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 Vargas,4,5 Stéphane Audic,4,5 BioMarKs Consortium, † Alexandra Stock,1 Frank Kauff,6 and Thorsten Stoeck1

1Department of Ecology, University of Kaiserslautern, Kaiserslautern, Germany
2Scientific Computing Group, Heidelberg Institute for Theoretical Studies, Heidelberg, Germany
3Institute of Theoretical Informatics, Department of Informatics, Karlsruhe Institute of Technology, Karlsruhe, Germany
4CNRS, UMR 7144, Adaptation et Diversité en Milieu Marin, Station Biologique de Roscoff, Roscoff, France
5UPMC Université de Paris 6, UMR 7144, Station Biologique de Roscoff, Roscoff, France
6Department 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, Institut 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.

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; Bittrner 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 from European nearshore sampling sites to infer that rather than being primarily terrestrial and freshwater, colpodean ciliates may have repeatedly transitioned from terrestrial to marine environments.

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 Brate et al. (2010), who used 454 pyrosequencing to uncover new Perkinsina 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).
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 colpodan 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 2a). 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 2a 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, Caricarthrix, Colpoda, Heterotrichia, Karyorelictea, Litostomatea, Oligohymenophorea, Nassophorea, Phyllopharyngea, Plagiopyle, Prosionome, and Plagiopyle); 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 2b).

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, Caricarthrix, CONthreeP, Phyllopharyngea, and Plagiopyle. Most of the remaining taxa have a clade placement that ranges between 90% and 100%. Values for Armophorea, Nassophorea, and Prosionome 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.

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

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

<table>
<thead>
<tr>
<th>Taxon</th>
<th>Number of Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Postciliodesmatophora</td>
<td>18</td>
</tr>
<tr>
<td>Heterotrichia</td>
<td>18</td>
</tr>
<tr>
<td>Karyorelictea</td>
<td>8</td>
</tr>
<tr>
<td>Intramacronucleata</td>
<td>90</td>
</tr>
<tr>
<td>Caricarthrix + Lamellircorticata + Spirotrichea</td>
<td>2</td>
</tr>
<tr>
<td>Caricarthrix + Spirotrichea</td>
<td>3</td>
</tr>
<tr>
<td>Lamellircorticata</td>
<td>11</td>
</tr>
<tr>
<td>Armophorea</td>
<td>11</td>
</tr>
<tr>
<td>Litostomatea</td>
<td>47</td>
</tr>
<tr>
<td>CONthreeP</td>
<td>7</td>
</tr>
<tr>
<td>Colpoda</td>
<td>24</td>
</tr>
<tr>
<td>Oligohymenophorea</td>
<td>64</td>
</tr>
<tr>
<td>Nassophorea</td>
<td>10</td>
</tr>
<tr>
<td>Phyllopharyngea</td>
<td>17</td>
</tr>
<tr>
<td>Plagiopyle</td>
<td>7</td>
</tr>
<tr>
<td>Prosionome</td>
<td>10</td>
</tr>
<tr>
<td>Outgroup</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>310</td>
</tr>
</tbody>
</table>

*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.
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*, etc.
### 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.

<table>
<thead>
<tr>
<th>Taxon</th>
<th>SSU</th>
<th>SSU-V4</th>
<th>SSU-V9</th>
<th>V4</th>
<th>V9</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PairDist</td>
<td>RAxML</td>
<td>MrBayes</td>
<td>PhyloBayes</td>
<td>PairDist</td>
</tr>
<tr>
<td>Postciliodesmatophora</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>61</td>
<td>69</td>
</tr>
<tr>
<td>Heterotrichea</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Karyorelictea</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Intramacronucleata</td>
<td>99</td>
<td>99</td>
<td>99</td>
<td>99</td>
<td>99</td>
</tr>
<tr>
<td>Cariacothrix + Spirotrichea + Lamellorticata</td>
<td>61</td>
<td>97</td>
<td>97</td>
<td>98</td>
<td>98</td>
</tr>
<tr>
<td>Cariacothrix + Spirotrichea</td>
<td>95</td>
<td>95</td>
<td>95</td>
<td>91</td>
<td>91</td>
</tr>
<tr>
<td>Sprirotrichea</td>
<td>97</td>
<td>92</td>
<td>92</td>
<td>93</td>
<td>93</td>
</tr>
<tr>
<td>Lamellorticata</td>
<td>77</td>
<td>96</td>
<td>96</td>
<td>77</td>
<td>77</td>
</tr>
<tr>
<td>Armophorea</td>
<td>81</td>
<td>81</td>
<td>81</td>
<td>81</td>
<td>81</td>
</tr>
<tr>
<td>Litostomatia</td>
<td>95</td>
<td>95</td>
<td>95</td>
<td>95</td>
<td>95</td>
</tr>
<tr>
<td>CONthreeP</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Colpodea</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>83</td>
<td>83</td>
</tr>
<tr>
<td>Oligohymenophorea</td>
<td>51</td>
<td>100</td>
<td>100</td>
<td>57</td>
<td>57</td>
</tr>
<tr>
<td>Nassophorea</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td>Phyllopharyngea</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Plagiopylera</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Prostomatea</td>
<td>70</td>
<td>70</td>
<td>70</td>
<td>70</td>
<td>70</td>
</tr>
<tr>
<td>Weighted average</td>
<td>85</td>
<td>93</td>
<td>93</td>
<td>93</td>
<td>93</td>
</tr>
</tbody>
</table>

### 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 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 ≥95%. Well-supported clades are dark gray; unsupported clades are light gray. Bipartitions ≤50% are shown as dashes.
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; Heterotrichia, 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 range between 90% and 100%. 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.

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,
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 nucleotide site is with the input tree.

The first analysis with the 50-bp sliding window used the SSU alignment in which all parts of both 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 bootstrap trees from the SSU alignment.

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. 3a). 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. 3b). Each alignment included 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.

![Diagram showing average placement error from a 50-nt sliding window as implemented in RAxML.

**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 bootstrap 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.
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. 3b) than in the analysis above (fig. 3a). 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 coast 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,
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 Aristostoma, 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,
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 Zulfal 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 Kiotricha). 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 \textit{Caricatricha} is 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. \textit{Vd’ačný} et al. (2010) inferred Lamellicorticata to be monophyletic when they excluded \textit{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 \textit{M. pulex} and \textit{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 ciliate (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 mitochondrial 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 (Swoford 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 2a and fig. 2). Also, there was also little difference in well-supported bipartitions from the trees inferred using these alignments (table 2a, 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 2a 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 archeal, 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. 

Recent studies have 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 (Bräte 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, or is it polyphyletic, polyphyletic 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 colpodeans that are dominantly marine (Foissner 1993; Dunthorn et al. 2009). There are reports of Rhyposophrya aplanata also being marine (Kahl 1933; Kieselbach 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 Choreaegypri 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 Spiretrichaea. 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 Nassophorea follows Lynn (2008); that is, the Synhymenia are included in the Nassophorea and not in the Phylopharyngea. We also use Vďačný et al.’s (2010) Lamelliloricuta. Following Garcia-Cueto 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 Lyon (2008): Armophorea, Cariacothyris, Colpodea, Heterotrichhea, Karyorelictea, Litostomatea, Nassophorea, Oligohymenophorea, Phyllorypharya, 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 v0.91b (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 ≥70% of the trees (Hillis and Bull 1993); for the Bayesian posterior probabilities, splits were considered to be supported if they were ≥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:

\[
\text{Clade placement} = \frac{\text{Number of species in taxon’s largest monophyletic group}}{\text{Number of species sampled in taxon}} \times 100\%.
\]

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 Cariacothyris 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:

\[
\text{Weighted-clade placement} = \frac{\sum_{i=1}^{12} \left( \frac{\text{Number of species in taxon } \times \text{taxon’s percent clade placement}}{100} \right) \times 0.01}{12}.
\]

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:

\texttt{raxmlHPC-SSE3 -f S -s <alignment.file> -W 50 -t <tree.file>}
\texttt{-m GTR\textsc{gamma} -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 align-
ment. The first alignment used with the sliding window algo-

rithm 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 algo-

rithm results as a function of the input tree, we also executed
sliding window analyses on 200 RAxML bootstrap trees in-
ferred 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 exclud-
ing 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 BioMarkks 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 repre-
sentative (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 modi-
ified by eye. A RAxML tree was inferred using the GTR-G model
(with the GTR-I-G model resulting in the same topo-
logy—data not shown). Alternative phylogenetic placements
of the MOTUs were calculated with the EPA from Berger et al.
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 confi-
dence, 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 cili-
ates 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 sup-
ported 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 BioMarkks project (grant 2008-6530)
to C.dV, T.S., and the BioMarkKS Consortium; Carl-Zeiss post-
doc 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 soft-
ware package, in such a way so that the overall mismatches
among all sequences are minimized according to an optimi-
ization 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 and insertions–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


Foissner W, Agatha S, Berger H. 2002. Soil ciliates (Protozoa, Ciliophora) from Namibia (Southwest Africa), with emphasis on two contrasting environments, the Etoha Region and the Namib Desert. Denisia. 5:1–1459.


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