Ciliophora, Colpodea). *Eur J Protistol.* 48:297–304.
Dunthorn M, Klier J, Bunge J, Stoeck T. 2012. Comparing the hypervariable V4 and V9 regions for assesment of ciliate environmental diversity. *J Eukaryot Microbiol.* 59:185–187.
Dunthorn M, Stoeck T, Wolf K, Breiner H-W, Foissner W. 2012. Diversity and endemism of ciliates inhabiting Neotropical phytotelmata. *Syst Biodivers.* 10:195–205.
Eddy SR. 1998. Profile hidden Markov models. *Bioinformatics* 14: 755–763.
Egge E, Bittner L, Andersen T, Audic S, de Vargas C, Edvardsen B. 2013. 454 pyrosequencing to describe microbial eukaryotic community composition, diversity and relative abundance: a test for marine haptophytes. *PLoS One* 8:e74371.
Felsenstein J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6: Distributed by the author. Seattle (WA): Department of Genome Sciences, University of Washington.
Fenchel T, Finlay BJ. 2006. The diversity of microbes: resurgence of the phenotype. *Phil Trans R Soc B.* 361:1965–1973.
Finlay BJ. 2002. Global dispersal of free-living microbial eukaryote species. *Science* 296:1061–1063.
Finlay BJ, Esteban GF, Brown S, Fenchel T, Hoef-Emden K. 2006. Multiple cosmopolitan ecotypes within a microbial eukaryote morphospecies. *Protist* 157:377–390.
Foissner W. 1987. Soil protozoa: fundamental problems, ecological significance, adaptations in ciliates and testaceans, bioindicators, and guide to the literature. *Prog Protistol.* 2:69–212.
Foissner W. 1993. Colpodea (Ciliophora). *Protozoenfauna* 4/1:i–x, 1–798.
Foissner W. 1999. Description of two new, mycophagous soil ciliates (Ciliophora, Colpodea): *Fungiphrya strobli* n. g., n. sp. and *Grossglockneria ovata* n. sp. *J Eukaryot Microbiol.* 46:34–42.
Foissner W, Agatha S, Berger H. 2002. Soil ciliates (Protozoa, Ciliophora) from Namibia (Southwest Africa), with emphasis on two contrasting environments, the Etosha Region and the Namib Desert. *Denisia* 5:1–1459.
Foissner W, Bourland WA, Wolf K, Stoeck T, Dunthorn M. 2013. New SSU-rDNA sequences for eleven colpodeans (Ciliophora, Colpodea) and description of *Apocyrtolophosis* nov. gen. *Eur J Protistol.* 50: 40–46.
Foissner W, Chao A, Katz LA. 2008. Diversity and geographic distribution of ciliates (Protista: Ciliophora). *Biodivers. Conserv.* 17: 345–363.
Foissner W, Stoeck T. 2009. Morphological and molecular characterization of a new protist family, Sandmanniellidae n. fam. (Ciliophora, Colpodea), with description of *Sandmanniella terricola* n. g., n. sp. from the Chobe floodplain in Botswana. *J Eukaryot Microbiol.* 56: 472–483.
Foissner W, Stoeck T, Agatha S, Dunthorn M. 2011. Intraclass evolution and classification of the Colpodea (Ciliophora). *J Eukaryot Microbiol.* 58:397–415.
Forster D, Behnke A, Stoeck T. 2012. Meta-analyses of environmental sequence data identify anoxia and salinity as parameters shaping ciliate communities. *Syst Biodivers.* 10:277–288.
Garcia-Cuetos L, Moestrup. Ø, Hansen PJ. 2012. Studies on the genus *Mesodinium*. II. Ultrastructural and molecular investigations of five marine species help clarifying the taxonomy. *J Eukaryot Microbiol.* 59: 374–400.
Gobet A, Quince C, Ramette A. 2010. Multivariate cutoff level analysis (MultiCoLA) of large community data sets. *Nucleic Acids Res.* 38: e155.
Gong J, Stoeck T, Yi Z, Miao M, Zhang N, Roberts DM, Warren A, Song W. 2009. Small subunit rDNA phylogenies show that the class Nassophorea is not monophyletic (Phylum Ciliophora). *J Eukaryot Microbiol.* 56:339–347.
Hillis DM, Bull JJ. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Syst Biol.* 42: 182–192.
Holzmann M, Habura A, Giles H, Bowser SS, Pawlowski J. 2003. Freshwater foraminiferans revealed by analysis of environmental DNA samples. *J Eukaryot Microbiol.* 50:135–139.
Huber JA, Morrison HG, Huse SM, Neal PR, Sogin ML, Mark Welch DB. 2009. Effect of PCR amplicon size on assessments of clone library microbial diversity and community structure. *Environ Microbiol.* 11: 1292–1302.
Johnson MD, Tengs T, Oldach DW, Delwiche CF, Stoecker DK. 2004. Highly divergent SSU rRNA genes found in the marine ciliates *Myrionecta rubra* and *Mesodinium pulex*. *Protist* 155:347–359.
Kahl A. 1933. Ciliata libera et ectocommensalia. In: Grimpe G, Wagler E, editors. Die Tierwelt der Nordund Ostsee. Lief. 23 (Teil, II, c3). Leipzig, pp. 147–183.
Katz LA, McManus GB, Snoeyenbos-West OLO, Griffin A, Pirog K, Costas B, Foissner W. 2005. Reframing the 'Everything is everywhere' debate: evidence for high gene flow and diversity in ciliate morphospecies. *Aquat Microb Ecol.* 41:55–65.
Kiesselbach A. 1936. Zur Ciliatenfauna der nördlichen Adria. *Thalassia* 2: 1–53.
Kreutz M, Foissner W. 2006. The sphagnum ponds of Simmelried in Germany: a biodiversity hot-spot for microscopic organisms. *Protozoolog Monogr.* 3:1–267.
Kunin V, Engelbrektson A, Ochman H, Hugenholtz P. 2010. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ Microbiol.* 12: 118–123.

Lara E, Berney C, Harms H, Chatzinotas A. 2007. Cultivation-independent analysis reveals a shift in ciliate 18s rRNA gene diversity in a polycyclic aromatic hydrocarbon-polluted soil. *FEMS Microbiol Ecol.* 62: 365–373.

Lara E, Mitchell EAD, Moreira D, López-García P. 2011. High diverse and seasonally dynamic protist community in a pristrine peat bog. *Protist* 162:14–32.

Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* 25:2286–2288.

Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol.* 21:1095–1109.

Lasek-Nesselquist E, Katz LA. 2001. Phylogenetic position of *Sorogena stoianovitchae* and relationships within the class Colpodea (Ciliophora) based on SSU rDNA sequences. *J Eukaryot Microbiol.* 48:604–607.

Lecroq B, Lejzerowicz F, Bachar D, Christen R, Esling P, Baerlocher L, Østerås M, Farinelli L, Pawlowski J. 2011. Ultra-deep sequencing of foraminiferal microbarcodes unveils hiddern richness of early monothalamous lineages in deep-sea sediments. *Proc Natl Acad Sci U S A.* 108:13177–13182.

Logares R, Audic S, Santini S, Pernice MS, de Vargas C, Massana R. 2012. Diversity patterns and activity of uncultured marine heterotrophic flahellates unveiled with pyrosequencing. *ISME J.* 6: 1823–1833.

Logares R, Bråte J, Bertilsson S, Clasen JL, Shalchian-Tabrizi K, Rengefors K. 2009. Infrequent marine-freshwater transitions in the microbial world. *Trends Microbiol.* 17:414–422.

Löytynoja A, Milinkovitch MC. 2001. SOAP, cleaning multiple alignments from unstable blocks. *Bioinformatics* 17:573–574.

Lozupone C, Hamady M, Knight R. 2006. UniFrac—an online tool for comparing microbial communuty diversity in a phylogenetic context. *BMC Bioinformatics* 7:371.

Lynn DH. 1976. Comparative ultrastructure and systematics of Colpodida. Structural conservatism hypothesis and a description of *Colpoda steinii* Maupas. *J Protozool.* 23:302–314.

Lynn DH. 2003. Morphology or molecules: how do we identify the major lineages of ciliates (Phylum Ciliophora)? *Eur J Protistol* 39:356–364.

Lynn DH. 2008. The ciliated protozoa: characterization, classification, and guide to the literature, 3rd ed. Dordrecht (The Netherlands): Springer.

Lynn DH, Wright ADG, Schlegel M, Foissner W. 1999. Phylogenetic relationships of orders within the class Colpodea (phylum Ciliophora) inferred from small subunit rRNA gene sequences. *J Mol Evol.* 48: 605–614.

Maddison DR, Maddison WP 2003. McClade. Version 4.0. Sunderland (MA): Sinauer Associates.

Maddison WP, Maddison DR. 2011. Mesquite: a modular system for evolutionary analysis. Version 2.75. [cited 2014 Feb 12]. Available from: http://mesquiteproject.org.

Massana R, Pedrós-Alió C. 2008. Unveiling new microbial eukaryotes in the surface ocean. *Cur Opin Microbiol.* 11:213–218.

Miao M, Shao C, Jiang J, Li L, Stoeck S, Song W. 2009. *Caryotricha minuta* (Xu et al., 2008) nov. comb., a unique marine ciliate (Protista, Ciliophora, Spirotrichea), with phylogenetic analysis of the ambiguous genus *Caryotricha* inferred from the small-subunit rRNA gene sequence. *Int J Syst Evol Microbiol.* 59:430–438.

Nebel M, Pfabel C, Stock A, Dunthorn M, Stoeck T. 2011. Delimiting operational taxonomic units for assessing ciliate environmental diversity of small subunit rRNA gene sequences. *Environ Microbiol Rep.* 3:154–158.

Nebel M, Wild S, Holzhauser M, Hüttenberger L, Reitzig R, Sperber M, Stoeck T. 2011. JAguc—a software package for environmental diversity estimates. *J Bioinform Comput Biol.* 9:749–773.

Nolte V, Pandey RV, Jost S, Medinger R, Ottenwälder B, Boenigk J, Schlötterer C. 2010. Contrasting seasonal niche separation between rare and abundant taxa conceals the extent of protist diversity. *Mol Ecol.* 19:2908–2915.

Not F, del Campo J, Balagué V, de Vargas C, Massana R. 2009. New insights into the diversity of marine picoeukaryotes. *PLoS One* 4: e7143.

Nylander JA 2004. MrModeltest v2. Distrubuted by the author. Uppsala (Sweden): Evolutionary Biology Center, Uppsala University.

Orsi W, Edgcomb V, Faria J, Foissner W, Fowle WH, Hohman T, Suarez P, Taylor C, Taylor GT, Vd'ačný P, et al. 2012. Class Cariacotrichea, a novel ciliate taxon from the anoxic Cariaco Basin, Venezuela. *Int J Syst Evol Microbiol.* 62:1425–1433.

Orsi W, Edgcomb V, Jeon S, Leslin C, Bunge J, Taylor GT, Varela R, Epstein S. 2011. Protistan microbial observatory in the Cariaco Basin, Caribbean. II. Habitat specialization. *ISME J.* 5: 1357–1373.

Orsi W, Song YG, Hallam S, Edgcomb V. 2012. Effect of oxygen munium zone formation on cummunites of marine protists. *ISME J.* 6: 1586–1601.

Pawlowski J, Christen R, Lecroq B, Bachar D, Shahbazkia HR, Amaral-Zettler L, Guillou L. 2011. Eukaryotic richness in the abyss: insights from pyrotag sequencing. *PLoS One* 6:e18169.

Pawlowski J, Lecroq B. 2010. Short rDNA barcodes for species identification in Foraminifera. *J Eukaryot Microbiol.* 57:197–205.

Penn O, Privman E, Ashkenazy H, Landan G, Graur D, Pupko T. 2010. GUIDANCE: a web server for asessing alignment confidence scores. *Nucleic Acids Res.* 38:W23–W28.

Penn O, Privman E, Landan G, Graur D, Pupko T. 2010. An alignment confidence score capturing robustness to guide tree uncertainty. *Mol Biol Evol.* 27:1759–1767.

Pernice MC, Logares R, Guillou L, Massana R. 2013. General patterns of diversity in major marine microeukaryote lineages. *PLoS One* 8: e57170.

Phadke SS, Zufall RA. 2009. Rapid diversification of mating systems in ciliates. *Biol J Linn Soc.* 98:187–197.

Quintela-Alonso P, Nitsche F, Arndt H. 2011. Molecular characterization and revised systematics of *Microdiaphanosoma arcuatum* (Ciliophora, Colpodea). *J Eukaryot Microbiol.* 58:114–119.

Rambaut A 2006. FigTree. Institute of Evolutionary Biology, University of Edinburgh. [cited 2014 Feb 12]. Available from: http://tree.bio.ed.ac.uk/software/figtree.

Richards TA, Vepritskiy AA, Gouliamova DE, Nierzwicki-Bauer SA. 2005. The molecular diversity of freshwater picoeukaryotes from an oligotrophic lake reveals diverse, distinctive and globally dispersed lineages. *Environ Microbiol.* 7:1413–1425.

Riley JL, Katz LA. 2001. Widespread distribution of extensive genome fragmentation in ciliates. *Mol Biol Evol.* 18:1372–1377.

Ronquist FR, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19: 1572–1574.

Scheckenbach F, Hausmann K, Wylezich C, Weitere M, Arndt H. 2010. Large-scale patterns in biodiversity of microbial eukaryotes from the abyssal sea floor. *Proc Natl Acad Sci U S A.* 107:115–120.

Schmidt SL, Foissner W, Schlegel M, Bernhard D. 2007. Molecular phylogeny of the Heterotrichea (Ciliophora, postciliodesmatophora) based on small subunit rRNA gene sequences. *J Eukaryot Microbiol.* 54:358–363.

Simon M, López-García P, Moreira D, Jardillier L. 2013. New haptophyte lineages and multiple independent colonizations of freshwater ecosystems. *Environ Microbiol Rep.* 5:322–332.

Small EB, Lynn DH. 1981. A new macrosystem for the phylum Ciliophora Doflein, 1901. *Biosystems* 14:387–401.

Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, Neal PR, Arrieta JM. 2006. Microbial diversity in the deep sea and the underexplored "rare biospshere". *Proc Natl Acad Sci U S A.* 103: 12115–12120.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.

Stoeck T, Anke B, Christen R, Amaral-Zettler L, Rodriguez-Mora MJ, Chistoserdov A, Orsi W, Edgcomb VP. 2009. Massively parallel tag

sequencing reveals the complexity of anaerobic marine protistan communities. *BMC Biol.* 7:72.

Stoeck T, Bass D, Nebel M, Christen R, Jones MDM, Breiner HW, Richards TA. 2010. Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Mol Ecol.* 19:21–31.

Stoeck T, Taylor GT, Epstein SS. 2003. Novel eukaryotes from the permanently anoix Carico Basin (Caribbean Sea). *Appl Environ Microbiol.* 69:5656–5663.

Stoeck T, Zuendorf A, Breiner HW, Behnke A. 2007. A molecular approach to identify active microbes in environmental eukaryotic clone libraries. *Microb Ecol.* 53:328–339.

Strüder-Kypke M, Wright A-DG, Foissner W, Lynn DH. 2006. Molecular phylogeny of Litostome ciliates (Ciliophora, Litostomatea) with emphasis on free-living haptorian genera. *Protist* 157:261–278.

Swofford DL, Olsen GJ, Waddell PJ, Hillis DM. 1996. Phylogenetic inference. In: Hillis DM, Moritz C, Mable BK, editors. . Molecular systematics. Sunderland (MA): Sinauer Associates. p. 407–514.

Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol.* 56:564–577.

Thompson JD, Higgins DG, Gibson TJ. 1994. Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22:4673–4680.

Utz LRP, Eizirik E. 2007. Molecular phylogenetics of subclass Peritrichia (Ciliophora: Oligohymenophorea) based on expanded analyses of 18s rRNA sequences. *J Eukaryot Microbiol.* 54:303–305.

Vďačný P, Orsi W, Foissner W. 2010. Molecular and morphological evidence for a sister group relationship of the classes Armophorea and Litostomatea (Ciliophora, Intramacronucleata, *Lamellicorticata infraphyl*. nov.), with an account of basal litostomateans. *Eur J Protistol.* 46:298–309.

von der Heyden S, Chao E, Cavalier-Smith T. 2004. Genetic diversity of goniomonads: an ancient divergence between marine and freshwater species. *Eur J Phycol.* 39:343–350.

Wuyts J, Perriere G, de Peer YV. 2004. The European ribosomal RNA database. *Nucleic Acids Res.* 32:D101–D103.

Yi Z, Dunthorn M, Song W, Stoeck T. 2010. Increased taxon sampling using both unidentified environmental sequences and identified cultures improves phylogenetic inference in the Prorodontida (Ciliophora, Prostomatea). *Mol Phylogenet Evol.* 57:937–941.

Zhan Z, Xu K, Dunthorn M. 2013. Evaluating molecular support for and against the monophyly of the Peritrichia and phylogenetic relationships within the Mobilida (Ciliophora, Oligohymenophorea). *Zool Scripta* 42:213–226.

Zhang Q, Simpson A, Song W. 2012. Insights into the phylogeny of systematically controversial haptorian ciliates (Ciliophora, Litostomatea) based on multigene analyses. *Proc R Soc B.* 279: 2625–2635.