The ISME Journal (2014) 8, 115–125

www.nature.com/ismej

ORIGINAL ARTICLE

Integrating metagenomic and amplicon databases to resolve the phylogenetic and ecological diversity of the *Chlamydiae*

Ilias Lagkouvardos[1], Thomas Weinmaier[2], Federico M Lauro[3], Ricardo Cavicchioli[3], Thomas Rattei[2] and Matthias Horn[1]

[1]*Division of Microbial Ecology, Department of Microbiology and Ecosystem Science, University of Vienna, Vienna, Austria;* [2]*Division of Computational System Biology, Department of Microbiology and Ecosystem Science, University of Vienna, Vienna, Austria and* [3]*School of Biotechnology and Biomolecular Sciences, The University of New South Wales, Sydney, New South Wales, Australia*

In the era of metagenomics and amplicon sequencing, comprehensive analyses of available sequence data remain a challenge. Here we describe an approach exploiting metagenomic and amplicon data sets from public databases to elucidate phylogenetic diversity of defined microbial taxa. We investigated the phylum *Chlamydiae* whose known members are obligate intracellular bacteria that represent important pathogens of humans and animals, as well as symbionts of protists. Despite their medical relevance, our knowledge about chlamydial diversity is still scarce. Most of the nine known families are represented by only a few isolates, while previous clone library-based surveys suggested the existence of yet uncharacterized members of this phylum. Here we identified more than 22 000 high quality, non-redundant chlamydial 16S rRNA gene sequences in diverse databases, as well as 1900 putative chlamydial protein-encoding genes. Even when applying the most conservative approach, clustering of chlamydial 16S rRNA gene sequences into operational taxonomic units revealed an unexpectedly high species, genus and family-level diversity within the *Chlamydiae*, including 181 putative families. These *in silico* findings were verified experimentally in one Antarctic sample, which contained a high diversity of novel *Chlamydiae*. In our analysis, the *Rhabdochlamydiaceae*, whose known members infect arthropods, represents the most diverse and species-rich chlamydial family, followed by the protist-associated *Parachlamydiaceae*, and a putative new family (PCF8) with unknown host specificity. Available information on the origin of metagenomic samples indicated that marine environments contain the majority of the newly discovered chlamydial lineages, highlighting this environment as an important chlamydial reservoir.

The ISME Journal (2014) **8,** 115–125; doi:10.1038/ismej.2013.142; published online 15 August 2013
Subject Category: Integrated genomics and post-genomics approaches in microbial ecology
Keywords: 16S rRNA; next-generation sequencing; amplicon sequencing; metagenomics

Introduction

The introduction of methods using next-generation sequencing to microbial ecology has enabled high-throughput assessment of complex microbial communities. This is achieved by sequencing either PCR-amplified marker genes (amplicon sequencing) (Huse *et al.*, 2008) or genomic DNA from environmental samples (metagenomics) (Tyson *et al.*, 2004; Venter *et al.*, 2004). These approaches have changed how the microbial biosphere is viewed and enabled novel insights to be gained into the composition and function of diverse microbial assemblages in habitats ranging from the deep sea to the human gut (Eckburg *et al.*, 2005; Sogin *et al.*, 2006). However, a limitation to effectively utilizing these vast data sets is their distribution among disparate sequence repositories, including GenBank/EMBL/DDBJ, IMG/m, CAMERA and VAMPS, (Wheeler *et al.*, 2008; Sun *et al.*, 2011; Markowitz *et al.*, 2012) (http://vamps.mbl.edu/). This lack of consolidation hampers exploration of the total available sequence information.

Chlamydiae are an assemblage of bacteria that depend on eukaryotic host cells for their reproduction. Evidence to date indicates the phylum is represented by members that are all obligate intracellular bacteria with a unique developmental life cycle. Their best known representatives are the human pathogens *Chlamydia trachomatis* and *Chlamydia pneumonia*, which cause trachoma and

Correspondence: M Horn, Division of Microbial Ecology, Department of Microbia, University of Vienna, Althan Street 14, Vienna 1090, Austria.
E-mail: horn@microbial-ecology.net
Received 16 May 2013; revised 12 July 2013; accepted 16 July 2013; published online 15 August 2013

sexually transmitted diseases, and pneumonia, respectively (Bebear and de Barbeyrac, 2009; Burillo and Bouza, 2010; Hu *et al.*, 2010). Although these medically important chlamydiae were described in 1907 (Halberstädter and Prowazek, 1907), the phylum was only represented by the single genus *Chlamydia* until 1995. The limited perception of chlamydial diversity gradually changed with the identification of environmental chlamydiae including *Simkania negevensis* (Kahane *et al.*, 1995), *Waddlia chondrophila* (Rurangirwa *et al.*, 1999) and amoeba-associated chlamydiae like *Parachlamydia acanthamoebae, Protochlamydia amoebophila* (Fritsche *et al.*, 1993; Amann *et al.*, 1997; Collingro *et al.*, 2005b) and *Criblamydia sequanensis* (Thomas *et al.*, 2006).

Analysis of these environmental chlamydiae helped to better understand the evolution of *Chlamydiae* as a whole (Horn, 2008). It was learned that the intracellular lifestyle of chlamydiae dates back to an ancient association with early unicellular eukaryotes in the Precambrian, hundreds of millions of years ago (Greub and Raoult, 2004; Horn, 2008; Kamneva *et al.*, 2012). This ancient intracellular lifestyle specialization might have contributed to the evolution of plants by facilitating the establishment of primary plastids (Brinkman *et al.*, 2002; Huang and Gogarten, 2007). In addition, several mechanisms for host interaction developed in these early associations are still used by extant chlamydial pathogens and symbionts (Hueck, 1998). Protists have thus been suggested to have provided 'evolutionary training ground' for contemporary intracellular bacteria (Molmeret *et al.*, 2005). There is evidence that some environmental chlamydiae are associated with disease in humans and animals, and their impact on public health is a source of discussion (Corsaro and Greub, 2006; Lamoth *et al.*, 2011).

The inability to cultivate chlamydiae outside eukaryotic host cells has hampered the characterization of novel chlamydiae. Co-cultivation with amoebae has been somewhat successfully used to facilitate retrieval of chlamydiae directly from environmental samples, but differences in host specificity limit the applicability of this approach (Collingro *et al.*, 2005a; Corsaro and Venditti, 2009; Corsaro *et al.*, 2010). *Chlamydiae* have also been largely missed in traditional 16S rRNA gene-based diversity surveys based on clone libraries, mainly because of their low abundance compared with free-living bacteria, but also because many general bacterial primers used in these studies have mismatches to known chlamydial 16S rRNA genes. Thus, only the application of primer sets specifically targeting members of the *Chlamydiae* enabled the identification of additional lineages within this phylum (Horn and Wagner, 2001). Such studies showed that chlamydiae are not only more diverse than originally thought, but are present in a variety of environments (Horn, 2008; Corsaro and Venditti, 2009; Corsaro *et al.*, 2010). To date, the phylum *Chlamydiae* has nine described families that range in size (Kuo and Stephens, 2008) from the well represented *Chlamydiaceae* and *Parachlamydiaceae* to the less represented *Rhabdochlamydiaceae* (Corsaro and Venditti, 2009), *Criblamydiaceae* (Corsaro *et al.*, 2009), *Simkaniaceae* (Everett *et al.*, 1999) and *Waddliaceae* (Rurangirwa *et al.*, 1999). The families with the least number of representatives (a single species) are *Clavochlamydiaceae* (Karlsen *et al.*, 2008), *Piscichlamydiaceae* (Draghi *et al.*, 2004) and the recently discovered *Parilichlamydiaceae* (Stride *et al.*, 2013).

In this study, we introduce an approach to combine all existing metagenomic and amplicon sequence data to assess the microbial diversity and ecology of the *Chlamydiae*. To achieve this, we collected all chlamydia-like protein and 16S rRNA gene sequences from publically available sequence databases by using similarity-based searches, filtering steps and large-scale phylogenetic analyses. Our study revealed the existence of an enormous, hidden, family-level diversity of *Chlamydiae*, particularly in marine habitats, and provided insights into the genomic diversity of the different families. Our approach is applicable to other microbial taxa; it demonstrates a useful computational strategy to explore taxonomic and genomic diversity and ecology of microbes that exist in available metagenomic sequence space.

Materials and methods

Identification and analysis of putative chlamydial proteins in metagenomic data

The database SIMAP (Rattei *et al.*, 2010) integrates data from multiple major public repositories of metagenomic sequences, such as IMG/M (Markowitz *et al.*, 2012), CAMERA (Sun *et al.*, 2011) and the whole-genome shotgun section of NCBI GenBank (Wheeler *et al.*, 2008). SIMAP consistently annotates all potential protein-coding sequences of these metagenomes and currently contains about 45 million non-redundant metagenomic proteins. Metagenomic proteins in SIMAP with significant similarity to known chlamydial proteins (*E*-value $<10^{-20}$, alignment coverage $>50\%$ for both query and subject) were extracted, and phylogenetic trees were calculated with their closest homologs using PhyloGenie (Frickey and Lupas, 2004) and a maximum likelihood method (RAxML; (Stamatakis, 2006)). Phylogenetic trees were then filtered with the PhyloGenie tool PHAT for well-supported (bootstrap values $>70\%$) monophyletic chlamydial clades containing metagenomic proteins. Only metagenomic proteins from well-supported clades were considered to be of putative chlamydial origin, and information on their closest phylogenetic relatives and their environmental origin were extracted for further analysis (Figure 1). A complete description of the method is provided in the Supplementary information (Supplementary methods).

Identification and analysis of chlamydial 16S rRNA gene sequences
NCBI (Wheeler *et al.*, 2008), CAMERA (Sun *et al.*, 2011) and IMG/m (Markowitz *et al.*, 2012) were searched with megablast using a representative chlamydial 16S rRNA gene sequence as reference (*Simkania negevensis*, NR_029194). All sequences with similarity greater than 60% to the reference sequence were collected. In addition, all amplicon 16S rRNA gene sequences obtained using the 454 Titanium technology were retrieved from VAMPS and SRA (Kodama *et al.*, 2012). Redundant (identical), low quality (> 0.4% ambiguous sites (N)) and short sequences (<300 nucleotides) were removed from the combined data set, and the remaining sequences were taxonomically classified using RDP classifier (Wang *et al.*, 2007) (Figure 1). Sequences recognized as members of the phylum *Chlamydiae* with confidence above 80% were then aligned using the SINA aligner.(Pruesse *et al.*, 2012). The final data set also included 12 16S rRNA gene sequences obtained in this study by PCR analysis of a water sample from Ace Lake in Antarctica (Supplementary Methods).

Two types of analyses were carried out with the aligned 16S rRNA gene sequences (Figure 1). First, near full-length sequences (>1100 nucleotides) were selected, and their phylogenetic relationships were reconstructed using Mr Bayes (Huelsenbeck and Ronquist, 2001). The obtained reference tree was visualized with iTOL (Letunic and Bork, 2007). Second, the multiple sequence alignment containing all sequences was trimmed around the region with the highest coverage. The sequences were again filtered for length and alignment quality and then used for the calculation of Operational Taxonomic Units (OTUs) using MOTHUR (Schloss *et al.*, 2009) and ESPRIT (Sun *et al.*, 2009). OTUs were classified according to the environmental origin of the sequences they include. Size, ecological classification and relative distance between OTUs were visualized in a NMDS (non-metric multi dimension scaling) plot using R. A more detailed description of the method is provided in the supplementary information (Supplementary Methods).

Results

Chlamydial proteins in metagenomic sequence data
To explore the diversity of putative chlamydial proteins in metagenomic sequence data, we conducted a comprehensive similarity-based search coupled to extensive phylogenetic analysis. A total of 31 279 proteins from various metagenomes contained in the SIMAP database (Rattei *et al.*, 2010) were identified to be most similar to known chlamydial homologs, representing 0.12% of the total metagenomic proteins included in these metagenomes (25 847 409 non-redundant proteins). After applying conservative alignment length and *E*-value filters, 5525 putative chlamydial protein sequences remained. This reduction was mainly due to the high number of short, incomplete protein sequences typically obtained in metagenomic studies. Phylogenetic analyses of those sequences further reduced this number to 1931 proteins that clustered monophyletically with known chlamydial homologs with significant bootstrap support (>70%). These proteins formed 1012 homologous groups with an average of two proteins per group. This indicates a

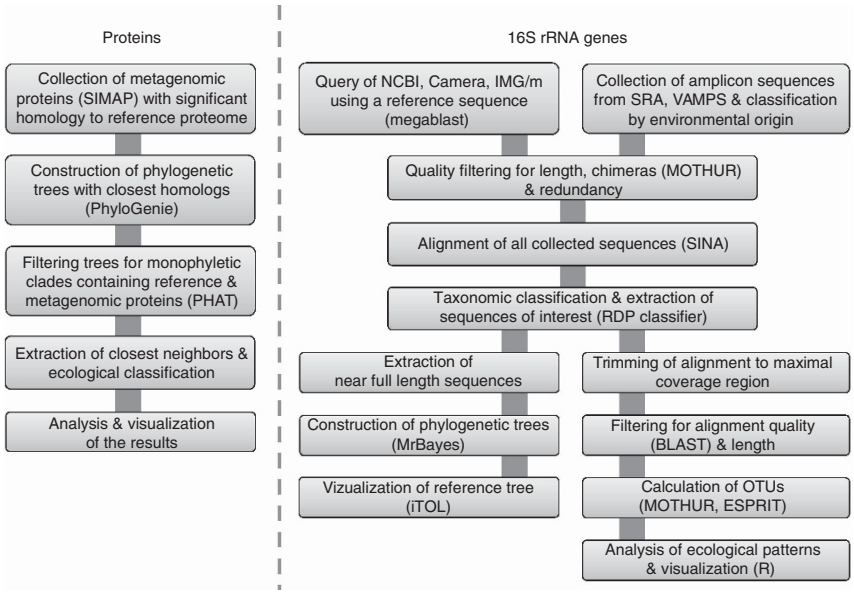

Figure 1 Flow chart illustrating the main steps in the analysis of metagenomic and amplicon sequence data for inferring diversity and ecology of defined microbial taxa. In this study, this approach was used for investigating the phylum *Chlamydiae*. A detailed description of each step is provided in supplementary information.

shallow representation, that is, a low coverage, of putative chlamydial homologs in the current extent of metagenomic sequence data.

For 392 putative chlamydial metagenomic proteins, only chlamydial homologs were detected. These proteins were classified as '*Chlamydiae* specific'. If other proteins exist with lower similarity than the criteria we used, they would have been excluded from our analysis. Within the complete set of putative chlamydial metagenomic proteins, we searched for homologs to known proteins that have been associated with host interaction and virulence of *Chlamydiae* (Collingro *et al.*, 2011). This search resulted in 76 metagenomic proteins that group with 29 virulence-associated proteins. Interestingly, at least one metagenomic protein was identified for each of the known virulence-associated proteins. Homologs of the plasmid-encoded protein pGP6 and the type III secretion system chaperone SctG were most frequently detected, with 9 and 7 metagenomic proteins, respectively (Supplementary Table S1).

Based on the closest neighbor in the phylogenetic trees, the majority of the putative chlamydial metagenomic proteins were most closely related to known proteins from members of the *Simkaniaceae* and *Parachlamydiaceae*, a trend that was also observed for the subset of '*Chlamydiae* specific' proteins (Figure 2). Noticeably, most putative chlamydial metagenomic proteins originated from marine samples (86%; Figure 2). Even considering that 60% of the total number of metagenomic proteins included in the analysis was of marine origin, this still indicates an overrepresentation of putative chlamydial proteins in those samples.

Identification of chlamydial 16S rRNA genes

To identify chlamydial 16S rRNA genes from amplicon and metagenomic studies, we integrated data from different sequence databases including VAMPS, SRA, NCBI, CAMERA and IMG/m (Wheeler *et al.*, 2008; Sun *et al.*, 2011; Kodama *et al.*, 2012; Markowitz *et al.*, 2012). A similarity-based search using relaxed criteria and subsequent taxonomic classification of the 16S rRNA gene sequences using the RDP classifier (Wang *et al.*, 2007), resulted in a set of 22 070 unique chlamydia-like 16S rRNA gene sequences with an average length of 471 nucleotides (Supplementary Table S2). Compared with the NCBI nt database alone, which is generally used to collect rRNA gene sequences for phylogenetic analysis, the inclusion of metagenomic-derived data from NCBI env, CAMERA and IMG/m more than doubled the number of chlamydial 16S rRNA gene sequences. However, despite this doubling of sequences, the vast majority (95%) of all recovered sequences originated from amplicon data sets in VAMPS and SRA (Supplementary Table S2).

A phylogenetic framework for the phylum Chlamydiae

To construct a robust phylogenetic framework for members of the *Chlamydiae*, we extracted all near full-length non-chimeric 16S rRNA gene sequences with at least 1100 nucleotides ($n = 271$) and used these for tree calculation (Figure 3). This sequence set was also used for estimation of family-level OTUs by applying a 10% distance cutoff, as proposed for the phylum *Chlamydiae* (Everett *et al.*, 1999). For clustering of the sequences into OTUs, two methods were used: ESPRIT (Sun *et al.*, 2009) and MOTHUR (Schloss *et al.*, 2009), which determine sequence similarity using pairwise alignments, and a multiple sequence alignments, respectively. The numbers of OTUs obtained with the two approaches differed. Although MOTHUR predicted 40 family-level OTUs, ESPRIT was more conservative and estimated 28 OTUs (Supplementary Table S3). Both tree topology and

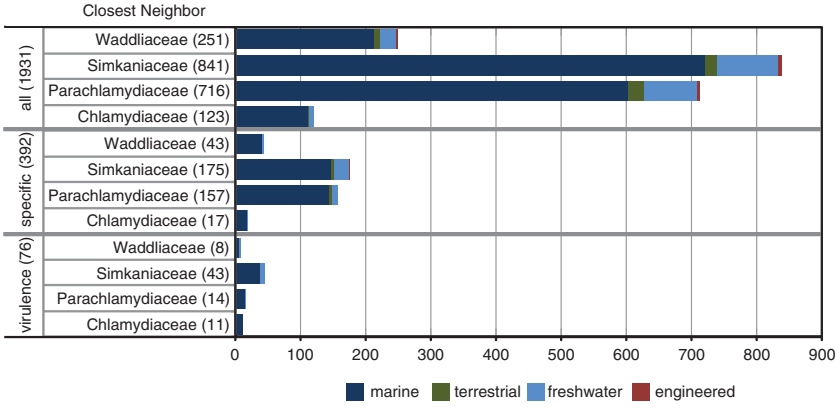

Figure 2 Ecological and taxonomic classification of putative chlamydial proteins in metagenomic sequence data. Proteins were classified based on their respective closest neighbor in maximum likelihood trees. Environmental origins grouped in four general categories are color coded. 'All' refers to all putative chlamydial proteins; 'specific' refers to the subgroup of proteins with exclusively known chlamydial homologs, 'virulence' includes all metagenomic proteins with homology to known chlamydial virulence-associated proteins. The number of proteins in each group is indicated in parenthesis. Most of the detected putative chlamydial metagenomic proteins originated from marine environments and are most similar to *Simkaniaceae* or *Parachlamydiaceae* homologs.

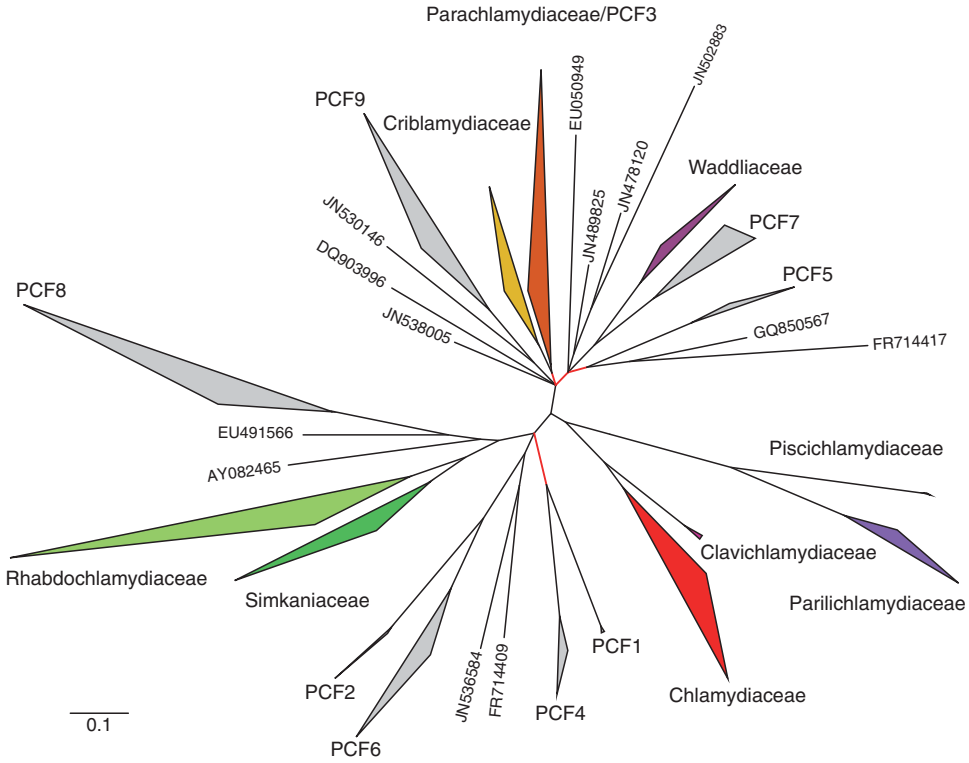

Figure 3 Relationships of described and predicted families in the phylum *Chlamydiae* based on near full-length 16S rRNA gene sequences (>1100 nt). The phylogenetic tree was calculated using Bayesian inference (MrBayes; (Huelsenbeck and Ronquist, 2001)). Branches with a posterior probability lower than 0.50 were collapsed. Those with posterior probability values between 0.50 and 0.70 are indicated with red color. The monophyly of all chlamydial families is well supported (>0.90); family level OTUs obtained by sequence similarity-based clustering with ESPRIT (Sun *et al*., 2009) and including only yet undescribed sequences are labeled as PCF. Details for the sequences included in tree calculation and clustering are available as Supplementary Table S3. Bar, 0.1 expected substitutions per site.

known chlamydial families were best represented by the grouping of sequences using ESPRIT (Supplementary Table S4, Figure 3). The only incongruence was observed for the *Criblamydiaceae* and the *Parachlamydiaceae,* which formed independent groups at a 9% distance cutoff but grouped together at 10%. In contrast, MOTHUR split the *Rhabdochlamydiaceae* into four separate groups and the *Parachlamydiaceae* into two. We thus used the more conservative approach of ESPRIT for assigning yet undescribed family-level OTUs as 'Predicted Chlamydial Families' (PCF) in the phylogenetic tree (Figure 3, Supplementary Table S4). In summary, our analysis of full-length 16S rRNA gene sequences from various databases showed that the total number of families in the *Chlamydiae* is two times higher ($n=17$) than described before, or more than three times higher ($n=28$) if singletons are considered (Figure 3, Supplementary Table S4).

The monophyly of all known chlamydial families is statistically well supported in the 16S rRNA gene-based phylogenetic tree (>0.90 posterior probability), but branching order is only partially resolved (Figure 3). Nevertheless, a phylogenetic relationship between *Parachlamydiaceae, Criblamydiaceae* and *Waddliaceae* together with PCF3, PCF5, PCF7 and PCF9 is well supported (0.97 posterior probability). Likewise, the families *Simkaniaceae*, *Rhabdochlamydiaceae* and the putative family PCF8 form a well-supported clade (0.94 posterior probability). In addition, the previously described relationships of *Clavichlamydiaceae* with *Chlamydiaceae* (Horn, 2008) and *Piscichlamydiaceae* with *Parilichlamydiaceae* (Stride *et al*., 2013) were recovered in the tree topology. We noted that three PCFs (PCF1, PCF4 and PCF2) consisted of sequences originating from a single environmental source, the marine-derived Lagoon Paola in Italy (Pizzetti *et al*., 2012).

Evidence for a vast diversity of Chlamydiae

The near full-length 16S rRNA gene sequences provide a robust framework for inferring phylogenetic relationships and diversity within the *Chlamydiae*, yet they represent only a minor fraction (1%) of all collected chlamydial 16S rRNA gene sequences. Although the majority of sequences in our data set are too short for robust phylogenetic analysis, they can be used to estimate the diversity of the phylum *Chlamydiae* using sequence similarity-based clustering into OTUs (Kim *et al*., 2011).

The meta-analysis of short 16S rRNA gene sequences derived from amplicon-based diversity surveys is complicated by the fact that not all

studies target the same regions of the 16S rRNA gene. We therefore performed a multiple sequence alignment of all 22 070 sequences collected from diverse sources, in order to identify the region with the highest coverage. Plotting these data showed that the variable region from V4 to V6 was best represented in our data (Supplementary Figure S1). We then determined whether this ~450 nucleotide length region was a good proxy for the full-length 16S rRNA gene in similarity-based OTU calculations for the phylum *Chlamydiae*. To evaluate this, the number of OTUs obtained with the full-length sequences was compared with the number of OTUs obtained with the same sequences after they were trimmed to V4 to V6. This analysis showed that the numbers of OTUs obtained with the full-length and trimmed data sets were comparable across the taxonomic levels that were resolved (Supplementary Table S3), indicating that the V4 to V6 region can be used for obtaining reasonably stringent and conservative predictions of chlamydial diversity. This is consistent with a previous study that found that the V4 to V6 region slightly underestimated diversity, predicting around 10% less OTUs compared with the full-length 16S rRNA gene for all similarity levels tested (Kim *et al.*, 2011).

After trimming and additional quality filtering, 14 311 partial 16S rRNA gene sequences remained in our data set. Removal of redundant sequences further reduced this data set to 12 636 sequences, which represented the final sequence collection used for OTU calculations. Clustering into OTUs using sequence similarity thresholds corresponding to different taxonomic levels in the phylum *Chlamydiae* (Everett *et al.*, 1999), showed that existing public metagenomic sequence data contained an as yet, undescribed, high level diversity of the *Chlamydiae* phylum (Table 1). More than 2000 OTUs were present at the species level, representing more than 250 chlamydial families.

In general, fewer OTUs were obtained with ESPRIT compared with MOTHUR (Table 1), which is consistent with our earlier observation during the analysis of full-length sequences (see above). As the pairwise alignment-based method implemented in ESPRIT resulted in more conservative diversity estimates of our data set, we only used the OTUs calculated by ESPRIT in subsequent analyses.

Table 1 Estimated diversity within the phylum *Chlamydiae* at different taxonomic levels based on clustering of partial metagenomic 16S rRNA gene sequences into OTUs

Cutoff	levels	*ESPRIT OTUs*	*MOTHUR OTUs*
Species	0.03	2031 (1161)	2276 (1378)
Genera	0.05	1236 (605)	1371 (702)
Families	0.1	262 (81)	349 (127)
Orders	0.15	17 (8)	51 (19)
Phyla	0.2	1 (0)	3 (1)

The number of singletons is indicated in parenthesis.

Insights into the ecology of Chlamydiae

Entries in public sequence databases generally contain additional information such as the origin of the investigated samples. These data can be used to analyze the environmental distribution of organisms detected in the samples. In our 16S rRNA gene data set, the majority of unique chlamydial sequences were derived from freshwater environments (67.6%), followed by marine environments (31%), while the number of sequences derived from terrestrial and engineered environments was negligible (<2%; Supplementary Figure S2). Despite this overrepresentation of freshwater sequences, at all taxonomic levels most OTUs contained only marine sequences (Supplementary Figure S2). Thus, although the number of freshwater sequences in our data set was higher, most of those sequences are more similar to each other and group in fewer OTUs than the marine sequences. This indicates that marine environments are more diverse in terms of *Chlamydiae* than freshwater or terrestrial habitats.

To illustrate the diversity of *Chlamydiae* and to visualize ecological patterns, we plotted family-level OTUs using non-parametric NMDS (Figure 4). This analysis shows that, even at the family level, there are a large number of OTUs (85% of all OTUs, Supplementary Figure S2) which contain sequences exclusively from a single environment category. This may be because these chlamydial families or their hosts are restricted to growth in specific environments. The dominance in numbers of marine OTUs (despite the majority of sequences originating from freshwater) is apparent in the NMDS plot. Marine OTUs are highly diverse and are distributed across the whole range of the plot. Yet, the largest OTUs comprising the highest numbers of unique sequences were of mixed origin. The three largest OTUs are the *Rhabdochlamydiaceae* (5004 sequences), followed by the *Parachlamydiaceae* (1834 sequences) and PCF8 (1594 sequences).

Experimental verification of chlamydial diversity in an Antarctic sample

We noted that among the samples included in this study, several contained a high diversity of novel family-level *Chlamydiae*. For example, a number of diverse chlamydial 16S rRNA gene sequences originated from the marine-derived Ace Lake in Antarctica (Lauro *et al.*, 2011). We thus chose this sample to evaluate whether the diversity of *Chlamydiae* predicted by our analysis could be confirmed experimentally. For this, we performed PCR using a *Chlamydiae*-specific primer set amplifying almost the complete 16S rRNA gene. From 25 clones showing different restriction fragment length polymorphism patterns, 12 unique chlamydial sequences were identified. All of these matched with 100% sequence similarity to partial metagenomic sequences from Ace Lake. The near full-length sequences that were obtained formed the

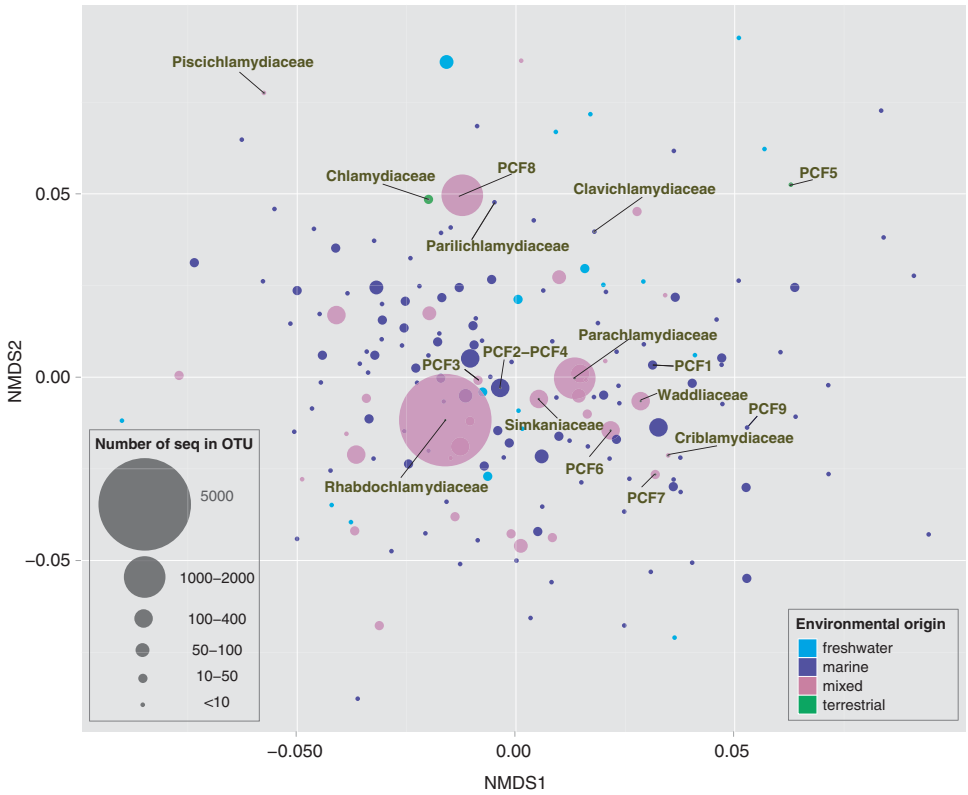

Figure 4 Diversity and ecology of chlamydial families based on NMDS of OTU distances. Filled circles represent family-level OTUs, with the size corresponding to the number of sequences included. The distance between circles indicates the relative distance between OTUs. Colors represent the environment from which the sequences that form the OTUs originated from. OTUs formed by a single sequence only (singletons) were not included in the plot. The majority of family-level OTUs contain only marine-derived sequences (dark blue circles) indicating a high diversity of marine *Chlamydiae* (see also Supplementary Figure S2). Three prominent OTUs comprise the majority of sequences, the *Rhabdochlamydiaceae*, followed by the *Parachlamydiaceae* and PCF8.

well-supported novel PCF6 clade (Figure 3), thus confirming the validity of the respective partial metagenomic sequences as being chlamydial. Therefore, the OTU classification of short metagenomic sequences correctly predicted the existence of a novel chlamydial family in the data from this lake.

Discussion

The aim of this study was to investigate the diversity of the phylum *Chlamydiae* and the genomic repertoire of its members using available sequence databases. However, there is no straight forward way to search metagenomes in public databases for proteins assigned to specific taxonomic groups. We thus used a similarity-based approach to extract an initial set of putative chlamydial proteins, and then analyzed them further using phylogenetic methods. The final set of metagenomic proteins that were classified as putative chlamydial constituted less than a tenth of the proteins originally identified by simple sequence similarity searches to known chlamydial proteins. This large reduction illustrates the uncertainty of similarity-based taxonomic classification, which is consistent with the notion that sequence similarity-based searches are often inadequate for finding the closest phylogenetic relative (Koski and Golding, 2001). However, a level of uncertainty remains even in the phylogeny-based classification of proteins. Phylogenetic monophyly of individual proteins (no matter how well supported) does not necessarily reflect organismal origin. Horizontal gene transfer between distantly related microbes or the absence of reference sequences may lead to protein phylogenies that are inconsistent with the organism tree, thereby providing mis-leading phylogenetic inference (Boucher *et al.*, 2003). Despite these limitations, the conservative set of putative chlamydial proteins identified in this study provides an improved means of evaluating the genomic diversity of *Chlamydiae*.

Compared with the total number of metagenomic proteins included in our analysis, only a small number of putative chlamydial proteins were identified, with a low redundancy in terms of homologous groups. This may indicate a low abundance of chlamydiae in the sampled environments and thus a low coverage of chlamydial genes in the available metagenomic sequence data. A low abundance of chlamydiae may reflect the fact that all known members of the *Chlamydiae* require a eukaryotic host (Horn, 2008) and thus may be expected to be rare members of microbial com-

munities. In addition, the cell size restriction imposed by many metagenome-sampling regimes (for example, 20 μm prefilter; Rusch *et al.*, 2007; Lauro *et al.*, 2011) would bias against hosts that may harbor intracellular chlamydiae.

It is difficult to isolate chlamydiae (in appropriate host cells) from environmental samples (Collingro *et al.*, 2005a; Corsaro and Venditti, 2009; Corsaro *et al.*, 2009; Hayashi *et al.*, 2010). It is thus possible that existing genome sequences of members of the *Chlamydiae* are not representative of environmental chlamydiae, as was recently reported for numerous taxa of marine bacteria by using single-cell genomics (Swan *et al.*, 2013), making it difficult to identify chlamydial genes from shotgun metagenome sequence data. Consistent with this, among the proteins identified by phylogenetic assignment as putative chlamydial, none were identical to known proteins, indicating that uncharacterized *Chlamydiae* are present in the source environments. Based on their closest relatives, the majority of these *Chlamydiae* are most closely related to known members of the *Simkaniaceae* or the *Parachlamydiaceae* (Figure 2). This either reflects the abundance of these or related families in the metagenomic samples or is an effect of the lack of reference genome sequences from other chlamydial families, such as the *Rhabdochlamydiaceae*.

To further explore the diversity of *Chlamydiae* we used the 16S rRNA gene as phylogenetic marker. Major 16S rRNA gene sequence databases such as SILVA (Pruesse *et al.*, 2007) and RDP (Cole *et al.*, 2009) mainly include sequence data from the Genbank/EMBL/DDBJ nt database, which does not contain metagenomic and amplicon sequences. In this study, we showed that collecting and integrating sequence data from different database sources is possible and facilitates a more comprehensive view of microbial diversity. In fact, 95% of the chlamydial sequences we identified originated from the VAMPS and SRA (Kodama *et al.*, 2012) databases.

Previous analyses of full-length sequences indicated that the diversity of *Chlamydiae* in these databases exceeds the diversity of described families by a factor of two to three (Corsaro *et al.*, 2003; Horn, 2008). In our present study, from the 28 family level lineages supported by full-length sequences, 21 are not represented by an isolate (Figure 3, Supplementary Table S4). The lack of matches to known members of the *Chlamydiae* was even more evident when we analyzed the complete data set of chlamydial 16S rRNA gene sequences, including also shorter sequences derived from amplicon-based studies. Even with the most conservative estimates, our analysis suggests the existence of more than 181 chlamydial families that are supported by at least two unique sequences (Table 1). Taking into account that the *Chlamydiae* included only a single family with a single genus until 1995, and only nine families until recently (Corsaro *et al.*, 2003; Horn, 2008), this discovery is highly unexpected—particularly as molecular, cultivation-independent tools for the identification of microbes has been available for more than two decades (Lane *et al.*, 1985; Amann *et al.*, 1995).

We selected one of the new family-level OTUs that was supported only by short metagenomic 16S rRNA gene sequences and analyzed the original sample using a *Chlamydiae*-specific PCR assay. The full-length sequences obtained by this experimental approach confirmed the presence of members of this OTU in the original sample. Subsequent phylogenetic analysis demonstrated that they formed an independent, family-level monophyletic group (PCF6 in Figure 3). This shows that amplicon-based OTU predictions can be verified experimentally and lends further support to the existence of the observed vast diversity of *Chlamydiae*.

All known *Chlamydiae* require a eukaryotic host for reproduction, and this lifestyle is considered an ancient feature of members of this phylum. The last common ancestor of all known *Chlamydiae* was thought to be already adapted to an intracellular lifestyle (Horn *et al.*, 2004; Kamneva *et al.*, 2012), and primordial chlamydiae might have contributed to the acquisition of primary plastids and the evolution of plants some 1.2 billion years ago (Huang and Gogarten, 2007; Ball *et al.*, 2013). If the members of the family-level chlamydial OTUs detected in our analysis have the same lifestyle as their known relatives, they also rely on eukaryotic hosts. As known chlamydiae show varying degrees of host specificity with many of them being restricted to a single host species (Horn *et al.*, 2000; Hayashi *et al.*, 2010; Coulon *et al.*, 2012), there should be a large number of eukaryotes that have not yet been identified as hosts for chlamydiae (Moonvan der Staay *et al.*, 2001). Interestingly, the most diverse chlamydial family with the highest number of unique sequences in our analysis is the *Rhabdochlamydiaceae*, whose known members infect arthropods (Kostanjsek *et al.*, 2004; Corsaro *et al.*, 2007), the most species-rich animal phylum comprising more than 80–90% of all described animals (Odegaard, 2000; Snelgrove, 2010). On the other hand, in agreement with our analysis of putative chlamydial proteins in metagenomic data sets, the majority of novel chlamydial families contain only sequences derived from marine environments, indicating an association with marine hosts. This would be consistent with the view that marine environments host an immense animal biodiversity that is comparable or even surpasses that to terrestrial habitats (Gray, 1997; Jaume and Duarte, 2006; Snelgrove, 2010).

In summary, arthropods might be important and so far neglected hosts for *Chlamydiae*, and there is a high diversity of novel, unexplored *Chlamydiae* particularly in marine environments. The absence of representative isolates for most chlamydial families and the lack of specific information about their actual hosts illustrate the huge gap we are facing in

studying and understanding chlamydial biology and evolution. Closing this gap will be a major challenge requiring the application of novel approaches and techniques such as single-cell genomics (Woyke *et al.*, 2009; Bruns *et al.*, 2010; Wang and Bodovitz, 2010; Siegl *et al.*, 2011; Li *et al.*, 2012; Stepanauskas, 2012; Seth-Smith *et al.*, 2013) and host-free cultivation and analysis of *Chlamydiae* (Haider *et al.*, 2010; Omsland *et al.*, 2013; Sixt *et al.*, 2013).

In more general terms, our study provided novel insights into the diversity and ecology of a selected group of microbes. This approach should be applicable to any other clade that is phylogenetically well defined. Standardized meta-information for metagenomics (Hirschman *et al.*, 2010; Gilbert *et al.*, 2011; Yilmaz *et al.*, 2011), and automatic retrieval and classification of publicly available sequences from different database sources would greatly facilitate this effort and would help to provide a more comprehensive and up-to-date estimate of microbial diversity.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

This work was funded by Austrian Science Fund (FWF) Grant Y277-B03 and the University of Vienna (Graduate School 'Symbiotic Interactions'). Matthias Horn acknowledges support from the European Research Council (ERC StG 'EvoChlamy'). Research in RC's laboratory is supported by the Australian Research Council and the Australian Antarctic Science Program.

References

Amann R, Springer N, Schonhuber W, Ludwig W, Schmid EN, Muller KD *et al.* (1997). Obligate intracellular bacterial parasites of acanthamoebae related to *Chlamydia* spp. *Appl Environ Microbiol* **63**: 115–121.

Amann RI, Ludwig W, Schleifer KH. (1995). Phylogenetic identification and in-situ detection of individual microbial-cells without cultivation. *Microbiol Rev* **59**: 143–169.

Ball SG, Subtil A, Bhattacharya D, Moustafa A, Weber AP, Gehre L *et al.* (2013). Metabolic effectors secreted by bacterial pathogens: essential facilitators of plastid endosymbiosis? *Plant Cell* **25**: 7–21.

Bebear C, de Barbeyrac B. (2009). Genital *Chlamydia trachomatis* infections. *Clin Microbiol Infect* **15**: 4–10.

Boucher Y, Douady CJ, Papke RT, Walsh DA, Boudreau MER, Nesbo CL *et al.* (2003). Lateral gene transfer and the origins of prokaryotic groups. *Ann Rev Genet* **37**: 283–328.

Brinkman FS, Blanchard JL, Cherkasov A, Av-Gay Y, Brunham RC, Fernandez RC *et al.* (2002). Evidence that plant-like genes in *Chlamydia* species reflect an ancestral relationship between *Chlamydiaceae*, cyanobacteria, and the chloroplast. *Genome Res* **12**: 1159–1167.

Bruns T, Becsi L, Talkenberg M, Wagner M, Weber P, Mescheder U *et al.* (2010). Microfluidic system for single cell sorting with optical tweezers. *Laser Appl Life Sci* **7376;** doi:10.1117/12.871450.

Burillo A, Bouza E. (2010). Chlamydophila pneumoniae. *Infect Dis Clin North Am* **24**: 61–71.

Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ *et al.* (2009). The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res* **37**: D141–D145.

Collingro A, Poppert S, Heinz E, Schmitz-Esser S, Essig A, Schweikert M *et al.* (2005a). Recovery of an environmental chlamydia strain from activated sludge by co-cultivation with *Acanthamoeba* sp. *Microbiol* **151**: 301–309.

Collingro A, Toenshoff ER, Taylor MW, Fritsche TR, Wagner M, Horn M. (2005b). '*Candidatus* Protochlamydia amoebophila', an endosymbiont of *Acanthamoeba* spp. *Int J Syst Evol Microbiol* **55**: 1863–1866.

Collingro A, Tischler P, Weinmaier T, Penz T, Heinz E, Brunham RC *et al.* (2011). Unity in variety - the pan-genome of the *Chlamydiae*. *Mol Biol Evol* **28**: 3253–3270.

Corsaro D, Feroldi V, Saucedo G, Ribas F, Loret JF, Greub G. (2009). Novel *Chlamydiales* strains isolated from a water treatment plant. *Environ Microbiol* **11**: 188–200.

Corsaro D, Greub G. (2006). Pathogenic potential of novel chlamydiae and diagnostic approaches to infections due to these obligate intracellular bacteria. *Clin Microbiol Rev* **19**: 283–297.

Corsaro D, Pages GS, Catalan V, Loret JF, Greub G. (2010). Biodiversity of amoebae and amoeba-associated bacteria in water treatment plants. *Int J Hygiene Environl Health* **213**: 158–166.

Corsaro D, Thomas V, Goy G, Venditti D, Radek R, Greub G. (2007). '*Candidatus* Rhabdochlamydia crassificans', an intracellular bacterial pathogen of the cockroach *Blatta orientalis* (Insecta: Blattodea). *Syst Appl Microbiol* **30**: 221–228.

Corsaro D, Valassina M, Venditti D. (2003). Increasing diversity within chlamydiae. *Critical Rev Microbiol* **29**: 37–78.

Corsaro D, Venditti D. (2009). Detection of *Chlamydiae* from freshwater environments by PCR, amoeba coculture and mixed coculture. *Res Microbiol* **160**: 547–552.

Coulon C, Eterpi M, Greub G, Collignon A, McDonnell G, Thomas V. (2012). Amoebal host range, host-free survival and disinfection susceptibility of environmental *Chlamydiae* as compared to *Chlamydia trachomatis*. *FEMS Immun Med Microbiol* **64**: 364–373.

Draghi A 2nd, Popov VL, Kahl MM, Stanton JB, Brown CC, Tsongalis GJ *et al.* (2004). Characterization of '*Candidatus* Piscichlamydia salmonis' (order *Chlamydiales*), a chlamydia-like bacterium associated with epitheliocystis in farmed Atlantic salmon (Salmo salar). *J Clin Microbiol* **42**: 5286–5297.

Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M *et al.* (2005). Diversity of the human intestinal microbial flora. *Science* **308**: 1635–1638.

Everett KD, Bush RM, Andersen AA. (1999). Emended description of the order *Chlamydiales*, proposal of *Parachlamydiaceae* fam. nov. and *Simkaniaceae* fam. nov., each containing one monotypic genus, revised

taxonomy of the family *Chlamydiaceae*, including a new genus and five new species, and standards for the identification of organisms. *Int J Syst Bacteriol* **49** Pt 2 415–440.

Frickey T, Lupas AN. (2004). PhyloGenie: automated phylome generation and analysis. *Nucleic Acids Res* **32**: 5231–5238.

Fritsche TR, Gautom RK, Seyedirashti S, Bergeron DL, Lindquist TD. (1993). Occurrence of bacterial endosymbionts in *Acanthamoeba* spp. isolated from corneal and environmental specimens and contact lenses. *J Clin Microbiol* **31**: 1122–1126.

Gilbert JA, Meyer F, Bailey MJ. (2011). The future of microbial metagenomics (or is ignorance bliss?). *ISME J* **5**: 777–779.

Gray JS. (1997). Marine biodiversity: patterns, threats and conservation needs. *Biodiv Conservation* **6**: 153–175.

Greub G, Raoult D. (2004). History of the ADP/ATP-translocase-encoding gene, a parasitism gene transferred from a Chlamydiales ancestor to plants 1 billion years ago (vol 69, pg 5530, 2003). *Appl Environ Microbiol* **70**: 6949–6949.

Haider S, Wagner M, Schmid MC, Sixt BS, Christian JG, Hacker G *et al.* (2010). Raman microspectroscopy reveals long-term extracellular activity of chlamydiae. *Mol Microbiol* **77**: 687–700.

Halberstädter L, Prowazek S. (1907). *Über Zelleinschlüsse parasitärer Natur beim Trachom*. Arbeiten aus dem Kaiserlichen Gesundheitsamte: Berlin, Germany, pp 44–47.

Hayashi Y, Nakamura S, Matsuo J, Fukumoto T, Yoshida M, Takahashi K *et al.* (2010). Host range of obligate intracellular bacterium *Parachlamydia acanthamoebae*. *Microbiol Immunol* **54**: 707–713.

Hirschman L, Sterk P, Field D, Wooley J, Cochrane G, Gilbert J *et al.* (2010). Meeting Report: 'Metagenomics, Metadata and Meta-analysis' (M3) Workshop at the Pacific Symposium on Biocomputing 2010. *Stand Genomic Sci* **vol. 2**: 357–360.

Horn M. (2008). *Chlamydiae* as symbionts in eukaryotes. *Annu Rev Microbiol* **62**: 113–131.

Horn M, Collingro A, Schmitz-Esser S, Beier CL, Purkhold U, Fartmann B *et al.* (2004). Illuminating the evolutionary history of chlamydiae. *Science* **304**: 728–730.

Horn M, Wagner M. (2001). Evidence for additional genus-level diversity of *Chlamydiales* in the environment. *FEMS Microbiol Lett* **204**: 71–74.

Horn M, Wagner M, Muller KD, Schmid EN, Fritsche TR, Schleifer KH *et al.* (2000). Neochlamydia hartmannellae gen. nov., sp nov (*Parachlamydiaceae*), an endoparasite of the amoeba *Hartmannella vermiformis*. *Microbiol* **146**: 1231–1239.

Hu VH, Harding-Esch EM, Burton MJ, Bailey RL, Kadimpeul J, Mabey DC. (2010). Epidemiology and control of trachoma: systematic review. *Trop Med Int Health* **15**: 673–691.

Huang J, Gogarten JP. (2007). Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids? *Genome Biol* **8**: R99.

Hueck CJ. (1998). Type III protein secretion systems in bacterial pathogens of animals and plants. *Microbiol Mol Biol Rev* **62**: 379–433.

Huelsenbeck JP, Ronquist F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**: 754–755.

Huse SM, Dethlefsen L, Huber JA, Welch DM, Relman DA, Sogin ML. (2008). Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *Plos Genet* **4**: e1000255.

Jaume D, Duarte CM. (2006). *General Aspects Concerning Marine and Terrestrial Biodiversity*. Fundación BBVA: Bilbao, Spain.

Kahane S, Metzer E, Friedman MG. (1995). Evidence that the novel microorganism 'Z' may belong to a new genus in the family. *Chlamydiaceae. FEMS Microbiol Lett* **126**: 203–207.

Kamneva OK, Knight SJ, Liberles DA, Ward NL. (2012). Analysis of genome content evolution in PVC bacterial super-phylum: assessment of candidate genes associated with cellular organization and lifestyle. *Genome Biol Evol* **4**: 1375–1390.

Karlsen M, Nylund A, Watanabe K, Helvik JV, Nylund S, Plarre H. (2008). Characterization of *'Candidatus* Clavochlamydia salmonicola': an intracellular bacterium infecting salmonid fish. *Environ Microbiol* **10**: 208–218.

Kim M, Morrison M, Yu Z. (2011). Evaluation of different partial 16S rRNA gene sequence regions for phylogenetic analysis of microbiomes. *J Microbiol Methods* **84**: 81–87.

Kodama Y, Shumway M, Leinonen R. (2012). The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res* **40**: D54–D56.

Koski LB, Golding GB. (2001). The closest BLAST hit is often not the nearest neighbor. *J Mol Evol* **52**: 540–542.

Kostanjsek R, Strus J, Drobne D, Avgustin G. (2004). *'Candidatus* Rhabdochlamydia porcellionis', an intracellular bacterium from the hepatopancreas of the terrestrial isopod *Porcellio scaber* (Crustacea: Isopoda). *Int J Syst Evol Microbiol* **54**: 543–549.

Kuo C-C, Stephens SR. (2008). Phylum XXIV. Chlamydiae. In: Krieg NR, Ludwig W, Whitman WB, Hedlund BP, Paster BJ, Staley JT *et al.* (eds) *Bergey's Manual of Systematic Bacteriology - The Bacteroidetes, Spirochaetes, Tenericutes (Mollicutes), Acidobacteria, Fibrobacteres, Fusobacteria, Dictyoglomi, Gemmatimonadetes, Lentisphaerae, Verrucomicrobia, Chlamydiae, and Planctomycetes*, 2nd edn Springer: New York, NY, USA, pp 843–877.

Lamoth F, Jaton K, Vaudaux B, Greub G. (2011). *Parachlamydia* and *Rhabdochlamydia*: Emerging agents of community-acquired respiratory infections in children. *Clin Infect Dis* **53**: 500–501.

Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, Pace NR. (1985). Rapid-determination of 16S ribosomal-RNA sequences for phylogenetic analyses. *Proc Natl Acad Sci USA* **82**: 6955–6959.

Lauro FM, DeMaere MZ, Yau S, Brown MV, Ng C, Wilkins D *et al.* (2011). An integrative study of a meromictic lake ecosystem in Antarctica. *ISME J* **5**: 879–895.

Letunic I, Bork P. (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**: 127–128.

Li MQ, Xu J, Romero-Gonzalez M, Banwart SA, Huang WE. (2012). Single cell Raman spectroscopy for cell sorting and imaging. *Curr Opin Biotechn* **23**: 56–63.

Markowitz VM, Chen IMA, Palaniappan K, Chu K, Szeto E, Grechkin Y *et al.* (2012). IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Res* **40**: D115–D122.

Molmeret M, Horn M, Wagner M, Santic M, Abu Kwaik Y. (2005). Amoebae as training grounds for intracellular bacterial pathogens. *App Environ Microbiol* **71**: 20–28.

Moon-van der Staay SY, De Wachter R, Vaulot D. (2001). Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature* **409**: 607–610.

Odegaard F. (2000). How many species of arthropods? Erwin's estimate revised. *Biol J Linnean Soc* **71**: 583–597.

Omsland A, Sager J, Nair V, Sturdevant DE, Hackstadt T. (2013). Developmental stage-specific metabolic and transcriptional activity of Chlamydia trachomatis in an axenic medium (vol 109, pg 19781, 2012). *Proc Natl Acad Sci USA* **110**: 1970–1970.

Pizzetti I, Fazi S, Fuchs BM, Amann R. (2012). High abundance of novel environmental chlamydiae in a Tyrrhenian coastal lake (Lago di Paola, Italy). *Environ Microbiol Rep* **4**: 446–452.

Pruesse E, Peplies J, Glockner FO. (2012). SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* **28**: 1823–1829.

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig WG, Peplies J *et al.* (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* **35**: 7188–7196.

Rattei T, Tischler P, Gotz S, Jehl MA, Hoser J, Arnold R *et al.* (2010). SIMAP-a comprehensive database of pre-calculated protein sequence similarities, domains, annotations and clusters. *Nucleic Acids Res* **38**: D223–D226.

Rurangirwa FR, Dilbeck PM, Crawford TB, McGuire TC, McElwain TF. (1999). Analysis of the 16S rRNA gene of micro-organism WSU 86-1044 from an aborted bovine foetus reveals that it is a member of the order *Chlamydiales*: proposal of *Waddliaceae* fam. nov., *Waddlia chondrophila* gen. nov., sp. nov. *Int J Syst Bacteriol* **49**: 577–581.

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S *et al.* (2007). The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol* **5**: 398–431.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB *et al.* (2009). Introducing mothur: open-Source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75**: 7537–7541.

Seth-Smith HM, Harris SR, Skilton RJ, Radebe FM, Golparian D, Shipitsyna E *et al.* (2013). Whole-genome sequences of *Chlamydia trachomatis* directly from clinical samples without culture. *Genome Res* **23**: 855–866.

Siegl A, Kamke J, Hochmuth T, Piel J, Richter M, Liang CG *et al.* (2011). Single-cell genomics reveals the lifestyle of *Poribacteria*, a candidate phylum symbiotically associated with marine sponges. *ISME J* **5**: 61–70.

Sixt B, Siegl A, Müller C, Watzka M, Wultsch A, Tziotis D *et al.* (2013). Metabolic features of Protochlamydia amoebophila elementary bodies - a link between activity and infectivity in Chlamydiae. *PLoS Pathogens* (in press).

Snelgrove PVR. (2010). *Discoveries of the Census of Marine Life: Making Ocean Life Count*. Cambridge University Press: Cambridge, New York, USA.

Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, Neal PR *et al.* (2006). Microbial diversity in the deep sea and the underexplored 'rare biosphere'. *Proc Natl Acad Sci USA* **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.

Stepanauskas R. (2012). Single cell genomics: an individual look at microbes. *Curr Opin Microbiol* **15**: 613–620.

Stride MC, Polkinghorne A, Miller TL, Groff JM, Lapatra SE, Nowak BF. (2013). Molecular characterization of 'Candidatus Parilichlamydia carangidicola,' a novel *Chlamydia*-like epitheliocystis agent in yellowtail kingfish, Seriola lalandi (Valenciennes), and the proposal of a new family, '*Candidatus* Parilichlamydiaceae' fam. nov. (order *Chlamydiales*). *Appl Environ Microbiol* **79**: 1590–1597.

Sun SL, Chen J, Li WZ, Altintas I, Lin A, Peltier S *et al.* (2011). Community cyberinfrastructure for advanced microbial ecology research and analysis: the CAMERA resource. *Nucleic Acids Res* **39**: D546–D551.

Sun YJ, Cai YP, Liu L, Yu FH, Farrell ML, McKendree W *et al.* (2009). ESPRIT: estimating species richness using large collections of 16S rRNA pyrosequences. *Nucleic Acids Res* **37**: e76.

Swan BK, Tupper B, Sczyrba A, Lauro FM, Martinez-Garcia M, Gonzalez JM *et al.* (2013). Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proc Natl Acad Sci USA* **110**: 11463–11468.

Thomas V, Casson N, Greub G. (2006). *Criblamydia sequanensis*, a new intracellular *Chlamydiales* isolated from Seine river water using amoebal co-culture. *Environ Microbiol* **8**: 2125–2135.

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM *et al.* (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**: 37–43.

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA *et al.* (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**: 66–74.

Wang DJ, Bodovitz S. (2010). Single cell analysis: the new frontier in 'omics'. *Trends Biotechnol* **28**: 281–290.

Wang Q, Garrity GM, Tiedje JM, Cole JR. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Enviro Microbiol* **73**: 5261–5267.

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V *et al.* (2008). Database resources of the national center for biotechnology information. *Nucleic Acids Res* **36**: D13–D21.

Woyke T, Xie G, Copeland A, Gonzalez JM, Han C, Kiss H *et al.* (2009). Assembling the marine metagenome, one cell at a time. *PLoS One* **4**: e5299.

Yilmaz P, Gilbert JA, Knight R, Amaral-Zettler L, Karsch-Mizrachi I, Cochrane G *et al.* (2011). The genomic standards consortium: bringing standards to life for microbial ecology. *ISME J* **5**: 1565–1567.

Supplementary Information accompanies this paper on The ISME Journal website (http://www.nature.com/ismej)