universität wien

DISSERTATION

Titel der Dissertation

Evolutionary history and phylogenetic diversity of *Chlamydiae*

verfasst von

Ilias Lagkouvardos

angestrebter akademischer Grad

Doctor of Philosophy (PhD)

Wien, 2014

Studienkennzahl lt. Studienblatt:	A 094 437
Dissertationsgebiet lt. Studienblatt:	Biologie
Betreuer:	Univ.-Prof. Dr. Matthias Horn

To all my teachers

Table of Contents

Chapter I

Preamble

Preamble

Chlamydiae are an assemblage of exclusively obligate intracellular bacteria that show a characteristic biphasic developmental cycle [1]. Their best known representatives are human pathogens but they are gradually recognized as been associated with a wide range of eukaryotic hosts spanning from unicellular protist to human [2]. The explosive increase in described hosts and diversity of the *Chlamydiae* phylum over the last decade raises the questions of what the genetic basis of this extensive symbiosis success and its actual boundaries really are. Genomic evidences support the view that at the time of speciation leading in the emergence of the chlamydia lineage the progenitor was already associated with eukaryotes [2]. This ancient relation had further implications since data suggest a role of chlamydia as mediators of the establishment of chloroplasts within the early eukaryotes that led to the emergence of modern plants [3-5]. It is becoming apparent that chlamydia had been a close partner of eukaryotes from their cradle, shaping their evolution in an extent that remains yet to be revealed. It is the goal of this work to add insights in the evolutionary history of the phylum *Chlamydiae* and to help answering the question of current phylogenetic diversity and environmental distribution of members of this intriguing phylum.

Based on 16S rRNA phylogenetic trees, it has been shown that the phylum *Chlamydiae* shares a common evolutionary origin with the phyla *Planctomycetes, Verrucomicrobia* and *Lentisphaera*. Together with the candidate phyla *Poribacteria* and *Omnitrophica* (OP3) these phyla were proposed to constitute the so called PVC superphylum [6]. In order to verify and consolidate this result and conclusively resolve the phylum's evolutionary history a comparative analysis among all members of the proposed superphylum was performed. Following the striking finding of a single protein that was uniquely shared among all members of the superphylum, the investigation was extended in the evolutionary and functional characterization of this protein. The conclusions of this part of investigation are presented in Chapter III.

By the time this thesis started, the common knowledge of the phylogenetic diversity of the phylum *Chlamydiae* was restricted to 8 described families [2]. Nevertheless,

sequence evidence for unknown chlamydial species had already pointed to the existence of far greater phylum diversity [7, 8]. Instead of embarking into sequencing hundreds of new samples, a meta-analysis approach of existing data was deemed to yield better results. Data from next generation sequencing projects of metagenomes (genomic sequencing of microbial communities) or amplicons (sequencing of amplified targeted regions like parts of the 16S rRNA gene) from thousands of samples have been accumulating for the past decade in sequence depositories like the Sequence Reads Archive [9]. Based on this realization, the primary aim was the assessment of the chlamydial diversity through the utilization of the accumulated metagenomic and amplicon wealth. The analysis goal was further extended to the actual profiling of the predicted novelty in order to identify ecological and phylogenetic patterns. Chapter IV describes the efforts spend in this direction and presents the results of these aspects of chlamydia evolution.

For a better understanding of members of the *Chlamydiae* phylum, representative isolates across all families are necessary. *Amoebae*, that act both as primary and transient hosts of several chlamydia, had served for many years as a convenient tool for isolation of chlamydia and other symbionts by isolating them directly from environmental samples or as proxy hosts in co-cultivation approaches [10]. Nevertheless the utilized isolation method was only slightly improved during the last 50 years following its introduction by Neff [11]. The realization of the vast non represented chlamydial families in the culture collections made apparent the importance of better isolation methods. In this direction a project aiming for the development of a more flexible and effective system for isolation and axenization of amoeba and their associated symbionts was undertaken and the results are presented in chapter V.

It has been suggested that amoebae serve as genomic melting pots [12, 13] facilitating the lateral gene transfer between amoeba-associated microbes. Nevertheless, the lack of genome information of chlamydia associated amoeba like *Acanthamoeba*, impairs the estimation of the amount of lateral gene transfers among amoebae and their symbionts. Furthermore, due to the fact that *Acanthamoeba* had been adopted as a model for the investigation of symbiosis and as a convenient proxy for macrophage

research the need of a quality reference genome became imperative. In order to answer these questions we joined a collaborative effort focused on the sequencing and analysis of the genome of *Acanthamoeba castellanii*, outlined in chapter VI.

Overall the set goals of this work were the elucidation of the evolutionary history of the *Chlamydiae* phylum through the investigation of its origin and the estimation of its current expansion in terms of diversity and environmental distribution. Furthermore we deemed important to invest in improving the available isolation and axenization of amoeba as a tool for the recovery of yet unknown chlamydial species. Finally, through the investigation of the genome of the versatile chlamydia host *Acanthamoeba castellanii*, the expansion of the experimental possibilities and the contextual linking with the genomic content of chlamydia will be possible.

References

1. AbdelRahman, Y.M. and R.J. Belland, *The chlamydial developmental cycle.* FEMS Microbiol Rev, 2005. **29**(5): p. 949-959.
2. Horn, M., *Chlamydiae as Symbionts in Eukaryotes.* Ann Rev Microbiol, 2008. **62**: p. 113-131.
3. Brinkman, F., et al., *Evidence that plant-like genes in Chlamydia species reflect an ancestral relationship between Chlamydiaceae, cyanobacteria, and the chloroplast.* Genome Res, 2002. **12**: p. 1159 - 1167.
4. Huang, J. and J. Gogarten, *Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids?* Genome Biol, 2007. **8**(6): p. R99.
5. Subtil, A., A. Collingro, and M. Horn, *Tracing the primordial Chlamydiae: extinct parasites of plants?* Trends Plant Sci, 2014. **19**(1): p. 36-43.
6. Wagner, M. and M. Horn, *The Planctomycetes, Verrucomicrobia, Chlamydiae and sister phyla comprise a superphylum with biotechnological and medical relevance.* Curr Opin Biotechnol, 2006. **17**(3): p. 241-9.
7. Horn, M. and M. Wagner, *Evidence for additional genus-level diversity of Chlamydiales in the environment.* FEMS Microbiol Lett, 2001. **204**(1): p. 71-74.
8. Corsaro, D., M. Valassina, and D. Venditti, *Increasing diversity within Chlamydiae.* Crit Rev Microbiol, 2003. **29**(1): p. 37-78.

9. Kodama, Y., M. Shumway, and R. Leinonen, *The sequence read archive: explosive growth of sequencing data.* Nucleic Acids Research, 2012. **40**(D1): p. D54-D56.
10. Kebbi-Beghdadi, C. and G. Greub, *Importance of amoebae as a tool to isolate amoeba-resisting microorganisms and for their ecology and evolution: the Chlamydia paradigm.* Environ Microbiol Rep, 2014.
11. NEFF, R.J., *Mechanisms of purifying amoebae by migration on agar surfaces.* Journal of Eukaryotic Microbiology, 1958. **5**(3): p. 226-231.
12. Ogata, H., et al., *Genome sequence of Rickettsia bellii illuminates the role of amoebae in gene exchanges between intracellular pathogens.* PLoS Genetics, 2006. **2**(5): p. 733-744.
13. Moliner, C., P.E. Fournier, and D. Raoult, *Genome analysis of microorganisms living in amoebae reveals a melting pot of evolution.* FEMS Microbiol Rev, 2010. **34**(3): p. 281-294.

Chapter II

Synopsis of the Publications

Synopsis of the Publications

The manuscript in **Chapter III** describes the discovery of a single protein (PVC signature protein) that is uniquely shared among all members of the *Planctomycetes*, *Verrucomicrobia* and *Chlamydiae* superphylum. This comparative genomic based finding, that effectively adds a unifying link among all members of the superphylum, was then further investigated by functionally characterizing this signature protein. The experimental analysis revealed a protein with unspecific nucleic acid binding properties that is expressed in representatives of all three major phyla. Reverse conservation analysis showed that among the bacteria-wide conserved proteins that lack homologs in members of the PVC superphylum the ribosomal protein L30 has extended similarities in its physicochemical characteristics and expression profile with the signature protein. Based on this finding, a role for the signature protein as an L30 analog was proposed. Moreover, the possible application of the signature protein as taxonomic and phylogenetic marker was explored.

Authors: **Ilias Lagkouvardos**, Marc-André Jehl, Thomas Rattei and Matthias Horn

Manuscript Title: **Signature protein of the PVC superphylum**.

Journal: Applied Environmental Microbiology **80**: 440-445 (2013).

Contributions: *All the biological experiments were performed by IL. M-AJ and TR performed the initial comparative study that identified the signature protein. TR provided a COG presence or absence matrix for all PVC and representative non-PVC bacteria. The rest of the bioinformatics analysis was performed by IL. The draft manuscript was composed by IL and then edited by all authors. MH initiated and supervised the project.*

Chapter IV contains a publication representing the most thorough sequence-based exploration of the *Chlamydiae* phylum diversity so far. Using all available genomic, metagenomic and amplicon data, estimates about the supported family, genus and species diversity were calculated. This eventually led in great expansion of our current perception of the actual chlamydial diversity in the environment. Moreover, in this analysis using the metadata related with the identified sequences it was inferred that the biggest and most diverse chlamydial families are the arthropod related *Rhabdochlamydiaceae* and the protists related *Parachlamydiaceae*. In addition the marine and marine derived habitats were shown to contain most of the so far unseen diversity indicating the role of these environments as underestimated reservoirs and targets for future chlamydia research.

Authors: **Ilias Lagkouvardos**, Thomas Weinmaier, Federico M Lauro, Ricardo Cavicchioli, Thomas Rattei and Matthias Horn.

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

Journal: ISME Journal **8**: 115-125 (2014)

***Contributions**: The data collection and analysis was performed by IL. TW and TR help with the extraction of chlamydial proteins from databases and the calculation of phylogenetic trees for each one of them. FL and RC provided DNA samples. The draft manuscript was composed by IL and edited by all authors. MH initiated and supervised the project.*

Chapter V includes a manuscript with the description of a modification of a standard amoeba isolation technique that is shown to improve the time and success of the method. This modification is based on the utilization of an *E. coli* mutant (*tolC* knockout) that is hypersensitive to antibiotics. Adding this mutant as a food source instead of the wild type that is commonly used during amoeba isolation enabled the immediate transfer of amoeba from agar plates with *E. coli* to culture media without the risk of contamination by supplementation with otherwise sub-lethal amounts of ampicillin. Such low amount of antibiotics is safe for any potential amoeba symbionts, eventually enhancing the possibility of isolates with such associations. In such an experiment it was shown that by using this method a wide variety of amoeba harboring symbionts could be isolated from two adjacent sampling sites. In many cases two phylogenetically distinct symbionts were found to reside in a single amoeba.

Authors: **Ilias Lagkouvardos**, Jie Shen and Matthias Horn

Manuscript title: **Improved axenization method reveals complexity of symbiotic associations between bacteria and acanthamoebae**

Journal: Environmental Microbiology Reports DOI: 10.1111/1758-2229.12162 (2014)

***Contributions**: The method was developed by IL. JS collected the samples and help with the time comparison experiment. The draft manuscript was composed by IL and edited by all authors. IL initiated the project and MH supervised it.*

The manuscript in **Chapter VI** represents the cumulative effort of several authors in the analysis of the genome of the free living amoeba *Acanthamoeba castellanii*. The analysis of the above mentioned genome revealed that at least 450 genes show phylogenetic evidences for prokaryotic origin, a number that exceeded all previous lateral gene transfers (LGT) in other eukaryotic organisms. These genes mostly encode proteins involved in metabolic processes and their phylogenetic profiles indicate an origin mainly from bacteria encountered in the ecological niche of the amoeba. An extensive sensory repertoire and a capacity for facultative anaerobic metabolism were also encoded in *Acanthamoeba* genome, revealing the genetic background of the flexibility of this highly successful amoeba.

Authors: Michael Clarke, Amanda J. Lohan, Bernard Liu, **Ilias Lagkouvardos**, Scott Roy, Nikhat Zafar, Claire Bertelli, Christina Schilde, Arash Kianianmomeni, Thomas R. Bürglin, Christian Frech, Bernard Turcotte, Klaus O. Kopec, John M. Synnott, Caleb Choo, Ivan Popanov, Aliza Finkler, Chris Soon Heng Tan, Andrew P. Hutchins, Thomas Weinmeier, Thomas Rattei, Jeffery S. C. Chu, Gregory Gimenez, Manuel Irimia, Daniel J. Rigden, David A. Fitzpatrick, Jacob Lorenzo-Morales, Alex Bateman, Cheng-Hsun Chiu, Petrus Tang, Peter Hegemann, Hillel Fromm, Didier Raoult, Gilbert Greub, Diego Miranda-Saavedra, Nansheng Chen, Piers Nash, Michael L. Ginger, Matthias Horn, Pauline Schaap, Lis Caler, Brendan Loftus

Manuscript title: **Genome of Acanthamoeba castellanii highlights extensive lateral gene transfer and early evolution of tyrosine kinase signaling**

Journal: Genome Biology, 14:R11 (2013)

***Contributions**: TR performed the original SIMAP searches and calculated the phylogenetic trees for the Acanthamoeba proteins. IL filtered the trees and analyzed the results of LGT. IL performed the functional, environmental and statistical analysis of the LGT result. The contributed section was composed and edited by IL, TR and MH. For more details on the contributed work please refer to the Appendix I.*

Chapter III

Signature protein of the PVC superphylum

Published in Applied Environmental Microbiology

2014

AEM
Journals.ASM.org

Signature Protein of the PVC Superphylum

Ilias Lagkouvardos,[a] Marc-André Jehl,[b] Thomas Rattei,[c] Matthias Horn[a]

Division of Microbial Ecology, Department of Microbiology and Ecosystem Science, University of Vienna, Vienna, Austria[a]; Department of Genome Oriented Bioinformatics, Wissenschaftszentrum Weihenstephan, Technische Universität München, Freising, Germany[b]; Division of Computational System Biology, Department of Microbiology and Ecosystem Science, University of Vienna, Vienna, Austria[c]

The phyla *Planctomycetes*, *Verrucomicrobia*, *Chlamydiae*, *Lentisphaerae*, and "*Candidatus* Omnitrophica (OP3)" comprise bacteria that share an ancestor but show highly diverse biological and ecological features. Together, they constitute the PVC superphylum. Using large-scale comparative genome sequence analysis, we identified a protein uniquely shared among all of the known members of the PVC superphylum. We provide evidence that this signature protein is expressed by representative members of the PVC superphylum. Its predicted structure, physicochemical characteristics, and overexpression in *Escherichia coli* and gel retardation assays with purified signature protein suggest a housekeeping function with unspecific DNA/RNA binding activity. Phylogenetic analysis demonstrated that the signature protein is a suitable phylogenetic marker for members of the PVC superphylum, and the screening of published metagenome data indicated the existence of additional PVC members. This study provides further evidence of a common evolutionary history of the PVC superphylum and presents a unique case in which a single protein serves as an evolutionary link among otherwise highly diverse members of major bacterial groups.

The bacterial phyla *Planctomycetes*, *Verrucomicrobia*, *Chlamydiae*, *Lentisphaerae*, "*Candidatus* Omnitrophica (OP3)," and "*Candidatus* Poribacteria" were proposed to share an ancestor on the basis of their monophyletic grouping in 16S rRNA-based phylogenetic trees (1). This diverse assemblage of phyla was termed the PVC superphylum and later received additional support from genomic and phylogenetic analysis of conserved proteins (2–4). Most recently, 16 housekeeping and ribosomal proteins were used to infer evolutionary relationships among the members of the PVC superphylum (5). This further established the common evolutionary origin of the members of the PVC superphylum.

Despite their common origin, the members of the PVC superphylum differ greatly with respect to life-style, physiology, and ecology (1). Each phylum includes members that attracted significant research interest because of their importance in carbon and nitrogen cycling (e.g., *Rhodopirellula* and "*Candidatus* Kuenenia" species [6, 7]), as pathogens or symbionts (e.g., *Chlamydia* and *Protochlamydia* species [8–10]), or as environmental microbes in aquatic and soil habitats (e.g., *Verrucomicrobia* [11, 12]). In addition to their ecological, biotechnological, and medical relevance, some members of the PVC superphylum show genetic and cellular features that are unusual for bacteria but reminiscent of eukaryotes or archaea (13–15). Because of these similarities, members of the PVC superphylum have been implicated in the emergence and evolution of eukaryotes, a hypothesis that is controversially discussed (14, 16–20).

In this study, we performed an extensive comparative genomic analysis in order to identify unifying links among the diverse members of the PVC superphylum. We describe the analysis and characterization of a protein, independently identified very recently (5), that is shared by all of the members of the superphylum but absent from all other bacteria. Computational analysis and functional assays provided evidence of a putative housekeeping function for this protein. Because of its conservation among the members of the PVC superphylum, we were able to use this protein to extract information about the occurrence and diversity of the members of the PVC superphylum from the available environmental metagenomes.

MATERIALS AND METHODS

Identification of the signature protein (SP). Predicted coding sequences from completely sequenced PVC and representative non-PVC genomes were obtained from the INSDC (21) and NCBI RefSeq (22) databases. All-versus-all pairwise sequence similarities were precalculated by the SIMAP database (23). From SIMAP we obtained all of the bidirectionally best-matching protein pairs (BBH) between all of the genomes, in which the alignment covered at least 50% of both protein sequences and the E value was not higher than 1e-04. The score of each BBH was additionally used as the threshold to determine inparalogs from the respective genomes. In order to cluster BBHs from the PVC superphylum into clusters of orthologous groups (COGs), we first determined all of the three-cliques (triangles) formed by PVC BBHs. Triangles were grouped into COGs if they shared a BBH. The remaining PVC BBHs were added to COGs if one of the proteins was already a member of a COG and the other was not. All of the other PVC BBHs were considered individual COGs. Inparalogs associated with BBH proteins were added to the respective COGs in all of the clustering steps mentioned above.

For each COG, we determined the presence or absence of the proteins encoded in PVC genomes. For COGs occurring in all of the PVC genomes, we determined their presence or absence in the representative non-PVC genomes from BBHs between PVC and non-PVC genomes. Only one COG, the PVC SP, was present in all of the PVC genomes and absent from all of the non-PVC genomes.

COG-based presence-or-absence analysis. The COGs of all of the bacterial genomes were obtained from the eggNOG (24) database. The BBHs between the PVC and non-PVC genomes described above were used to determine the presence or absence of each COG in the PVC genomes not yet contained in eggNOG. A matrix was then created with all of the COGs in the first column and the organism names in the top row.

Received 8 August 2013 Accepted 25 October 2013
Published ahead of print 1 November 2013
Address correspondence to Matthias Horn, horn@microbial-ecology.net.
Supplemental material for this article may be found at http://dx.doi.org/10.1128/AEM.02655-13.

doi:10.1128/AEM.02655-13

Then the table was filled with a 1 or a 0 for each COG for each genome on the basis of its presence or absence, respectively, allowing a quick overview of COG conservation across PVC and non-PVC bacteria as selective sums.

For each COG without representatives in the PVC superphylum, the *Escherichia coli* representative was found and used as the query in searches against the NCBI Refseq database (22) with BLAST (25). The first 10 proteins of nonredundant origin (different organisms) were collected. With these sets, the average protein size and isoelectric point (pI) were calculated for each COG. The pI was calculated by solving the Henderson-Hasselbach equation by a local Perl script.

Screening of metagenome data. All of the assembled metagenomes available at the JGI Genome Portal (26) were downloaded and organized into BLAST databases with makeblastdb (included in the BLAST+ suit) according to their originating environments. The nucleotide sequence databases were searched for the presence of the SP by using tBLASTx (25) with default settings and all of the known SP sequences as queries. The output files were then merged, and the matching translated sequences were collected. All of the redundant sequences (exact or substring match) and those that contained stop codons or were shorter than 45 amino acids were removed. The remaining sequences were submitted to the Conserved Domains Database (27), and the presence of the SP domain was verified in all of them.

Phylogenetic analysis. Amino acid sequences from sequenced members of the PVC superphylum with or without metagenomic proteins were aligned by using MUSCLE (28) in MEGA5 (29), and their evolutionary history was inferred by the unweighted-pair group method using average linkages (UPGMA) (30) or FastTree (31). The evolutionary distances were computed with the JTT (32) for UPGMA and the WAG model (33) for FastTree, while a gamma value of 20 was used for both. Phylogenetic trees were visualized with iTOL (34).

Reverse transcriptase PCR. *Verrucomicrobium spinosum* DSM 4136 and *Rhodopirellula baltica* SH1 were inoculated from colonies grown on agar plates to flasks containing 100 ml of the appropriate media described by Schlesner (35, 36), respectively, and grown while shaking at 22°C. Initially, growth characteristics were determined by measuring optical density at 600 nm (OD_{600}) in a spectrophotometer. Cultures were harvested after 3 days (exponential growth phase) and 5 to 6 days (stationary phase), respectively. Cells were lysed by bead beating (FastPrep FP120, Savant), and total RNA was extracted with TRIzol (Molecular Research Center, Inc.) according to the manufacturer's instructions. Primers were designed to target the genes encoding the *V. spinosum* and *R. baltica* SPs, respectively (VssignF, 5′-TCCCAGCATCGTAGTCTCAA-3′; VssignR, 5′-TAAGCTTC CGGCTTGGTCT-3′; RbsignF, 5′-TAAGAGTCGCAACGTCCTGA-3′; RbsignR, 5′-TTCTTCTTGTCGTCGGCTTC-3′). The housekeeping gene coding for glyceraldehyde 3-phosphate dehydrogenase from *V. spinosum* was used as a positive control (37) (VsqapdhF, 5′-CGGTCTCTTTACCGAAGC TG-3′; VsqapdhR, 5′-CGTTGGAGATGATGTTGTGG-3′). Reverse transcriptase PCR was performed with Moloney murine leukemia virus polymerase (Invitrogen) and an annealing temperature of 55°C for 35 cycles.

Cloning, expression, and purification of recombinant proteins. The genes coding for the SP of *R. baltica* (GenBank/EMBL/DDBJ accession number KF733603) and *Protochlamydia amoebophila* (YP_008052) were synthesized (GenScript Corp.) flanked by restriction sites for EcoRI (Thermo Scientific) and XhoI (Thermo Scientific) that were used for subsequent cloning into the pGEX 4T-1 vector (GE Healthcare) containing an N-terminal glutathione *S*-transferase (GST) tag at the multiple cloning site. The final constructs were then transformed into electrocompetent *E. coli* strain BL21 (λDE3). Transformed *E. coli* cells were grown overnight in 5 ml of Luria-Bertani (LB) medium containing 50 μg/ml ampicillin (LB-Amp) at 37°C on a shaker (120 rpm), and the next day, 1 ml of each culture was used to inoculate flasks containing 100 ml of LB-Amp. The cells were incubated for 2 h (OD_{600} of ~0.4), and then the expression of the proteins was induced by 100 μM (final concentration) isopropyl-β-D-thiogalactopyranoside. After 2 h of induction, cells expressing the GST signature fusion protein were collected by centrifugation in 50-ml tubes at 6,000 rpm for 10 min at 4°C. The supernatant was discarded, and the tubes containing the cell pellets were stored at −20°C.

For protein purification, the collected cell pellets were resuspended by vortexing in 4.5 ml binding buffer (125 mM Tris, 150 mM NaCl, 1% Triton X [pH 8], protease inhibitors [Roche Diagnostics]) plus 0.5 ml lysozyme from a 2-mg/ml stock solution. The tubes were incubated horizontally on a rocking platform for 15 min at room temperature and then placed on ice. The final cell disruption was performed with three rounds of sonication for 30 s at 70% strength (Bandelin Electronic) with intervals of cooling. The lysates were centrifuged at 12,000 rpm for 10 min, and the supernatant was transferred to new tubes. After three rounds of washing in 20 ml of binding buffer, 2 ml of a glutathione-coated magnetic bead slurry (Pierce) was mixed with the lysate and kept shaking horizontally for 1 h. With an appropriate magnetic stand, the beads were washed three times with washing buffer (125 mM Tris, 500 mM NaCl, 1% Triton X [pH 8]). Finally, 4 ml of elution buffer (125 mM Tris, 500 mM NaCl, 50 mM reduced glutathione [pH 9], protease inhibitors) was added and the beads were incubated for an additional 15 min before elution, three times, keeping the eluates separated. The purity and quantity of purified proteins were determined by 12.5% SDS-PAGE and staining with colloidal Coomassie blue (Invitrogen).

For desalting, 2-ml volumes of pooled protein purifications were placed in an Ultracell 10K spin column (Millipore) and phosphate-buffered saline (PBS; pH 7.4) was used to fill the column to 15 ml. The column was centrifuged for 30 min at 5,000 × *g*. The desalting was repeated with another 15 ml of PBS, resulting in 200 μl of desalted and concentrated protein.

Electrophoretic mobility shift assay. To evaluate the effect of SP on the mobility of nucleic acids, purified proteins were mixed with DNA or RNA samples and gel loading dye (New England BioLabs). The mixtures were then loaded onto 1% agarose gels, run for 1 h at 120 V, and visualized by staining with ethidium bromide. When cleaved protein was used, 3 μl of thrombin (GE Healthcare Life Sciences) was added to 30 μl of a desalted and concentrated stock of fusion protein and left overnight at room temperature. Complete cleavage was then verified by SDS-PAGE.

Nucleotide sequence accession number. The gene sequence coding for the SP of *R. baltica* was deposited in GenBank/EMBL/DDBJ under accession number KF733603.

RESULTS AND DISCUSSION

To investigate the evolutionary history of the PVC superphylum, early after the original proposal, we performed a comparative genome analysis to identify orthologous genes conserved among all of the PVC members. We discovered a single protein-coding gene of unknown function that is uniquely shared among all of the members of the superphylum that we refer to as the SP of the PVC superphylum (I. Lagkouvardos, T. Rattei, and M. Horn, 8th German Chlamydia Workshop, Munich, Germany, 24 to 26 February 2010). In the following, we verified its presence in all of the further sequenced PVC genomes published since with PSI-BLAST (25) and found the SP in all of the 55 available genome sequences. The only exceptions were (i) missing gene predictions (e.g., for *R. baltica* SH1) that we identified only with tblastn and (ii) incomplete genome sequences (e.g., the *Poribacteria* draft genome that has been estimated to represent 75% of the complete genome [38]) that we did not consider suitable for presence-or-absence analysis (see Table S1 in the supplemental material). Recently, 16 housekeeping and ribosomal proteins were used to infer evolutionary relationships among the members of the PVC superphylum (5), which further established the common evolutionary origin of the PVC superphylum. By searching for conserved signature insertions

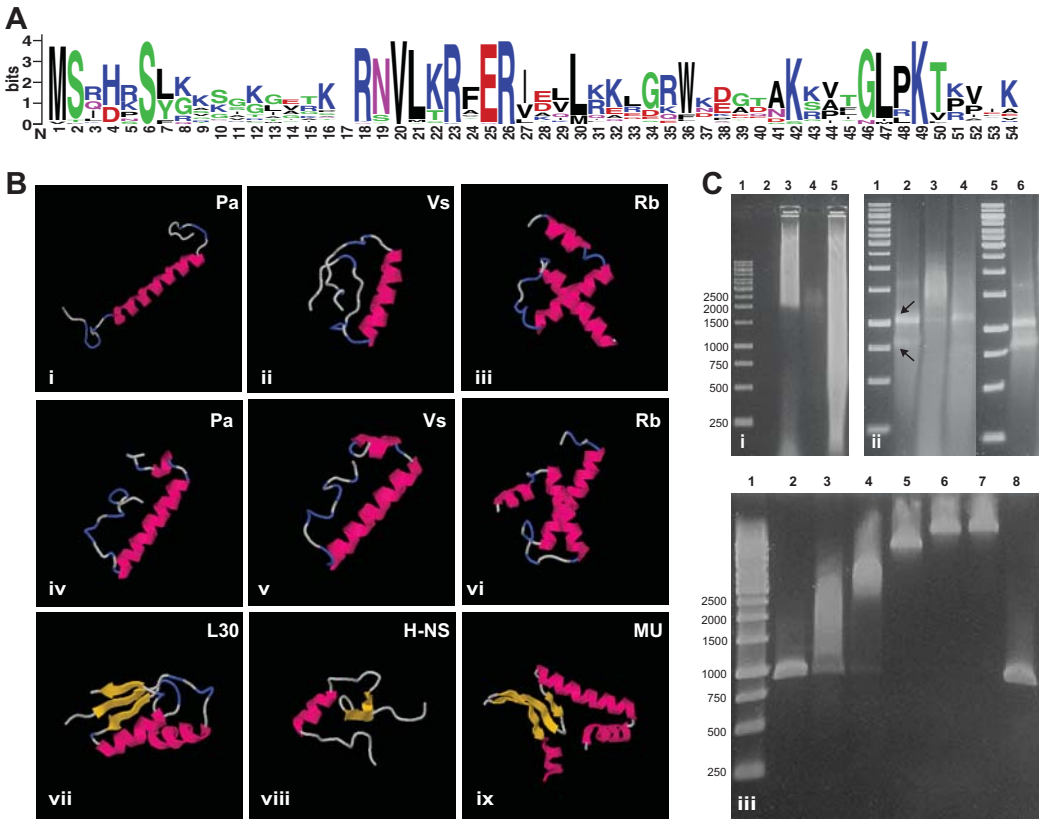

FIG 1 Features of the PVC superphylum SP. (A) Conservation of the SP amino acid sequence. A sequence logo based on a MUSCLE alignment of all of the known SPs generated by WebLogo 3 is shown (28, 48). The overall height of the alignment positions indicates sequence conservation, while the height of each symbol indicates the relative frequency of each amino acid at the respective position. Symbol colors reflect amino acid chemical properties. Highly conserved positions can be observed along the complete length of the alignment, with a longer conserved region in the middle, corresponding to a predicted α-helix. (B) Predicted secondary and tertiary structures of representative SPs compared to those of small DNA/RNA binding proteins of *E. coli*. Predictions were performed with I-TASSER (49) (i to iii) and the QUARK server (50) (iv to vi). Structures: i and iv, SP of *Protochlamydia amoebophila*. UWE25 (GenBank/EMBL/DDBJ accession number YP_008052); ii and v, SP of *V. spinosum* (WP_009960041); iii and vi, SP of *R. baltica* (KF733603); vii, *E. coli* ribosomal protein L30 (Protein Data Bank accession number 2AW4); viii, *E. coli* DNA binding protein H-NS (Protein Data Bank accession number 1HNS); ix, *E. coli* histone like protein HU (Protein Data Bank accession number 1MUL). Pink, alpha-helix; yellow, beta-sheet; blue, turn; gray, unstructured. Independently of the software, a central alpha-helix is predicted for all of the SPs. The SP of all of the *Planctomycetes* shows a C-terminal lysine-rich extension that forms additional secondary-structure elements. (C) Nucleic acid mobility retardation by SP of *R. baltica* and *P. amoebophila*. Agarose gel i, retardation assay with sheared genomic DNA. Lanes: 1, molecular size markers; 2, empty; 3, genomic DNA with GST-tagged SP of *R. baltica*, 4: GST-tagged SP of *R. baltica* without DNA; 5, genomic DNA only. Agarose gel ii, retardation assay with purified total RNA. Lanes: 1 and 5, molecular size markers; 2, RNA only; 3, RNA with GST-tagged SP of *R. baltica*; 4, RNA with GST-tagged SP of *P. amoebophila*; 6, RNA with GST only. Arrows indicate bands representing the 16S and 23S rRNAs, respectively. (Bottom agarose gel) retardation assay with PCR products. Lanes: 1, molecular size markers; 2, only PCR product; 3 to 7, PCR product with increasing concentrations of GST-tagged SP of *P. amoebophila*; 8, PCR product with GST only. The same molecular size marker was used in all of the experiments, and fragment sizes in base pairs are shown on the left. The retardation assays suggest unspecific binding of SP to DNA and RNA.

or deletions, the same study independently recovered the SP to be encoded in all of the known members (except for *Poribacteria*) (5). The presence of a protein in all of the PVC members that does not show any sequence similarity to other known proteins serves as a unifying link among the members of this diverse assemblage of microbes and suggests a conserved function.

Asking whether the SP is expressed, we searched available transcriptomic and proteomic data on members of the PVC superphylum. Members of the phylum *Chlamydiae* are represented the best in such studies, with few reports on *Planctomycetes*. We found evidence of its expression only in members of the phylum *Chlamydiae*, where the SP seems to be expressed constitutively in small amounts similar to those of some housekeeping proteins (see Table S2 in the supplemental material). To compensate for the lack of evidence of transcription in *Planctomycetes* and *Verrucomicrobia*, we performed reverse transcriptase PCR assays with RNA from *R. baltica* SH1 and *V. spinosum* DSM 4136 isolated in the logarithmic and stationary growth phases, respectively. This demonstrated that the SP is also expressed in these organisms (see Fig. S1 in the supplemental material). Taken together, these findings are evidence of the expression of SP by representatives of all of the major phyla within the PVC superphylum.

The SP is a small, 50- to 60-amino-acid protein exhibiting considerable conservation of its sequence (55% average amino acid sequence similarity among all of the representatives; Fig. 1A) and physicochemical properties. *In silico* prediction of its localization, isoelectric point, and secondary structure revealed a highly basic cytosolic protein (pI 10 to 11) (39) consisting of an alpha helix followed by a putative second alpha helix, depending on the prediction software (Fig. 1B), which is reminiscent of the DNA binding helix-turn-helix motif (40). Structure prediction and

physicochemical characteristics thus point toward a nucleic acid-associated protein such as histone-like proteins, transcription factors, or ribosomal proteins. Consistent with this observation, the SP has been recognized as a protein family in TIGRFAM (TIGR04137 [41]), where a possible rRNA interaction is proposed.

To verify the *in silico* prediction and to investigate the *in vitro* activity of the SP, we heterologously expressed the SPs of *P. amoebophila* (as a representative of *Chlamydiae*) and *R. baltica* (*Planctomycetes*) as glutathione *S*-transferase (GST) fusion proteins in *E. coli*. The expressed proteins were purified with glutathione-coated magnetic beads and subsequently used for gel retardation assays. When the fusion proteins were incubated with various DNA and RNA products (sheared genomic DNA, total RNA, or PCR products), the mobility of the nucleic acids in agarose gels was retarded (Fig. 1C). This was also observed after the removal of the GST tag by protease treatment but never when only the GST tag was used (Fig. 1C; see Fig. S2 in the supplemental material). Dose-dependent retardation was observed when increasing amounts of SP were added to PCR products (Fig. 1C). Together, these findings demonstrate an unspecific and concentration-dependent DNA and RNA binding activity of the SP from *R. baltica* and *P. amoebophila in vitro*. This mode of nucleic acid interaction seems to rule out a role for the SP as a transcription factor, which typically shows highly specific DNA binding activity.

To investigate whether the SP could function as a histone-like protein, we analyzed *E. coli* cells overexpressing *R. baltica* or *P. amoebophila* SP. Overexpression of histones generally leads to nucleation of chromatin, which can be detected by staining with DNA-specific dyes (40). However, no nucleation was observed during the overexpression of both SPs in *E. coli* (see Fig. S3 in the supplemental material). Although we cannot exclude a histone function for the SP *in vivo* in *R. baltica* or *P. amoebophila*, expression in the heterologous host does not support such a role. In addition, *P. amoebophila*, showing a condensed nucleoid in the elementary body stage, encodes other histone-like proteins that are likely to be involved in chromatin condensation (42, 43).

The occurrence and documented expression of the SP in all of the members of the PVC superphylum point toward a highly conserved function. This function could be unique to the superphylum, or the SP could substitute for the role of an otherwise conserved and essential protein in non-PVC organisms. To search for proteins that are well conserved in most other organisms but do not occur in PVC members, we conducted a COG-based comparative analysis of all of the available PVC genomes and a representative set of non-PVC genomes. This analysis revealed several highly conserved bacterial functions with no representation by a protein homolog in the PVC superphylum (Table 1). Of those bacterial homologs missing from PVC bacteria, the ribosomal protein L30, which is also present in archaea and eukaryotes, shows a striking physicochemical similarity to the SP of the PVC superphylum. Despite the absence of any amino acid sequence similarity, the two proteins have similar size, pI, and expression profiles (Table 1). Together with its observed nucleic acid binding activity, this suggests the possibility that the SP is a functional analog of ribosomal protein L30, which is missing from all of the members of the superphylum. Further experimental investigation is needed to verify the presence and function of the SP in the ribosome of PVC members.

The high sequence conservation and exclusive presence of the SP in all of the members of the PVC superphylum suggest that it may serve as an additional phylogenetic marker for the superphylum. In fact, the topology of amino acid-based phylogenetic trees resembles that of the 16S rRNA gene (see Fig. S4 in the supplemental material). Simple clustering by UPGMA recovered all of the PVC phyla with good bootstrap support, and the structures within the different phyla are largely similar (see Fig. S4A). The 16S rRNA tree topology was less well recovered in approximately maximum-likelihood SP trees with FastTree (31). Here the *Verrucomicrobia* SPs were not monophyletic but included the *Lentisphaerae* sequences (see Fig. S4B). Still, the overall congruence between 16S rRNA gene- and SP-based trees allowed us to exploit the SP for the analysis of metagenomic data sets from various environmental samples to obtain insights into the diversity of the PVC superphylum. To this end, metagenomic data sets available in IMG/m (44) and SIMAP (45) were first screened with PSI-BLAST (46). In addition, tblastx was used to detect the SP even in the absence of correctly predicted coding sequences. A total of 233 nonredundant SP sequences were detected, mainly in metagenomes originating from freshwater (36%), soil (34%), and marine (21%) samples (see Table S1). Phylogenetic analysis of these sequences showed that the majority of the metagenomic SPs are related to one of the known phyla within the PVC superphylum (Fig. 2). Within the different phyla, however, several novel evolutionary lineages could be observed, significantly expanding the known diversity of the PVC superphylum as inferred from SP phylogeny. *Lentisphaerae* was the least diverse phylum, followed by *Chlamydiae* and "*Candidatus* Omnitrophica (OP3)"; *Planctomycetes* and *Verrucomicrobia* were the most diverse. Interestingly, the majority of the *Verrucomicrobia* sequences originated from soil metagenomes, while most of the "*Candidatus* Omnitrophica (OP3)" and *Chlamydiae* sequences originated from freshwater samples (including sediments); no trend was observed for the other phyla. Although this analysis cannot be used to quantitatively assess the abundance of PVC microbes in the different habitats, the observed ecological patterns are consistent with those of

TABLE 1 Functional categories conserved among bacterial genomes absent from members of the PVC superphylum[a]

COG	No. found in: PVC[b]	No. found in: Other bacteria[c]	Function	Avg length (amino acids)	Avg pI
COG0806	0	466	16S rRNA processing protein RimM	182	4.8
COG0779	0	439	Ribosome maturation factor RimP	154	4.5
COG1559	0	405	Aminodeoxychorismate lyase	339	8.9
COG1841	0	375	Ribosomal protein L30/L7E	60	11.0
COG1660	0	334	Predicted P loop-containing kinase	294	5.8
COG2884	0	327	Cell division ATP-binding protein FtsE	224	9.5
COG2177	0	318	Cell division protein FtsX	304	7.6
COG0595	0	300	mRNA degradation RNase J1/J2	537	5.4
SP	56	0	DNA/RNA binding	60	11.6

[a] COGs absent from all members of the PVC superphylum but conserved in at least 60% of all non-PVC bacteria analyzed are listed together with basic physicochemical properties. The SP of the PVC superphylum is shown for comparison.
[b] Total $n = 56$.
[c] Total $n = 490$.

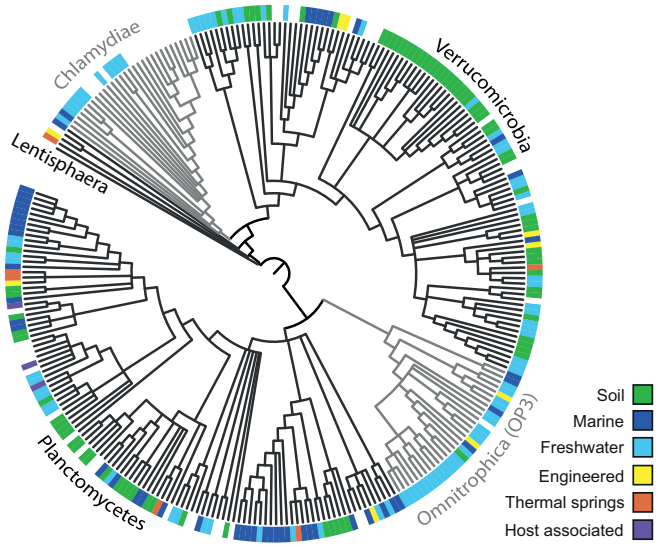

FIG 2 Evolutionary relationships of all of the known PVC superphylum SPs and their metagenomic homologs. The environmental origin of SPs is color coded at the tips of the tree for metagenomic sequences but not for SPs originating from complete genome sequences. An approximate maximum-likelihood tree is shown; nodes with less than 70% support are collapsed.

known members of the superphylum. For example, the relatively low number of metagenomic SPs branching with known members of the *Chlamydiae* phylum is consistent with the generally low abundance of chlamydial protein sequences detected in metagenomes in a recent study (47). An explanation for this could be the low abundance of members of the phylum *Chlamydiae* (which are typically associated with eukaryotic hosts) in environmental samples, which would result in a low coverage of chlamydial genomes in metagenomic data sets. Overall, the suitability of the SP as a phylogenetic marker allows the identification of genomic fragments containing the SP as originating from a PVC member and thus helps in the binning of metagenomic data and in the estimation of the overall presence of PVC members in such data sets. In addition, concatenation of the SP with other conserved proteins should help in the construction of robust phylogenetic trees to analyze the diversity and evolutionary history of the PVC superphylum (5).

In summary, all of the known members of the PVC superphylum produce a small, conserved SP with nucleic acid binding activity. There is evidence of the expression of this protein by some PVC members, and its physicochemical properties, predictions of its structure, and the absence of ribosomal protein L30 from all of the members of the superphylum suggest that the SP has a conserved function and is possibly associated with the ribosome. We demonstrated that the SP is a useful marker for the analysis of metagenomic data and that it may serve to investigate the diversity and ecology of bacteria related to this medically and biotechnologically important superphylum.

ACKNOWLEDGMENTS

We gratefully acknowledge Michael Wagner for helpful discussions.

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).

REFERENCES

1. **Wagner M, Horn M.** 2006. The Planctomycetes, Verrucomicrobia, Chlamydiae and sister phyla comprise a superphylum with biotechnological and medical relevance. Curr. Opin. Biotechnol. **17:**241–249. http://dx.doi.org/10.1016/j.copbio.2006.05.005.
2. **Pilhofer M, Rappl K, Eckl C, Bauer AP, Ludwig W, Schleifer KH, Petroni G.** 2008. Characterization and evolution of cell division and cell wall synthesis genes in the bacterial phyla *Verrucomicrobia*, *Lentisphaerae*, *Chlamydiae*, and *Planctomycetes* and phylogenetic comparison with rRNA genes. J. Bacteriol. **190:**3192–3202. http://dx.doi.org/10.1128/JB.01797-07.
3. **Griffiths E, Gupta RS.** 2007. Phylogeny and shared conserved inserts in proteins provide evidence that Verrucomicrobia are the closest known free-living relatives of chlamydiae. Microbiology **153:**2648–2654. http://dx.doi.org/10.1099/mic.0.2007/009118-0.
4. **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. http://dx.doi.org/10.1093/gbe/evs113.
5. **Gupta RS, Bhandari V, Naushad HS.** 2012. Molecular signatures for the PVC clade (Planctomycetes, Verrucomicrobia, Chlamydiae, and Lentisphaerae) of bacteria provide insights into their evolutionary relationships. Front. Microbiol. **3:**327. http://dx.doi.org/10.3389/fmicb.2012.00327.
6. **Shu QL, Jiao NZ.** 2008. Different Planctomycetes diversity patterns in latitudinal surface seawater of the open sea and in sediment. J. Microbiol. **46:**154–159. http://dx.doi.org/10.1007/s12275-008-0002-9.
7. **Strous M, Pelletier E, Mangenot S, Rattei T, Lehner A, Taylor MW, Horn M, Daims H, Bartol-Mavel D, Wincker P, Barbe V, Fonknechten N, Vallenet D, Segurens B, Schenowitz-Truong C, Medigue C, Collingro A, Snel B, Dutilh BE, Op den Camp HJM, van der Drift C, Cirpus I, van de Pas-Schoonen KT, Harhangi HR, van Niftrik L, Schmid M, Keltjens J, van de Vossenberg J, Kartal B, Meier H, Frishman D, Huynen MA, Mewes HW, Weissenbach J, Jetten MSM, Wagner M, Le Paslier D.** 2006. Deciphering the evolution and metabolism of an anammox bacterium from a community genome. Nature **440:**790–794. http://dx.doi.org/10.1038/nature04647.
8. **Bebear C, de Barbeyrac B.** 2009. Genital Chlamydia trachomatis infections. Clin. Microbiol. Infect. **15:**4–10. http://dx.doi.org/10.1111/j.1469-0691.2008.02647.x.
9. **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. http://dx.doi.org/10.1128/CMR.19.2.283-297.2006.
10. **Horn M.** 2008. Chlamydiae as symbionts in eukaryotes. Annu. Rev. Microbiol. **62:**113–131. http://dx.doi.org/10.1146/annurev.micro.62.081307.162818.
11. **Zwart G, Crump BC, Agterveld MPKV, Hagen F, Han SK.** 2002. Typical freshwater bacteria: an analysis of available 16S rRNA gene sequences from plankton of lakes and rivers. Aquat. Microb. Ecol. **28:**141–155. http://dx.doi.org/10.3354/ame028141.
12. **Bergmann GT, Bates ST, Eilers KG, Lauber CL, Caporaso JG, Walters WA, Knight R, Fierer N.** 2011. The under-recognized dominance of Verrucomicrobia in soil bacterial communities. Soil Biol. Biochem. **43:**1450–1455. http://dx.doi.org/10.1016/j.soilbio.2011.03.012.
13. **Santarella-Mellwig R, Franke J, Jaedicke A, Gorjanacz M, Bauer U, Budd A, Mattaj IW, Devos DP.** 2010. The compartmentalized bacteria of the Planctomycetes-Verrucomicrobia-Chlamydiae superphylum have membrane coat-like proteins. PLoS Biol. **8:**e1000281. http://dx.doi.org/10.1371/journal.pbio.1000281.
14. **Devos DP, Reynaud EG.** 2010. Evolution. Intermediate steps. Science **330:**1187–1188. http://dx.doi.org/10.1126/science.1196720.
15. **Fuerst JA, Sagulenko E.** 2011. Beyond the bacterium: planctomycetes challenge our concepts of microbial structure and function. Nat. Rev. Microbiol. **9:**403–413. http://dx.doi.org/10.1038/nrmicro2578.
16. **Fuchsman CA, Rocap G.** 2006. Whole-genome reciprocal BLAST analysis reveals that *Planctomycetes* do not share an unusually large number of genes with *Eukarya* and *Archaea*. Appl. Environ. Microbiol. **72:**6841–6844. http://dx.doi.org/10.1128/AEM.00429-06.
17. **McInerney JO, Martin WF, Koonin EV, Allen JF, Galperin MY, Lane N, Archibald JM, Embley TM.** 2011. Planctomycetes and eukaryotes: a case of analogy not homology. Bioessays **33:**810–817. http://dx.doi.org/10.1002/bies.201100045.

18. **Budd A, Devos DP.** 2012. Evaluating the evolutionary origins of unexpected character distributions within the bacterial Planctomycetes-Verrucomicrobia-Chlamydiae superphylum. Front. Microbiol. **3:**401. http://dx.doi.org/10.3389/fmicb.2012.00401.
19. **Forterre P.** 2011. A new fusion hypothesis for the origin of Eukarya: better than previous ones, but probably also wrong. Res. Microbiol. **162:**77–91. http://dx.doi.org/10.1016/j.resmic.2010.10.005.
20. **Fuerst JA, Sagulenko E.** 2012. Keys to eukaryality: planctomycetes and ancestral evolution of cellular complexity. Front. Microbiol. **3:**167. http://dx.doi.org/10.3389/fmicb.2012.00167.
21. **Nakamura Y, Cochrane G, Karsch-Mizrachi I.** 2013. The International Nucleotide Sequence Database Collaboration. Nucleic Acids Res. **41:**D21–24. http://dx.doi.org/10.1093/nar/gks1084.
22. **Pruitt KD, Tatusova T, Maglott DR.** 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. **35:**D61–D65. http://dx.doi.org/10.1093/nar/gkl842.
23. **Rattei T, Tischler P, Gotz S, Jehl MA, Hoser J, Arnold R, Conesa A, Mewes HW.** 2010. SIMAP—a comprehensive database of pre-calculated protein sequence similarities, domains, annotations and clusters. Nucleic Acids Res. **38:**D223–226. http://dx.doi.org/10.1093/nar/gkp949.
24. **Powell S, Szklarczyk D, Trachana K, Roth A, Kuhn M, Muller J, Arnold R, Rattei T, Letunic I, Doerks T, Jensen LJ, von Mering C, Bork P.** 2012. eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. Nucleic Acids Res. **40:**D284–289. http://dx.doi.org/10.1093/nar/gkr1060.
25. **Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:**3389–3402. http://dx.doi.org/10.1093/nar/25.17.3389.
26. **Grigoriev IV, Nordberg H, Shabalov I, Aerts A, Cantor M, Goodstein D, Kuo A, Minovitsky S, Nikitin R, Ohm RA, Otillar R, Poliakov A, Ratnere I, Riley R, Smirnova T, Rokhsar D, Dubchak I.** 2012. The genome portal of the Department of Energy Joint Genome Institute. Nucleic Acids Res. **40:**D26–32. http://dx.doi.org/10.1093/nar/gkr947.
27. **Marchler-Bauer A, Zheng CJ, Chitsaz F, Derbyshire MK, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Lanczycki CJ, Lu F, Lu SN, Marchler GH, Song JS, Thanki N, Yamashita RA, Zhang DC, Bryant SH.** 2013. CDD: conserved domains and protein three-dimensional structure. Nucleic Acids Res. **41:**D348–D352. http://dx.doi.org/10.1093/nar/gks1243.
28. **Edgar RC.** 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32:**1792–1797. http://dx.doi.org/10.1093/nar/gkh340.
29. **Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S.** 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. **28:**2731–2739. http://dx.doi.org/10.1093/molbev/msr121.
30. **Sneath PHA, Sokal RR.** 1973. Numerical taxonomy; the principles and practice of numerical classification. W. H. Freeman, San Francisco, CA.
31. **Price MN, Dehal PS, Arkin AP.** 2010. FastTree 2—approximately maximum-likelihood trees for large alignments. PLoS One **5:**e9490. http://dx.doi.org/10.1371/journal.pone.0009490.
32. **Jones DT, Taylor WR, Thornton JM.** 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. **8:**275–282.
33. **Whelan S, Goldman N.** 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. **18:**691–699. http://dx.doi.org/10.1093/oxfordjournals.molbev.a003851.
34. **Letunic I, Bork P.** 2007. Interactive tree of life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics **23:**127–128. http://dx.doi.org/10.1093/bioinformatics/btl529.
35. **Schlesner H.** 1987. *Verrucomicrobium spinosum* gen. nov., sp. nov.—a fimbriated prosthecate bacterium. Syst. Appl. Microbiol. **10:**54–56. http://dx.doi.org/10.1016/S0723-2020(87)80010-3.
36. **Schlesner H.** 1994. The development of media suitable for the microorganisms morphologically resembling *Planctomyces* spp., *Pirellula* spp., and other *Planctomycetales* from various aquatic habitats using dilute media. Syst. Appl. Microbiol. **17:**135–145. http://dx.doi.org/10.1016/S0723-2020(11)80042-1.
37. **Thellin O, Zorzi W, Lakaye B, De Borman B, Coumans B, Hennen G, Grisar T, Igout A, Heinen E.** 1999. Housekeeping genes as internal standards: use and limits. J. Biotechnol. **75:**291–295. http://dx.doi.org/10.1016/S0168-1656(99)00163-7.
38. **Siegl A, Kamke J, Hochmuth T, Piel J, Richter M, Liang CG, Dandekar T, Hentschel U.** 2011. Single-cell genomics reveals the lifestyle of Poribacteria, a candidate phylum symbiotically associated with marine sponges. ISME J. **5:**61–70. http://dx.doi.org/10.1038/ismej.2010.95.
39. **Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, Dao P, Sahinalp SC, Ester M, Foster LJ, Brinkman FSL.** 2010. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. Bioinformatics **26:**1608–1615. http://dx.doi.org/10.1093/bioinformatics/btq249.
40. **Harrison SC.** 1991. A structural taxonomy of DNA-binding domains. Nature **353:**715–719. http://dx.doi.org/10.1038/353715a0.
41. **Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, Finn RD, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Laugraud A, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Mulder N, Natale D, Orengo C, Quinn AF, Selengut JD, Sigrist CJA, Thimma M, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C.** 2009. InterPro: the integrative protein signature database. Nucleic Acids Res. **37:**D211–D215. http://dx.doi.org/10.1093/nar/gkn785.
42. **Horn M, Collingro A, Schmitz-Esser S, Beier CL, Purkhold U, Fartmann B, Brandt P, Nyakatura GJ, Droege M, Frishman D, Rattei T, Mewes HW, Wagner M.** 2004. Illuminating the evolutionary history of chlamydiae. Science **304:**728–730. http://dx.doi.org/10.1126/science.1096330.
43. **Sixt BS, Heinz C, Pichler P, Heinz E, Montanaro J, Op den Camp HJM, Ammerer G, Mechtler K, Wagner M, Horn M.** 2011. Proteomic analysis reveals a virtually complete set of proteins for translation and energy generation in elementary bodies of the amoeba symbiont Protochlamydia amoebophila. Proteomics **11:**1868–1892. http://dx.doi.org/10.1002/pmic.201000510.
44. **Markowitz VM, Ivanova NN, Szeto E, Palaniappan K, Chu K, Dalevi D, Chen IM, Grechkin Y, Dubchak I, Anderson I, Lykidis A, Mavromatis K, Hugenholtz P, Kyrpides NC.** 2008. IMG/M: a data management and analysis system for metagenomes. Nucleic Acids Res. **36:**D534–D538. http://dx.doi.org/10.1093/nar/gkm869.
45. **Rattei T, Tischler P, Arnold R, Hamberger F, Krebs J, Krumsiek J, Wachinger B, Stumpflen V, Mewes W.** 2008. SIMAP—structuring the network of protein similarities. Nucleic Acids Res. **36:**D289–D292. http://dx.doi.org/10.1093/nar/gkm963.
46. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ.** 1990. Basic local alignment search tool. J. Mol. Biol. **215:**403–410. http://dx.doi.org/10.1016/S0022-2836(05)80360-2.
47. **Lagkouvardos I, Weinmaier T, Lauro FM, Cavicchioli R, Rattei T, Horn M.** 15 August 2013. Integrating metagenomic and amplicon databases to resolve the phylogenetic and ecological diversity of the chlamydiae. ISME J. (Epub ahead of print.) http://dx.doi.org/10.1038/ismej.2013.142.
48. **Crooks GE, Hon G, Chandonia JM, Brenner SE.** 2004. WebLogo: A sequence logo generator. Genome Res. **14:**1188–1190. http://dx.doi.org/10.1101/gr.849004.
49. **Zhang Y.** 2008. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics **9:**40. http://dx.doi.org/10.1186/1471-2105-9-40.
50. **Xu D, Zhang Y.** 2012. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins **80:**1715–1735. http://dx.doi.org/10.1002/prot.24065.

The signature protein of the PVC superphylum

Lagkouvardos *et al.*

Supplementary Material

Table S1: Signature proteins identified in genome sequences of members of the PVC superphylum and in metagenomic data. The SP was identified in the genome sequences of all PVC members; only one representative SP is shown if sequences were identical among the same species.
See file "Table S1.xlsx".

Table S2: Detection of the PVC signature protein in transcriptomic and proteomic studies. Due to the missing gene prediction in the genome of *Rhodopirellula baltica* SH1 the respective SP was never detected.

Organism	Method	Detection	Reference
Chlamydia trachomatis	microarray	–	Belland, 2003a (1)
Chlamydia trachomatis	microarray	+	Belland, 2003b (2)
Chlamydia trachomatis	RNA-Seq	+	Albrecht, 2010 (3)
Chlamydia trachomatis	proteomics	–	Shaw, 2002 (4)
Chlamydia trachomatis	proteomics	+	Skipp, 2005 (5)
Chlamydia pneumoniae	proteomics	–	Vandahl, 2001(6)
Chlamydia pneumoniae	proteomics	–	Molestina, 2002 (7)
Chlamydia pneumoniae	proteomics	–	Wehr, 2004 (8)
Chlamydia pneumoniae	proteomics	–	Mukhopadhyay, 2006 (9)
Chlamydia pneumoniae	microarray	+	Maurer, 2007 (10)
Chlamydia pneumoniae	RNA-Seq	+	Albrecht, 2011 (11)
Chlamydia pneumoniae	proteomics	–	Saka, 2011 (12)
Protochlamydia amoebophila	proteomics	–	Heinz, 2010 (13)
Protochlamydia amoebophila	proteomics	–	Sixt, 2011 (14)
Protochlamydia amoebophila	microarray	+	Haider, unpublished
Rhodopirellula baltica	proteomics	–	Gade, 2005a (15)
Rhodopirellula baltica	proteomics	–	Gade, 2005b (16)
Rhodopirellula baltica	proteomics	–	Hieu, 2008 (17)
Rhodopirellula baltica	RNA-Seq	–	Wecker, 2009 (18)
Rhodopirellula baltica	RNA-Seq	–	Wecker, 2010 (19)

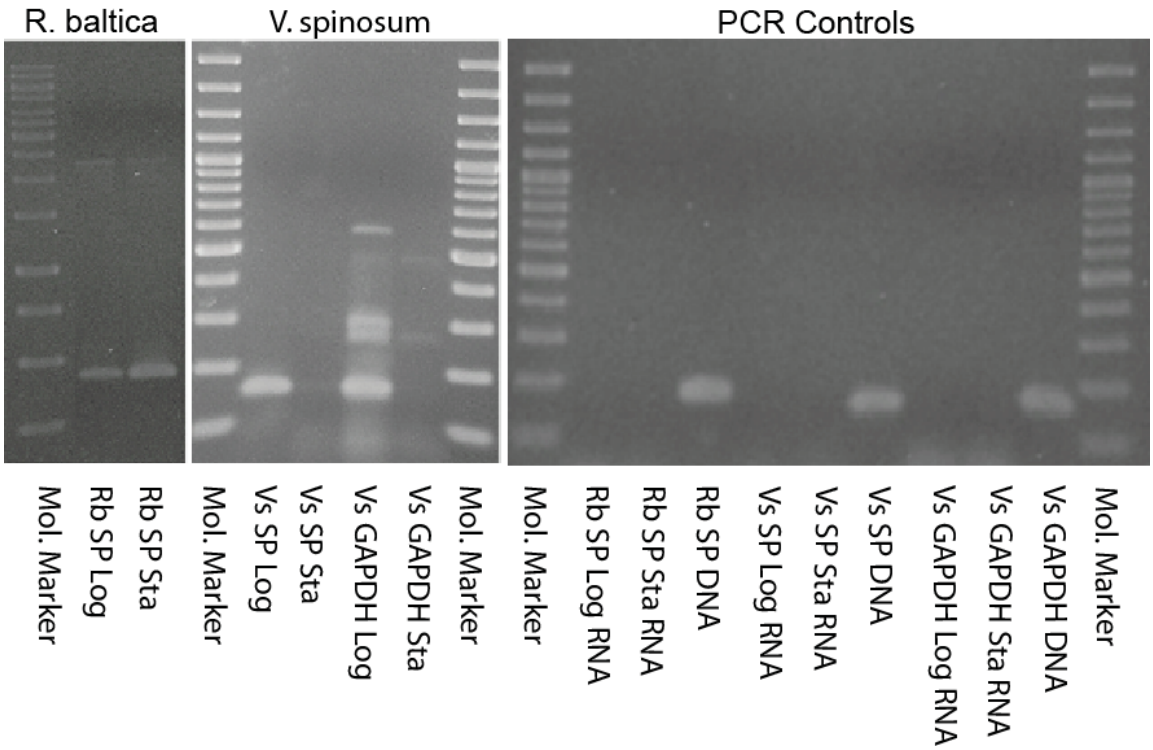

Figure S1: The SP of *Rhodopirellula baltica* (Rb) and *Verrucomicrobium spinosum* (Vs) is transcribed. Reverse transcriptase PCR using RNA isolated during logarithmic (3 days) and stationary growth (6 days). The GAPDH gene of *Verrucomicrobium spinosum* was used as positive control (left panel). A PCR using the same RNA samples demonstrates the absence of DNA in the RNA preparations (right panel). Note that the amount of RNA obtained from *V. spinosum* in the stationary phase was too low for successful detection of SP and the control. All RT-PCR products were cloned and verified by sequencing.

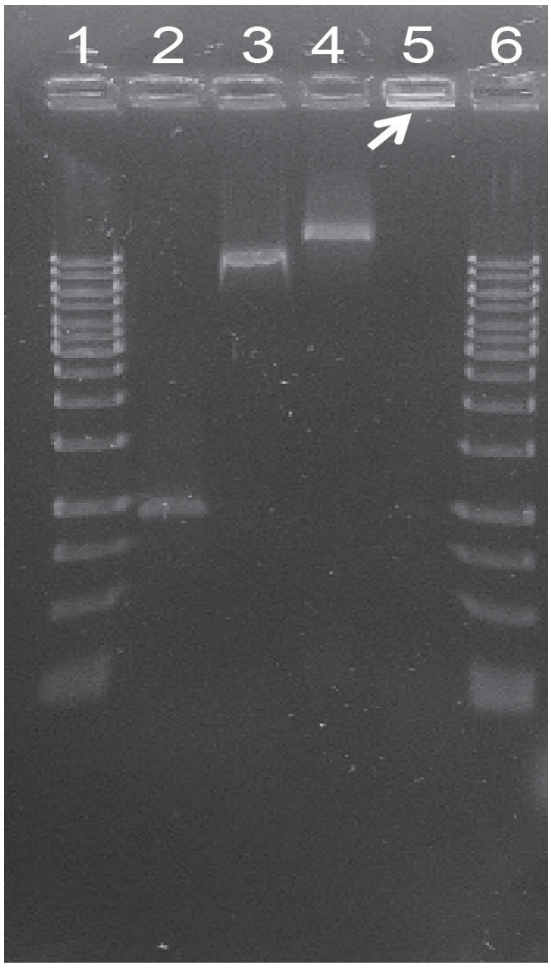

Figure S2: DNA mobility retardation by GST tagged and untagged SP of *R. baltica*. 1,6, molecular marker; 2, DNA (PCR product) only; 3-4 DNA with different doses of GST tagged SP of *R. baltica*; 5, DNA with the same amount of protein as in 4 but thrombin digested. The DNA incubated with the digested mixture of GST and Rb SP didn't enter the gel (arrow). The complete digestion of the GST tag was verified by SDS-PAGE.

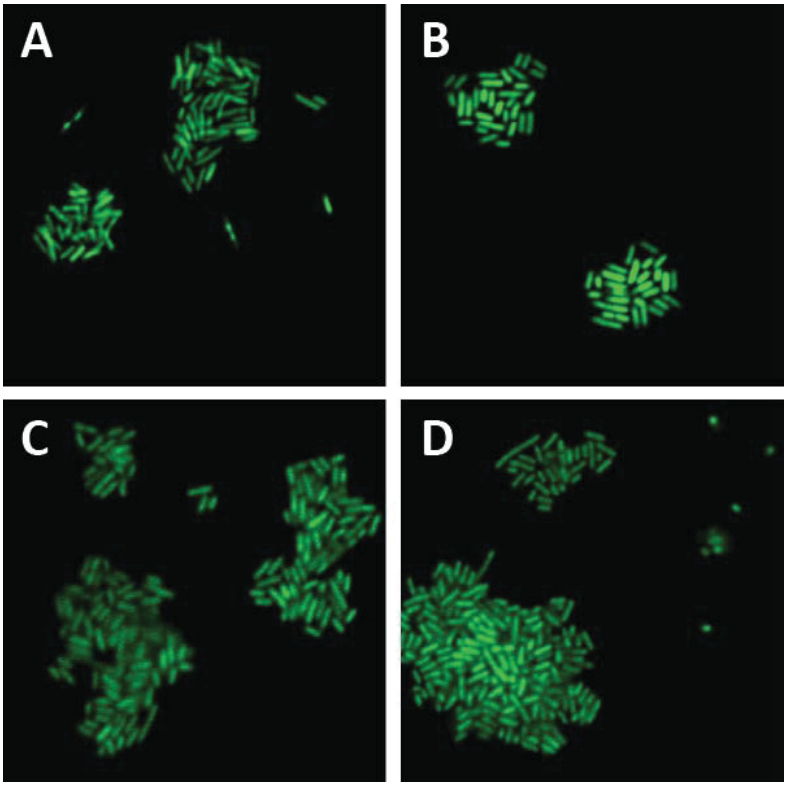

Figure S3: Effect of heterologous overexpression of signature proteins in *E. coli* BL21. A, GST-tagged SP of *P. amoebophila*; B, GST-tagged SP of *R. baltica*; C, GST only; D, non-induced *E.coli* cells. *E. coli* were induced for 2 hours using 1mM IPTG. Expression of the SP was verified by SDS-PAGE. No nucleation was observed after staining with SYBR Green.

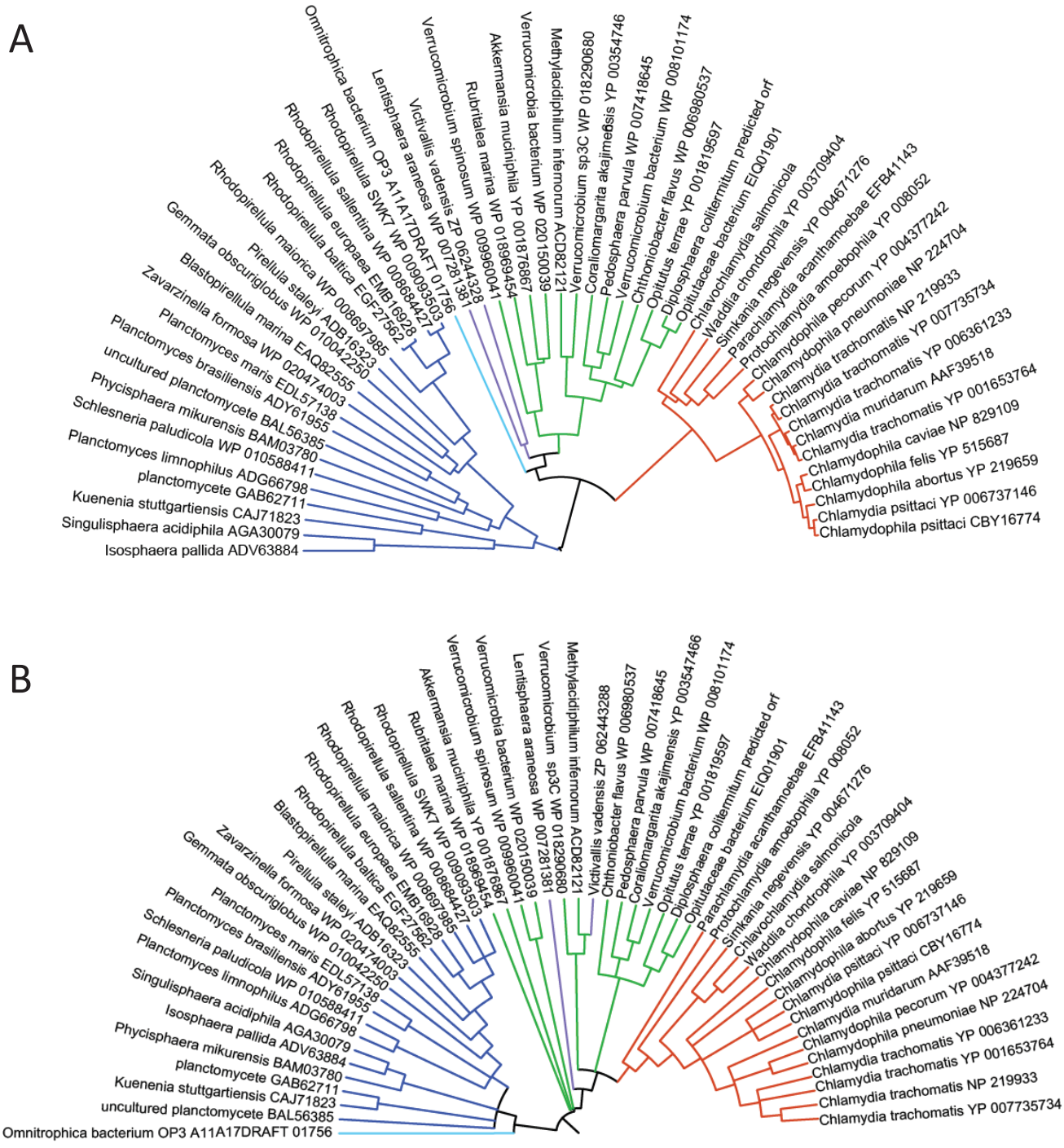

Figure S4. Evolutionary relationships of PVC superphylum signature proteins from published genome sequences. All SP sequences were aligned using MUSCLE (20) in MEGA5 (21) and their evolutionary history was inferred using (A) UPGMA or (B) FastTree (22). The evolutionary distances were computed using the JTT(23) for UPGMA and WAG model (23) for FastTree, while a gamma value of 20 was used for both. Nodes with less than 70% bootstrap support are collapsed in the maximum likelihood tree.

References

1. **Belland RJ, Zhong GM, Crane DD, Hogan D, Sturdevant D, Sharma J, Beatty WL, Caldwell HD.** 2003. Genomic transcriptional profiling of the developmental cycle of Chlamydia trachomatis. P Natl Acad Sci USA **100:**8478-8483.
2. **Belland RJ, Nelson DE, Virok D, Crane DD, Hogan D, Sturdevant D, Beatty WL, Caldwell HD.** 2003. Transcriptome analysis of chlamydial growth during IFN-gamma-mediated persistence and reactivation. P Natl Acad Sci USA **100:**15971-15976.
3. **Albrecht M, Sharma CM, Reinhardt R, Vogel J, Rudel T.** 2010. Deep sequencing-based discovery of the Chlamydia trachomatis transcriptome. Nucleic Acids Res **38:**868-877.
4. **Shaw AC, Gevaert K, Demol H, Hoorelbeke B, Vandekerckhove J, Larsen MR, Roepstorff P, Holm A, Christiansen G, Birkelund S.** 2002. Comparative proteome analysis of Chlamydia trachomatis serovar A, D and L2. Proteomics **2:**164-186.
5. **Skipp P, Robinson J, O'Connor CD, Clarke IN.** 2005. Shotgun proteomic analysis of Chlamydia trachomatis. Proteomics **5:**1558-1573.
6. **Vandahl BB, Birkelund S, Demol H, Hoorelbeke B, Christiansen G, Vandekerckhove J, Gevaert K.** 2001. Proteome analysis of the Chlamydia pneumoniae elementary body. Electrophoresis **22:**1204-1223.
7. **Molestina RE, Klein JB, Miller RD, Pierce WH, Ramirez JA, Summersgill JT.** 2002. Proteomic analysis of differentially expressed Chlamydia pneumoniae genes during persistent infection of HEp-2 cells. Infect Immun **70:**2976-2981.
8. **Wehr W, Meyer TF, Jungblut PR, Muller EC, Szczepek AJ.** 2004. Action and reaction: Chlamydophila pneumoniae proteome alteration in a persistent infection induced by iron deficiency. Proteomics **4:**2969-2981.
9. **Mukhopadhyay S, Good D, Miller RD, Graham JE, Mathews SA, Timms P, Summersgill JT.** 2006. Identification of Chlamydia pneumoniae proteins in the transition from reticulate to elementary body formation. Mol Cell Proteomics **5:**2311-2318.
10. **Maurer AP, Mehlitz A, Mollenkopf HJ, Meyer TF.** 2007. Gene expression profiles of Chlamydophila pneumoniae during the developmental cycle and iron depletion-mediated persistence. Plos Pathog **3:**752-769.
11. **Albrecht M, Sharma CM, Dittrich MT, Muller T, Reinhardt R, Vogel J, Rudel T.** 2011. The transcriptional landscape of Chlamydia pneumoniae. Genome Biol **12**.
12. **Saka HA, Thompson JW, Chen YS, Kumar Y, Dubois LG, Moseley MA, Valdivia RH.** 2011. Quantitative proteomics reveals metabolic and pathogenic properties of Chlamydia trachomatis developmental forms. Mol Microbiol **82:**1185-1203.
13. **Heinz E, Pichler P, Heinz C, den Camp HJMO, Toenshoff ER, Ammerer G, Mechtler K, Wagner M, Horn M.** 2010. Proteomic analysis of the outer membrane of Protochlamydia amoebophila elementary bodies. Proteomics **10:**4363-4376.
14. **Sixt BS, Heinz C, Pichler P, Heinz E, Montanaro J, den Camp HJMO, Ammerer G, Mechtler K, Wagner M, Horn M.** 2011. Proteomic analysis reveals a virtually complete set of proteins for translation and energy generation in elementary bodies of the amoeba symbiont Protochlamydia amoebophila. Proteomics **11:**1868-1892.
15. **Gade D, Stuhrmann T, Reinhardt R, Rabus R.** 2005. Growth phase dependent regulation of protein composition in Rhodopirellula baltica. Environ Microbiol **7:**1074-1084.

16. **Gade D, Theiss D, Lange D, Mirgorodskaya E, Lombardot T, Glockner FO, Kube M, Reinhardt R, Amann R, Lehrach H, Rabus R, Gobom J.** 2005. Towards the proteome of the marine bacterium Rhodopirellula baltica: Mapping the soluble proteins. Proteomics **5:**3654-3671.
17. **Hieu CX, Voigt B, Albrecht D, Becher D, Lombardot T, Glockner FO, Amann R, Hecker M, Schweder T.** 2008. Detailed proteome analysis of growing cells of the planctomycete Rhodopirellula baltical SH1(T). Proteomics **8:**1608-1623.
18. **Wecker P, Klockow C, Ellrott A, Quast C, Langhammer P, Harder J, Glockner FO.** 2009. Transcriptional response of the model planctomycete Rhodopirellula baltica SH1(T) to changing environmental conditions. BMC Genomics **10:**410.
19. **Wecker P, Klockow C, Schuler M, Dabin J, Michel G, Glockner FO.** 2010. Life cycle analysis of the model organism Rhodopirellula baltica SH 1(T) by transcriptome studies. Microb Biotechnol **3:**583-594.
20. **Edgar RC.** 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res **32:**1792-1797.
21. **Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S.** 2011. MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Mol Biol Evol **28:**2731-2739.
22. **Price MN, Dehal PS, Arkin AP.** 2010. FastTree 2-Approximately Maximum-Likelihood Trees for Large Alignments. Plos One **5**.
23. **Whelan S, Goldman N.** 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol **18:**691-699.

Table S1: Signature proteins iden fied in genome sequences of members of the PVC superphylum and in metagenomic data.
The SP was iden fied in the genome sequences of all PVC members; only one representa ve SP is shown if sequences were iden cal among the same species.

		Source	Source Type	Identifier	Identifier type	Sequence
Chlamydiae	1	Chlamydia muridarum Nigg	Genomic	AAF39518	NCBI accession	MSRHRSYGKSIKGETKRNVLKRFERIEVLRKLGRWDDATAKKATGLLKTPVIK
	2	Chlamydia psittaci M56	Genomic	YP_006737146	NCBI accession	MSRHRSYGKSIKGETKRNVLKRFERIDLLRKLGRWNDATAKKATGLPKTPAIK
	3	Chlamydia trachomatis D/UW-3/CX	Genomic	NP_219933	NCBI accession	MSRHRSYGKSVKGETKRNVLKRFERIEVLRKLGRWDDATAKKATGLLKTPAIK
	4	Chlamydia trachomatis E/SW3	Genomic	YP_006361233	NCBI accession	MSRHRSYGKSVKGETKRNVLKRFERIEVLRKLGRWDDATAKKATGLLKTPVIK
	5	Chlamydia trachomatis Ia/SotoniaI	Genomic	YP_007735734	NCBI accession	MSRHRSYGKSVKGETKRNVLKRFERIKVLRKLGRWDDATAKKATGLLKTPVIK
	6	Chlamydia trachomatis L2b/UCH-1/proctitis	Genomic	YP_001653764	NCBI accession	MSRHRSYGKSFKGETKRNVLKRFERIEVLRKLGRWDDATAKKATGLLKTPAIK
	7	Chlamydophila abortus S26/3	Genomic	YP_219659	NCBI accession	MSRHRSYGKSIKGETKRNVLKRFERIDLLRKLSRWNDTTAKKATGLPKTPVIK
	8	Chlamydophila caviae GPIC	Genomic	NP_829109	NCBI accession	MSRHRSYGKSIKGETKRNVLKRFERIDVLRKLGRWNDVTAKKATGLPKTPVMK
	9	Chlamydophila felis Fe/C-56	Genomic	YP_515687	NCBI accession	MSRHRSYGKSIKGETKRNVLKRFERIEVLRKLGRWNDATAKKATGLPKTPVMK
	10	Chlamydophila pecorum E58	Genomic	YP_004377242	NCBI accession	MSRHRSYGKSVKGVTKRNVLKRFERIEVLRKLGRWDDATAKKATGLPKTPILK
	11	Chlamydophila pneumoniae CWL029	Genomic	NP_224704	NCBI accession	MSRHRSYGKSVKGVTKRNVLKRFERVEVLRKLGRWNDSTAKKVTGLPKTPILK
	12	Chlamydophila psittaci RD1	Genomic	CBY16774	NCBI accession	MSRHRSYGKSIKGETKRNVLKRFERIDLLRKLGRWNDATAKKATGLPKTPVIK
	13	Chlavochlamydia salmonicola	Genomic	Unpublished	local	MSRHPSYGKSTKTAKKRNVLKRFERIDVLKKLGRWSKEKNSRVTGLPKTDFIE
	14	Parachlamydia acanthamoebae Hall's coccus	Genomic	EFB41143	NCBI accession	MSRHPSFGSSSKAATRNVLKRFERIDVLRKLGRWKDGENKRVTGLPKTLIQQ
	15	Protochlamydia amoebophila UWE25	Genomic	YP_008052	NCBI accession	MSRHPSFGKAGKTATKRNVLKRFERIDVLKKIGRWKDSENKRITGLPKTPVL
	16	Simkania negevensis Z	Genomic	YP_004671276	NCBI accession	MSRHPSYGKSNKGGKKQNVLKRFQRIDILRKLGRWKDGENKKVTGLPKTPVK
	17	Waddlia chondrophila WSU 86-1044	Genomic	YP_003709404	NCBI accession	MSRHPSFGKSNKGGQKRNVLKRFERLDVLKKLGRWEKDKDKKITGLPKTPVM
Lentisphaerae	18	Lentisphaera araneosa HTCC2155	Genomic	WP_007281381	NCBI accession	MSIHRSLKVKGNTAGKKNVLKRFERVDQLIQEGRLKPGDQVLGLPKTKVNI
	19	Victivallis vadensis ATCC BAA-548	Genomic	ZP_06244328	NCBI accession	MSQHPSLRVGGKIRKIRNVMKRYERIDVLLADGRIKEDKVLGLPKTKPTEI
Omnitrophica	20	Omnitrophica bacterium sp.	Genomic	A11A17DRAFT_01756	locus_tag	MSQHSSLKARGFGVKHRNVLKRFERIKKLKADGRWQDVMSPFGLPKIKSVKIKVKKIKAEAKTDAAAPAGAPAAAAPAKTEAKP
Verrucomicrobia	21	Akkermansia muciniphila ATCC BAA-835	Genomic	YP_001876867	NCBI accession	MSKHSSLKATGTVGGKRSVLKRFERVKLLKERGEWKKGQSPLGLPKTKHEA
	22	Chthoniobacter flavus Ellin428	Genomic	WP_006980537	NCBI accession	MSQHRSLKGQSTIAAKRNVLKRFERVELLQQRGQYKQGDKVVGLRKTKPNE
	23	Coraliomargarita akajimensis DSM 45221	Genomic	YP_003547466	NCBI accession	MSQHNSFKASGGGGKKNRTVLKRFERVELLRKRGEWKEGDRVIGLKKTQPEA
	24	Diplosphaera colitermitum TAV2	Genomic	Predicted_orf	local	MSQHNSLQGPRGIVIKRNVLKRFERVALLKKRSQWKEGDRVMGLRKTKPDV
	25	Methylacidiphilum infernorum V4	Genomic	ACD82121	NCBI accession	MSQHRSYRTGSLLVAKRNVLKRYERINILKKQGKWKEGDKVLGLPKTKPI
	26	Opitutaceae bacterium TAV1	Genomic	EIQ01901	NCBI accession	MSQHNSLQGPRGIVIKRNVLKRFERVALLKKRGQWKDGERVMGLRKTKPDV
	27	Opitutus terrae PB90-1	Genomic	YP_001819597	NCBI accession	MSQHKSLQGSSGIVIKRNVLKRFERVDILKKRGQWKAGDRVQGLRKTKPDV
	28	Pedosphaera parvula str. Ellin514	Genomic	WP_007418645	NCBI accession	MSQHRSLRAVATMGGKRNVLKRFERVGLLKKRGQWKEGDRITGLRKTKPEA
	29	Rubritalea marina DSM 17716	Genomic	WP_018969454	NCBI accession	MSKHSSLKAKGSAQGKRSVMKRFERVKALQESGKWQEGQSPIGLPKTKVTK
	30	Verrucomicrobia bacterium SCGC AAA164-M04	Genomic	WP_020150039	NCBI accession	MSKHNSLKRKGGATGKRSVMKRFERVKLMKERGDWKEGQSAIGLPKTKAEG
	31	Verrucomicrobiae bacterium DG1235	Genomic	WP_008101174	NCBI accession	MSQHPSYKSNASTGAKRNVLKRFERVELLKKRGEWKDGDRVSGLRKTKPEA
	32	Verrucomicrobium sp. 3C	Genomic	WP_018290680	NCBI accession	MSQHRSLRSGSLLAAKRNVLKRLERIALLKKRGRWKDGMRVFGLPKTRPV
	33	Verrucomicrobium spinosum DSM 4136	Genomic	WP_009960041	NCBI accession	MSQHRSLKSSGGSVGTKRSVLKRGERIKLMKARGQWNEGRSLYNLPKTKPEA
Planctomycetes	34	Blastopirellula marina DSM 3645	Genomic	EAQ82555	NCBI accession	MTIDKSLKVKRGGISSRSVLTRVERLEQMRKAGKFNPETDSPIGIPKTRVIKISMKKKKKKKEEAT
	35	Gemmata obscuriglobus UQM 2246	Genomic	WP_010042250	NCBI accession	MSIDKSLKRKAGMSRQRCVLTRAERITKMLENGKFGAESSPYGLPKTRVQKIALKKKAKKEEAEGGDDKAKKKK
	36	Isosphaera pallida ATCC 43644	Genomic	ADV63884	NCBI accession	MSIDKSLRKGGGLGGQRNVLKRAERMALLQEDGRWTENHGPYNLPKTKYRKLKPGQSGPKRPESAS
	37	Kuenenia stuttgartiensis	Genomic	CAJ71823	NCBI accession	MSIDKSLKPKGKLSRPRNVYRKVERIAILKAENRWDQDTSVFGLPKVKVEKLKRKGKAPKKVEPEETGKAGKAGKAGKAKK
	38	Phycisphaera mikurensis NBRC 102666	Genomic	BAM03780	NCBI accession	MSLDSSLKAGGGLKRHRNVLTRPERIELLAKKDAFDKKQGDPLGLPKVGNRKLTTGKK
	39	Pirellula staleyi DSM 6068	Genomic	ADB16323	NCBI accession	MPIDNSLKVKAGAISNRNVLTRAERLEKLKEADRWKEGDKVLGMPKVRVVKISMKKKKKVKTEEEGGAAGAKGAKGGKAAPKK
	40	Planctomyces brasiliensis DSM 5305	Genomic	ADY61955	NCBI accession	MSLNSSLKTGSKMKRPRNVLKRHERIEQLQKQDRWIEGQAPIGLPKVRVFKAVIGKKKKKTKEEK
	41	Planctomyces limnophilus DSM 3776	Genomic	ADG66798	NCBI accession	VSIDKSLKRKGRLARSRNVLSRSERILQMIGEDKLGTQDSPFGLPKLRIVKMVIGKKKKKKAAEDKKDDKKKAAKKK
	42	Planctomyces maris DSM 8797	Genomic	EDL57138	NCBI accession	VSLDKSLKSKSTLVRARNVLKRAERIEKLKFEDRWVEGQGALGLPKVRVEKISIGKKKKKKKDDDED
	43	planctomycete KSU-1	Genomic	GAB62711	NCBI accession	MSIDKSLKTKGKLVRPRNVLTRIERIKLLREEKKWEPTISVFGIPKVKVLKLKQRGKAEKKAKPEATAAAAGGEVSATAGKEKAKK
	44	Rhodopirellula baltica WH47	Genomic	EGF27562	NCBI accession	MTMDRSLKVQAGAIKSRNVLTRAERVERMKDLDKFNEESSIIGMPKTRVLKVSLKKKKKVKKADEADDKKKK
	45	Rhodopirellula europaea 6C	Genomic	EMB16928	NCBI accession	MTMDRSLKVQAGAIKSRNVLTRAERVERMKDLDKFNEESSIIGMPKTRVLKVSLKKKKKVKKADDADDKKKK
	46	Rhodopirellula maiorica	Genomic	WP_008697985	NCBI accession	MTMDRSLKVQAGAIKSRNVLTRAERIARLTELDRFDPEASIVGMAKVRVPKVSLKRKKKVKKEDDPKDAKKKK
	47	Rhodopirellula sallentina	Genomic	WP_008684427	NCBI accession	MTMDRSLKVQAGAIKSRNVLTRAERIARLKELDKFDEDASIIGMPKTRVQKISLKKKKKVKKAED
	48	Rhodopirellula sp. SWK7	Genomic	WP_009093503	NCBI accession	MTMDRSLKVQAGAIKSRNVLTRAERIARLKELDKFDENASIIGMPKTRVQKISLKKKKKVKKAEEK
	49	Schlesneria paludicola DSM 18645	Genomic	WP_010588411	NCBI accession	MTIEKSLKRKGRLARLRSVLSRDERITQMKADERWADGTSPFGLPKLRVVRLVVGKKKKKKTAEGDKKDDKKKAAKKK
	50	Singulisphaera acidiphila DSM 18658	Genomic	AGA30079	NCBI accession	MSIDKSLKKGGGLSRTRNVLTRPERLALLQEDERWKPAQGVFNLPKTKFRRLAPGQSGPKRPS
	51	Uncultured planctomycete	Genomic	BAL56385	NCBI accession	MSLDKSLRRKGRIVRARNVLTRAERIKLMQEERWREGMSPFGLPKIRVIKLLKKSKRSAKEEKAAAAAGGAAGAAASPTQEKK
	52	Zavarzinella formosa	Genomic	WP_020474003	NCBI accession	MSIDKSLKRKSGGISNRSVLTRTERINTLKDADKWVEGRSPFGLPKVRVLKISLKKAKKEKEEEDTKAPAKKK
	53	Alaska permafrost	Soil	2124908040 Predicted orf	local	MSLHSSLRXDKAGAAQRTVLTRIERIKELMKKGQWSDESSVTNLPKTKQVRV
	54	Alaska permafrost	Soil	2124908041 Predicted orf	local	MSQHRSLRAASTLGGKRNVLKRFERVELLKKRGQWKAGDRITGLRKTKPEA
	55	Alaska permafrost	Soil	2124908043 Predicted orf	local	MSQHRSLKGASTITAKRNVLKRFERVELLKKRGQWKEGGKVVGLPKTKPDV
	56	Alaska permafrost	Soil	2124908044 Predicted orf	local	MSQHRSLKGASTITAKRNVLKRFERVELLKKRGQWKDGAKVMGLPKTKPDA
	57	Alaska permafrost	Soil	2140918006 Predicted orf	local	MTQHSSLKSANKITAKRNVMKRFERIDFLRASGKWKEGDRGLGLPKTKPQV
	58	Annamox bioreactor	Engineered	2017108002 Predicted orf	local	MSIDKSLKTMGKLVRQRNVLTRPERIKYLMDEGLWDEKKSVYGLPKV
	59	Ant colony	Soil	2032320007 Predicted orf	local	MSLDRSLKTAGSLIKHRNVLTRPERVAKLAAAGKFDLSQDDPVGIPKVLNRRVAAGGK
	60	Ant colony	Soil	2038011000 Predicted orf	local	MTIDKSLKVKGSGIQNRNVLTRAERIEKLVAADKFKEGDKVLGLPKSAF
	61	Ant colony	Soil	2038011000 Predicted orf	local	MSMDPSLKSASSLVRHRNVLTRGERLERLAGQGKWDESKAVFGLPKVGNRKLVVGGKPKKEE
	62	Ant colony	Soil	2038011000 Predicted orf	local	MSIDPSLKTAGNLTTKRSVMKREERIAALKDEKRHDDSKGALNLPKTKVRE
	63	Ant colony	Soil	2040502000 Predicted orf	local	MSLDKSLKKAGSMARARNVLTRAERMAVLQEDDRWRPEQGVYNLPKTKYRRLAPGQSGPKRPD
	64	Ant colony	Soil	2040502000 Predicted orf	local	MTIDKSLKVRRGATSNRSVLTRAERLERLKETDRWTDGMSPLGLPKVRVRKPAM
	65	Ant colony	Soil	2040502000 Predicted orf	local	MTMDQSLKAKKGLARARGVLSRDERLAKLREMDRWSEGDAPVGLPK
	66	Ant colony	Soil	2040502000 Predicted orf	local	MSQHPSFKAGGSGNQKKRNVLKRFERVNLLRKRGQWKDGNRVTGLLKTLPEE
	67	Ant colony	Soil	2040502000 Predicted orf	local	MSQHSSLQGAKGIVIKRNVLKRFERVATLKKRGQWKDGDRIMGLRKTKPDV
	68	Ant colony	Soil	2040502000 Predicted orf	local	MSQHRSLRAASNTGGKRNVMKRFERVAVLKKRGQWKDGDRVTGLRKTKPEV
	69	Antarctic aquatic	Marine	orf_1032546/47558590	SIMAP name/ID	MSRHRSYGKATTGAKKRNVLKRFERVQLMKDRDQWKERISVFSLPKQSCIIEKTLF
	70	Antarctic aquatic	Marine	orf_1032914/47562567	SIMAP name/ID	MSQHKSLQGPKGLVVKRNVLKRFERVELLRKRGQWKEGDRVQGLRKTTPDV
	71	Antarctic aquatic	Marine	orf_84015/42269853	SIMAP name/ID	MGLHPSLKRAKNLSGSRSVMKRSERLKWLKDKGDWSEKDPVLGLPKIKIVKLKALKKEKAKEKEEGATEEKKETSAQTQKPKD
	72	Antarctic Lake	Freshwater	2100351015 Predicted orf	local	MTLDKSLRVRRGLIRSRSVLSRAERLVKLQESERWKEGDSPLGLPKVRVAKLA
	73	Arabidopsis rhizosphere	Soil	2105370224	IMG/m id	MSQHRSLRAASTAGAKRNVMKRYERVELLKKRGQWKTGDRVTGLRKTKPEA
	74	Arabidopsis rhizosphere	Soil	2105812886	IMG/m id	MSQHRSLKGASTIAAKRNVLKRFERVELLKKRGQWKDGQKVVGLPKTKPDV
	75	Arabidopsis rhizosphere	Soil	2209111006 Predicted orf	local	MSQHRSLKGASTIAARRNVLKRFERVELLKKRGQWKEGQKVIGLPKTKPDV
	76	Arabidopsis rhizosphere	Soil	2209111006 Predicted orf	local	MSQHRSLRGASTIAAKRNVLKRFERVDLLKKRGQWKETSKVIGLPKTKPDV
	77	Arctic Sediment	Marine	2088090012 Predicted orf	local	MSIDRSLKIKGALSRHRNVLTRAERIEKLKEEERWSEGESLLGLPKVAHRKSHAGRK
	78	Arctic Sediment	Marine	2088090012 Predicted orf	local	MTIDKSLKIRRGGSGNRSVLTRIERLERLKEADRWKEGDSPLGLPKVRVRKLAM
	79	Arctic Sediment	Marine	2088090012 Predicted orf	local	MTIDRTLKKHGGLKGQRSVLTRAERIAKLTEDGKFDPEKDSVLGLPKVKVKA
	80	Arctic Sediment	Marine	2100351001 Predicted orf	local	MSMHSSLKGAAKIRTRRNVLKRFERVEVLKRRANGRMATVRMACRKPSPKILE
	81	Arctic Sediment	Marine	2100351001 Predicted orf	local	MSIDPSLKIKGALTRHRNVLKRAERIEVLKDEERWSEGDSLLGLEKVAHRKSHAGRKEKKAPD
	82	Arctic Sediment	Marine	2100351001 Predicted orf	local	MSIHRSFRSGNRLEKHRNVLSRIERIARLETVHKWEASADSVFKLPKIRNIKIR
	83	Arctic Sediment	Marine	2100351006 Predicted orf	local	MSIDRSLKIKGALSRHRNVLSRAERIEKLKDEERWSEGDSLLGLPKVAHRKSHAGRK
	84	Arctic Sediment	Marine	2100351006 Predicted orf	local	MTMDRTLKIHGGLARARSVLTRTERIQRLTDEGKFDPDQNSPFGLPKTRVHHSKVGVKAKKVEQ
	85	Arctic Sediment	Marine	2100351006 Predicted orf	local	MSIDRSLKIRGALERHRNVLSRAERIERLKDEEQWTEGDSLFGLPKVANRKSHSGR
	86	Arctic Sediment	Marine	2100351006 Predicted orf	local	SIHRSLKSKGKLERHRSVLTRTEQLDTLKNEERWQDGDSVYGLPKVRVIRVKRKKAAKEEE
	87	Arctic Sediment	Marine	2100351006 Predicted orf	local	MSIDRSLKVRGALTRHRNVLTRAERIEQLKDEEKWTEEDSLLGLPKVAHRKSHAGRK
	88	Arctic Sediment	Marine	2100351006 Predicted orf	local	MSIDRSLKIKGALTRHRNVLTRAERVEQLKDEDRWSEENSLLGLPKVAHRKSHAGR
	89	Arctic Sediment	Marine	2100351006 Predicted orf	local	MGLHPSLKRAEKLGGSRSVMKRTERIKWLKERGKWNENDTVLGLPKIKS
	90	Arctic Sediment	Marine	2100351011 Predicted orf	local	MSIDRSLKIKSALQRHRNVLTRAERIEQLKEEERWSEEDSLLGLPKVAHRKSHVGRKAAKEE
	91	Arctic Sediment	Marine	2100351011 Predicted orf	local	MSIDRSLKIKGALERHRNVLSRAERIEKLKEEERWDEGDAVLGLPKVAHRK
	92	Arctic Sediment	Marine	2100351011 Predicted orf	local	MSIDRSLKIKGTLTRHRNVLSRAERIEQLKDEEKWSEGDSVLGMPKVEHRKSHAGKK
	93	Arctic Sediment	Marine	2100351011 Predicted orf	local	VAVDKSLKTKGSLLRRRNVLSREERLQMLEKDGRWQDGDSVFGLPKVA
	94	Arctic Sediment	Marine	2100351011 Predicted orf	local	MSQHSSLKFGSKDKQHRSVLTRAERIKKLKGKDKWQEGNSVFGLPKVKMLRFKVKKEKA
	95	Arctic Sediment	Marine	2100351011 Predicted orf	local	MTIDKSLRVSRSLSRNRSVLTRNERLLRLKEADRWKEGDSPLGLPKVRVQRLVV
	96	Arctic Sediment	Marine	2100351011 Predicted orf	local	MSLHPSLKVDTAGAQQRTVLTRIERIKDLMKRDLWHDEHSVTGLPKT
	97	Arctic Sediment	Marine	2100351012 Predicted orf	local	MSLDKSLKSQNSLVRHRNVLSRAERLEKLEEEQRWSEGASVLGLAKVVHRKAAL
	98	Arctic Sediment	Marine	2100351012 Predicted orf	local	MSKHASLKIAGGAGGKRSVLKRFERIKLLKERGQWKEGKSPIGLPKTKPES
	99	Arctic Sediment	Marine	2100351012 Predicted orf	local	MSIDPSLKIKGALSRHRNVLTRAERIELLKEEERWSEEDSLLGLPKVAHRKSHAGRK
	100	Arctic Sediment	Marine	2100351012 Predicted orf	local	MTIDKSLKIRRGGAGNRSVLTRGERLEKLKETDRWKEGDSPLGLPKVRVRKLVMKKKKKKKVEEED
	101	Arctic Sediment	Marine	2100351012 Predicted orf	local	MSIDRSLKIKGALERHRNVLSRAERIEKLKEEERWDEGDSVLGLPKVAH
	102	Benzene degrading bioreactor	Engineered	2020627003 Predicted orf	local	MSIHPSLKIAAKGAQQRTVLTRIERIKNLMEKDLWVKDRNVTGLPKLKIL
	103	Benzene degrading bioreactor	Engineered	2061766000 Predicted orf	local	MAIHPSLNFSEKDQKQRSVLKRIERIKHMIEKNEWKEGDKIFGLPKIKSV
	104	Biofuel	Engineered	2156126002 Predicted orf	local	MSQHPSLRVGGKIRKIRNVMKRYERIDVLRAEGRIPEDKILGLPKTKPIEV
	105	Bioreactor	Engineered	orf_25308/24645314	SIMAP name/ID	MSQHPSLKGQGAIKARRNVLKRFERVELMKSVASTSPATVSSVFPRPNPRIN
	106	Bioreactor	Engineered	orf_25309/24645322	SIMAP name/ID	MSQHPSLKASKSVGSKRSVLKRYERVKLLQARGLWKEGRSPIGLPKTKPEE
	107	Bioreactor	Engineered	orf_25310/24645328	SIMAP name/ID	MSQHPSLKGKGAIKAKRSVLKRFERVELMKKRGVFKEGQRVIGLPKRSPRNSRSFQFEFRQARLRRAFFFMRINRFLAAAGL
	108	Colorado Soil	Soil	2209111000 Predicted orf	local	MSQHRSLKGASTITAKRNVLKRFERVELLKKRGQWKAGAKVIGLPKTKPDA

Metagenomes	109	Corn Field	Soil	2044078000 Predicted orf	local	MSQHRSLKGASTIAAKRNVLKRFERVELLKKRGQWKEGQKVIGLPKTKPDV
	110	Corn Field	Soil	2044078001 Predicted orf	local	MSQHKSLQGASGIVVKRNVLKRFERVAVLKKRGQWKEGDRVAGLRKTKPDV
	111	Freshwater lake	Freshwater	2149837010 Predicted orf	local	MTMHSSLKNAAKIQIRRNVLKRFERVDQLKAEGRWKDGDRGFGLVKTKPAE
	112	Freshwater lake	Freshwater	2166559021 Predicted orf	local	MSLHPSLKVDKAGAQQRTVLSRIERVKDLMKKGQWEDGRDVLSLPKTRIIKIKARKSKAVKEA
	113	Freshwater lake	Freshwater	2166559022 Predicted orf	local	MTMDQSLKVQAGAIKSRNVLTRAERVARLIELERFDEEANVVGMPKVRVMKVSL
	114	Freshwater lake	Freshwater	2166559022 Predicted orf	local	MSQHKSLQGSSGLVIKRNVLKRFERVDLLKKRGQWKAGDRVQGLRKTKPDV
	115	Freshwater lake	Freshwater	2166559022 Predicted orf	local	MTIDKSLKVKAGAAKSRNVLNRTERLTQLKETDRWRDGDSLLGMPKVRVMKISM
	116	Freshwater lake	Freshwater	2166559023 Predicted orf	local	MSQHPSLRIKSAGSKHRNVLKRYERIKKLEEAEKWDAEKDSVFNLPKVKSQKVK
	117	Freshwater lake	Freshwater	2166559023 Predicted orf	local	MSQHNSLKGSAKITAKRNVLKRFERVDLMRKRGQWKDDSRGFGLPKTKPAE
	118	Freshwater lake	Freshwater	2189573023 Predicted orf	local	MTLDKSLKVSGGSIKSRNIMTRVERLTRLKELERWKEGDPVLGIPKTR
	119	Freshwater lake	Freshwater	2189573023 Predicted orf	local	MTIDTSLKVRVGGAKTRNVLNRNERLAKLIEAGKWQDGSSVLGMPKVRV
	120	Freshwater lake	Freshwater	2199034001 Predicted orf	local	MSQHKSLQGASGIVRKRNVLKRFERVEILKKRGQWKEGDRVQGLRKTKPGV
	121	Freshwater lake	Freshwater	2199352004 Predicted orf	local	MTQHNSFKSSASSSGAKRNVLKRFERVELLKKRGTVKDLKRVTKLPKTKPDA
	122	Freshwater lake	Freshwater	2199352004 Predicted orf	local	MTLDKSLKVRAGAIKSRNVMTRAERIKRLQDLDKWTDESIVLGMAKVRVPKISL
	123	Guaymas Basin hydrothermal plume	Thermalspring	2061766003 Predicted orf	local	MSIHSSLHGADSLVGERSVLSRIERIEQLKKEGRFDPEEDSVLGLPKVRTKFR
	124	Hypersaline mat	Marine	orf_17822/19125034	SIMAP name/ID	MSLDRSLKTSGALSRHRNVLSRAERLEKLKEQERWKDGDSVLGLPKVSHRKTGG
	125	Hypersaline mat	Marine	orf_7405/19144357	SIMAP name/ID	MSLDSSLKTGGGLAKHRNVLTRAERIARLTDQEKFDMEGDSPLGLPKVANRKVVIGGKKKKKPGEEGDEDAAE
	126	Lake Washington Sediment	Freshwater	2006442541	IMG/m id	MSQHRSLRGATTLGAKRNVLKRYERVDMLKKRSEWKEGNRVTGLRKTKPPA
	127	Lake Washington Sediment	Freshwater	orf_116748/16198401	SIMAP name/ID	MSIDKSLKRKNTLARARSVLSRGERIKVLTDQEKWVDGRSPFGLPKVRVQKHVVKKAKKEKVEAVEGAEGAAAAPAAA
	128	Lake Washington Sediment	Freshwater	orf_94721/17099471	SIMAP name/ID	VPIDKSLRRRGRLVRSRNVLTREERVDQMKREETWIAGRSPMGLPKTRVVKVAAKKKKKAKEEGAEEGAVAPAGVLRPLLLPL
	129	Lake Washington Sediment	Freshwater	2006207001 Predicted orf	local	VPIDKSLRRRGRLVRSRNVLTREERVDQMKREETWIAGRSPMGLPKTR
	130	Lake Washington Sediment	Freshwater	2006207001 Predicted orf	local	HRSLKGASTITAKRNVLKRFERVELLKKRGQFKEGMKVIGLPETKPDA
	131	Lake Washington Sediment	Freshwater	2006207003 Predicted orf	local	MSIDKSLKRKNTLARARSVLSRGERIKVLTDQEKWVDGRSPFGLPKVRVQK
	132	Lake Washington Sediment	Freshwater	2006207003 Predicted orf	local	MSQHRSLKGAANVAAKRNVLKRFERIEVLKRRGQWKEGDKVLGLRK
	133	Lake Washington Sediment	Freshwater	2046860008 Predicted orf	local	MSQHRSLRAASTSGANRNVLKRFERVALLKKRGQWKTGDRVTGLRKTKADA
	134	Lake Washington Sediment	Freshwater	2084038009 Predicted orf	local	MSQHPSLRSDSVGAKHRNVFKRLERVKKLQEMGKWSDRGTAYNLPKVKSLKVKMK
	135	Lake Washington Sediment	Freshwater	2084038009 Predicted orf	local	MSQHPSLRSSDKDKAQRSVLKRYERVKILKDKEKWTEEDSVYKLPKLKIIKFKVKKR
	136	Lake Washington Sediment	Freshwater	2084038009 Predicted orf	local	MSQHPSLRSSDNDKKQRSVLKRYERISLLKDKEKWKEDDSVFGLPKVKIVKLQS
	137	Lake Washington Sediment	Freshwater	2084038009 Predicted orf	local	MSRHPSYGKSSKSAKKRNVLKRFERVDVLRKLGRWKDGENKKVTGLPKTPIL
	138	Lake Washington Sediment	Freshwater	2084038009 Predicted orf	local	MSVHKSLRTGGSLVKHRNVLTRSERIKFLEAAGEWREGQHKVLGLRKVKSQ
	139	Lake Washington Sediment	Freshwater	2084038009 Predicted orf	local	MSQHPSLRSSDKDKKARSVLKRYERIKTLQDKEKWDEKKDSVFGLPKVKVTRFHL
	140	Lake Washington Sediment	Freshwater	2084038009 Predicted orf	local	MSQHPSLRSSDKDKAQRSVLKRFERIKILKDKESWTESDSVYKLPKLKIITKFKV
	141	Lake Washington Sediment	Freshwater	2088090005 Predicted orf	local	MSQHRSLRAASTLGGKRNVLKRFERVEILKKRGEWKEGGRITGLRKTKAVVA
	142	Lake Washington Sediment	Freshwater	2088090005 Predicted orf	local	MSLHRSLRAASTSGANRNVLKRFERVALLKKRGQWKTGDRVTGLRKTKADA
	143	Lake Washington Sediment	Freshwater	2088090005 Predicted orf	local	MSQHRSLRAASVLGGKRNVLKRFERVEILKKRGQWKEGDRVTGLRKTRRET
	144	Lake Washington Sediment	Freshwater	2088090005 Predicted orf	local	MSQHRSLRAASTLGGKRNVLKRFERTALLEKRGQRKAGDKITGLRKTKPEA
	145	Lake Washington Sediment	Freshwater	2088090005 Predicted orf	local	MSQHRSLRAASTLGGKRNVMKRFERVEILKKRGQWKPGDRITGLRKTKADA
	146	Lake Washington Sediment	Freshwater	2088090005 Predicted orf	local	MSQHASLKAADRIVEKRNVLKRFERIEIMRANKKWKKGDRGLGLPKTKSE
	147	Lake Washington Sediment	Freshwater	2088090005 Predicted orf	local	MSQHRSLRAASTLGGKRNVMKRFERIEVLKKRGQWKAGDRITGLRKTKASV
	148	Lake Washington Sediment	Freshwater	2088090005 Predicted orf	local	MSMHSSLKSGTKMAAKRNVMKRFERIDLLKAKGKWKAGDRAFGLPKIKPAA
	149	Lake Washington Sediment	Freshwater	2088090006 Predicted orf	local	MTLDKSLRVSTSSIQNRNVLTRAERVLKLKDTERWKEGDQVLGLPKVRVSKVSLKKK
	150	Lake Washington Sediment	Freshwater	2088090006 Predicted orf	local	MAQHRSLRAASTLGGKRNVMKRFERIEILKKRGQWKDGDRITGLRKTKADV
	151	Lake Washington Sediment	Freshwater	2088090006 Predicted orf	local	MSQHRSLRAASTLGGKRNVLKRFERTAILEKRGQRKAGDKITGLRKTKPDA
	152	Lake Washington Sediment	Freshwater	2088090006 Predicted orf	local	MSQHRSLRAASTLGGKRNVLKRFERTAILEKRGQRKSGDKITGLRKTKPDA
	153	Lake Washington Sediment	Freshwater	2088090006 Predicted orf	local	MSLHTSLKIDTAGAQQRSVLTRIERMKEMMKKGLWKEDRSAVGLPKTKIVKVKAR
	154	Lake Washington Sediment	Freshwater	2088090006 Predicted orf	local	MTMDKSLRTRRGLVRSRSVLTRAERLIQLKESARWKEGDSPLGLPKVRVFKLAM
	155	Lake Washington Sediment	Freshwater	2088090007 Predicted orf	local	MSLHPSLKVDLAGAQHKTVYSRIQRVKSLMKKGKWLEGQAVTGLPK
	156	Lake Washington Sediment	Freshwater	2088090007 Predicted orf	local	MSQHKSLQGISGLVVKRNVLKRFERVDLLRKRGLWKVGDRVQGLRKTKPDV
	157	Lake Washington Sediment	Freshwater	2088090007 Predicted orf	local	MSQHKSLQGSSGLVIKRNVLKRFERVDILKKRGQWKAGDRVQGLRKTKPDV
	158	Lake Washington Sediment	Freshwater	2088090007 Predicted orf	local	MSRHPSYGKSNKGATKRNVLKRFERVDVLRKLGRWKDGENKKVTGLPKTPVT
	159	Lake Washington Sediment	Freshwater	2088090007 Predicted orf	local	MSRHPSYGKSSKGGKKRNVLKRFERVNVLRELGRWKDGENTRVTGLPKTPIL
	160	Lake Washington Sediment	Freshwater	2088090007 Predicted orf	local	MSQHSSLKGSGKITAKRNVLKRFERVNLLKKRGQWKDGDRGLGLRKTKPEV
	161	Lake Washington Sediment	Freshwater	2088090007 Predicted orf	local	MSQHRSLRSSGTIAAKRNVLKRFERVDLLTKRGQFKEGRSVLGLPKTKPPE
	162	Lake Washington Sediment	Freshwater	2088090007 Predicted orf	local	MSIHPSLTISEKDKKARSVLKRTERIRQMWEKNKWKEGDSVYGLPKIKTLR
	163	Lake Washington Sediment	Freshwater	2088090007 Predicted orf	local	MSLDSSLKVKGALTRHRNVLTRAERVEKLKDEEKWDSNSSVLGLPKVGHRKIRAGRK
	164	Lake Washington Sediment	Freshwater	2088090007 Predicted orf	local	MSIHSSLKTKSGNLASHRGVFTREERISKLTEAEKFGAGKSPLGLPKVLSVKLGGAKK
	165	Lake Washington Sediment	Freshwater	2088090007 Predicted orf	local	SKHNSLKSNATVGGKRSVLKRFERVKLLKERGEWKAGRSPVGLPKTKHEG
	166	Lake Washington Sediment	Freshwater	2088090007 Predicted orf	local	MSIHPSLKSTQKIKKQRSVLKRTERLRMMMEKDQWKEGDDVFGLPKI
	167	Lake Washington Sediment	Freshwater	2088090009 Predicted orf	local	MTMDKSLRVRKGSSSARGVLSRAERIAKMKEQERWTEGRSPLGLPKVRVQKLSM
	168	Lake Washington Sediment	Freshwater	2088090009 Predicted orf	local	MSQHSSLRGASALGAKRNVLKRFERVDLLKKRGQWKAGDRVTGLRKTKPDA
	169	Lake Washington Sediment	Freshwater	2088090009 Predicted orf	local	MSQHRSLKGSNKIQAKRNVMKRFERVEALKKQGRWKDGQRVFGLPKTKPTS
	170	Lake Washington Sediment	Freshwater	2088090009 Predicted orf	local	MSIDRSLKVKDALSRHRNVLSRAERLEMLQEQERWQEGGSVLGLPKVAHRKTVAGKK
	171	Lake Washington Sediment	Freshwater	2100351007 Predicted orf	local	MSMHPSLKASEKGKKQRSVLGRIERLKVMLDKEQWKEGSAVYGLPKIK
	172	Lake Washington Sediment	Freshwater	2100351007 Predicted orf	local	MSLHSSLKAADKIKTRRNVLKRFERIEVLKNRAVSRRASAYSAFRR
	173	Lake Washington Sediment	Freshwater	2100351007 Predicted orf	local	MSQHPSLRSASKIASKRNVLKRFERVDLLASRGKWKEGDRGIGLPKTKS
	174	Lake Washington Sediment	Freshwater	2100351007 Predicted orf	local	MSRHTSFGKGSKGIKKRNVLKRFERVDKLRSLGRWNEKDNQRVTGLPKTPVS
	175	Lake Washington Sediment	Freshwater	2100351007 Predicted orf	local	MSIHPSLVISEKDKMARSVLKRTERIRQMHEKGKWKEGDSVYGLPKMKTLRIKIKKKEKVAKAETTT
	176	Lake Washington Sediment	Freshwater	2100351007 Predicted orf	local	MSQHPSLRSSEKDKKQRSVLKRYERIKTLQDKEKWDEKKDSVFGLPKVKVTRFKIKKEKAAA
	177	Lake Washington Sediment	Freshwater	2100351007 Predicted orf	local	MTQHPSLKGSQLGTKFRAVLKRYEKVKELEEKDKWNEEKGSIYKLPKLRRIKFKI
	178	Lake Washington Sediment	Freshwater	2100351007 Predicted orf	local	MSQHPSLRVDKVGARHRNVLKRYERVDRLKEQSKWQEARSVFGLPKVKSQ
	179	Lake Washington Sediment	Freshwater	2100351007 Predicted orf	local	MSIHPSFGKASGAMREKRSVMKRFERVEELRKRGQFKPGDRVFGLRKTSCL
	180	Lake Washington Sediment	Freshwater	2100351007 Predicted orf	local	MTRHASYGKSTKSSKKRNVLKRYERIDVLRKLGRWIDGANKKVTGLPKTPVV
	181	Lake Washington Sediment	Freshwater	2100351007 Predicted orf	local	MSQHPSLRSSDKDKKQRSVLKRYERIKTLQDKEKWDEEKDSVFGLPKVKVTRFKI
	182	Lake Washington Sediment	Freshwater	2100351007 Predicted orf	local	MSQHPSLRSSDKDKAQRSVLKRYERVKILKDKEKWTEEDSVYKLPKLKIIKFKVKKEKA
	183	Lake Washington Sediment	Freshwater	2124908000 Predicted orf	local	MSQHPSLRSSGKDKKQRSVLKRFERVKTLQDKEKWDEEKDSVFGLPKVKVTRFKI
	184	Lake Washington Sediment	Freshwater	2124908000 Predicted orf	local	MSQHPSLRSSDKDKKQRSVLKRYERIKTLQDKEKWDEKKDSVFGLPKVKVTRFKIKKEKAAA
	185	Lake Washington Sediment	Freshwater	2124908000 Predicted orf	local	MSRHPSFGKSTASGSKRNVLKRFERVDALKKLGRWKNEATRVTGLPKTPI
	186	Lake Washington Sediment	Freshwater	2124908000 Predicted orf	local	MSQHPSLRSSDNDKKNRSVLKRYERISLLKDKEKWKEDDSVFGLPKVRS
	187	Lake Washington Sediment	Freshwater	2124908000 Predicted orf	local	MSQHPSLRSSDNDKKNRSVLKRYERISLLKDKEKWKEDDSVFGLPKVRS
	188	Marine Sediment	Marine	2022819504	IMG/m id	VSIDKSLSNKGKLIRHRSVLTRSERVKALTNEGVWGEERSVFGLPKVKTIKMRKKGKSTKEKGEEASSADGKKA
	189	Marine Water	Marine	EBF37225.1	NCBI accession	MTMDKSLRARAGLIRSRSVLTRAERIERLKASDRWKEGDSPFGLDKVRVYKLAIKKKKKKKEEE
	190	Marine Water	Marine	EDJ30329.1	NCBI accession	MTMDKSLRVRKGASRPRGVLTRAERIAKLEELGKWEDGQRPLGLPKVRVYKVVLKKKKKAKT
	191	Marine Water	Marine	EDJ07517.1	NCBI accession	MTMDKSLRVRKGSSRARGVLTRAERIAKLQELGTWQDGQRPLGLPKVRVYKVVLKKKKKAKEEPEAEGGKKKK
	192	Marine Water	Marine	ECM33690.1	NCBI accession	MTMNKSLRVRKGGGNRGVLSRAERITKLKELEKWKEGQSPLGLAKVRVQKISMKKKKAKDEPEATEST
	193	Marine Water	Marine	EBV87401.1	NCBI accession	MTMDKSLRVRKGATSARGVLTRAERIAKLKEQDIWKDGRSPLGLPKVRVQKMAMKKKVKKKDEAAEGAAAAKPAAGGAAGAK
	194	Marine Water	Marine	EBH72892.1	NCBI accession	MTMNKSLRVRKGGGNRGVLSRAERITKLKELEKWEEGQSPLGLAKVRVEKISMKKKKPAKDEPDATEGK
	195	Marine Water	Marine	EBT86298.1	NCBI accession	MTMNKSLRVRKGGGGNRGVLSRAERITKLKELEKWEEGQSPLGLPKVRVQKISMKKKKKAKDDSDAAEST
	196	Marine Water	Marine	ECG67749.1	NCBI accession	MTMNKSLRVRKGGGGNRGVLSRAERITKLKELEKWEEGQSPLGLPKVRVQKISMKKKKKAKDDSDTTAEST
	197	Marine Water	Marine	EDJ11357.1	NCBI accession	MTMNKSLRIRKGGGGSRGVLSRAERITKLKELERWEDGQSPLGLPKVRVYKISLKKKKKTKADEGEEDGKKAKKK
	198	Marine Water	Marine	orf_97861/16136525	SIMAP name/ID	MSRHRSLSSSNKLGAKRNVLKRFERVALLKERGQWKEGERVTGLRKTKPAE
	199	Marine Water	Marine	orf_75586/31654001	SIMAP name/ID	VSLDKSLRRGNKMSGTRNVLKRAERVSQLEAEDRWTEKSTVLGMPKVRVKRTVIGKKEEEEIQGRRFRVI
	200	Marine Water	Marine	2156126010 Predicted orf	local	MTMNKSLRVRKGSGGSRGVLSRAERITKLKDLEKWEEGQSPLGLPKVRVYKISM
	201	Marine Water	Marine	2156126011 Predicted orf	local	MSQHPSLKVGGAAEAKRTVLTRYERVELLRKRGELKEGDRITGLKKTKPGE
	202	Marine Water	Marine	2156126013 Predicted orf	local	MSQHPSLKVGGAAEAKRSVLTRYERVELLRKRGELKEDDRITGLKKTKPEE
	203	Marine Water	Marine	2162886003 Predicted orf	local	MSQHPSLKVGGAAEAKRTVLTRYERVELLRKRGELKEDDRITGLKKTKPGE
	204	Marine Water	Marine	2162886004 Predicted orf	local	MSQHPSLKPPGGLMKKRNVLKRFQRIDLLRQRNQWKEGDRVVSLPKTKPAE
	205	Marine Water	Marine	2162886004 Predicted orf	local	MSQHPSLKVGGAAEAKRTVLTRYERVELLRKRGELKEDDRITGLKKTKPEE
	206	Marine Water	Marine	2189573014 Predicted orf	local	MSQHPSLKVGGAAEAKRSVLTRYERVELLRKRGELKEGDRITGLKKTKPEE
	207	Marine Water	Marine	2189573019 Predicted orf	local	MSIHKSLSLGSGISTERSVFSRRERVDRLIKEGKLSPEDSPIGLPKVRTQFKVVSRKAKKAK
	208	Marine Water	Marine	2189573019 Predicted orf	local	MSLHRSLKTKPAGLNQHRNVLTRAERIEHLADADEFDKESGCPIGLPKVGSR
	209	Meryland Soil	Soil	2032320004 Predicted orf	local	MTIDKSLKVKRGASRNRSVLTRAERLQKLKIADRWQEGDSPLGLPKVRVRKLTI
	210	Meryland Soil	Soil	2035918006 Predicted orf	local	MSMHSSLKGAAKIRTKRNVLKRFERIDVLKKDGKWKEGDRAFNLPKTKSEA
	211	Meryland Soil	Soil	2035918006 Predicted orf	local	MSIDSSLRLKDALVRHRNVLTRAERISKLKDEERWEEDYSVLGLPKSGTQK
	212	Meryland Soil	Soil	2035918006 Predicted orf	local	MSIHRSLRTVSSALNQHRNVLKRSERITLMTDRDAFDPKNDSPIGLVKYANRKATTGKK
	213	Meryland Soil	Soil	2035918006 Predicted orf	local	RSLRAASTLGGKRNVLKRFERVELLKKRGQWKSGDRITGLRKTKPDV
	214	Miscanthous Field	Soil	2044078003 Predicted orf	local	MPIDKTLKVKAGGIKNRNVLKRAERVEKLRELEKFQSGMSVLGLPKVRVIKLAI
	215	Miscanthous Field	Soil	2044078003 Predicted orf	local	MTIDKSLKVKRGATRNRSVLTRVERISRLREADRWTEGDSPLGLPKVRVRKLTM
	216	Miscanthous Field	Soil	2162886011 Predicted orf	local	MSQHRSLKGASTIAARRNVLKRFERVELLKKRGQWKDSKVIGLPKTKPDV
	217	Miscanthous Field	Soil	2162886011 Predicted orf	local	MSQHRSLKGASTIAAKRNVLKRFERVDLLKKRGQWKETSKVVGLPKTKPDL
	218	Nevada Hot Spring	Thermalspring	2149837004 Predicted orf	local	MTMDKSLRVRKALVRNRSVLTRAERIQRLLAMDRWQEGDSPLGLPKVRVQKISM
	219	Nevada Soil	Soil	2032320002 Predicted orf	local	MTIDRSLRVRSGSIANRNVLTRGERLQKLKESERWQPGSSVYGLPKV
	220	Nevada Soil	Soil	2032320002 Predicted orf	local	MSMDRSLKTKGNYSGTRSVLTRAERIAKMQTDKKFDPKKDKALGLPKTLIGKV
	221	Nevada Soil	Soil	2032320002 Predicted orf	local	MSQHRSLKGQSTIAAKRNVLKRFERVEVLKSRGQYKDGDKVIGLRKTKPNE

	222	Nevada Soil	Soil	2065487013 Predicted orf	local	MTIDKSLRTRRGVTRSRNVLTRAERIEKLQQQDRWTEEDGPFSLPKVRVYQVVI
	223	Nevada Soil	Soil	2081372006 Predicted orf	local	MSMDRSLKAGAGLIRHRNVLTRDERLLRLQGDGKWDETKSVLGLVKVGNRKMIIGGK
	224	Nevada Soil	Soil	2119805009 Predicted orf	local	MSIDKSLKKASSMARSRNVLTRAERLIILQDDERWTPALGVYNLPKTKYRRLPPGQSGPRRVEPK
	225	Nevada Soil	Soil	2119805009 Predicted orf	local	MTMDKSLKIRRGLIRARGVLTRDERILRLKEADRWQEGASPLGLPKVRVFKLTM
	226	Nevada Soil	Soil	2119805010 Predicted orf	local	MSQHRSLRAASTLGGKRNVLKRFERTELLKKRGQWKAGDRITGLRKTKPES
	227	Nevada Soil	Soil	2119805010 Predicted orf	local	MSQHRSLKGTSTIAARRNVLKRFERVELLKKRGQWKKDSKVIGLPKTKPDV
	228	Nevada Soil	Soil	2119805010 Predicted orf	local	MSQHRSLRAAATLGGKRNVLKRFERLEILKRRGQWKEGERITGLRKTKADA
	229	Nevada Soil	Soil	2119805010 Predicted orf	local	MSQHQSLKGANKIQARRNVLKRFERVESLKKQGRWKDGDRVFGLPKTKPE
	230	Nevada Soil	Soil	2119805012 Predicted orf	local	MSQHRSLKGASTITAKRNVLKRFERVELLKKRGQWKDGAKVLGLPKTKPDA
	231	North Carolina Soil	Soil	2035918004 Predicted orf	local	MSIDKSLRRKNSLQRARNVLTRGERIKTLQNEERWQVGRSPFGLPKV
	232	North Carolina Soil	Soil	2035918004 Predicted orf	local	MSLDKSLKKAGSLARARNVLSRAERLALLQEDELWKPAAGVYNLPKTKYRRLAPGQSGPKRPAATS
	233	North Carolina Soil	Soil	2040502001 Predicted orf	local	MSQHRSLRSSGTIAAKRNVLKRFERVDLLKKRGQWKEGMRVLGLPKTKPEA
	234	North Carolina Soil	Soil	2124908001 Predicted orf	local	MSQHRSLKGASTITTKRNVLKRFERVELLKKRGIFKPGDKVIGLPKTKPDA
	235	North Carolina Soil	Soil	2124908001 Predicted orf	local	MSQHRSLRAASTLGGKRNVLKRYERTALLKKRNQWKDGDRITGLRKTKPEA
	236	North Carolina Soil	Soil	2124908001 Predicted orf	local	MSQHRSLKGTGTIAARRNVLKRFERVELLKKRGQWKDGQKVIGLPKTKPDV
	237	Oak Ridge ground water	Freshwater	orf_339907/27887473	SIMAP name/ID	MSLHKSLIPPSKFGGHRNVLRRSERVAKLESEQKLKPEDSVYGLQKVKMIKVRQKKKPAEEKEAVAPGAAQGAAAPGAPDP
	238	Oak Ridge Soil	Soil	2032320005 Predicted orf	local	MSQHRSLKGASTIAAKRNVLKRFERVEILKKRGQWKDGQKVIGLPKTKPDV
	239	Oak Ridge Soil	Soil	2032320006 Predicted orf	local	SLKGTSTIAAKRNVLKRFERVELLKKRGQWKETSKVIGLPKTKPDV
	240	Oak Ridge Soil	Soil	2032320006 Predicted orf	local	MSQHPSLRAASTTGGKRNVLKRFERVELLKKRGQWKEGDRVVGLRKTKPSE
	241	Plant Endophytes	Host-associated	2509281318	IMG/m id	MTIDKSLKIKAGAAKTRNVLTRPERLTKLIAEDRWEEGDPVYGMPKVRVAKLALKKKKKVKKEDEEEK
	242	Plant Endophytes	Host-associated	2509285195	IMG/m id	MTMDQSLKVKAGAIRSRNVLTRAERVARLKELEKFDENNSIVGMPKVRVQKISLKKKKKVKKADDEK
	243	Plant Endophytes	Host-associated	2509291597	IMG/m id	MAIDKSLKVKAGATANRSVLTRVERIEKLRETGAFDEDSSPFGLQKVRVRKLTMKKKKKPKKADEDDK
	244	Poplar decaying biomass	Soil	2049071722	IMG/m id	MSQHRSLKGQSTIAAKRNVLKRFERVEILKSRGQYKQGDKVIGLRKTKPNE
	245	Poplar decaying biomass	Soil	2049496894	IMG/m id	MSQHKSLRGASTIAAKRNVLKRFERVDLMKKRGVWKEGTRVLGLPKTKPEV
	246	Poplar decaying biomass	Soil	2049658830	IMG/m id	MSQHRSLKGASTITVKRNVLKRFERVELMKKRGVWKADSRVIGLPKTKPEE
	247	Poplar decaying biomass	Soil	2049745478	IMG/m id	MSQHPSLKGKGSIQAKRSVLKRFERVELMKKRGQFKPGQRVIGLPKTKPEE
	248	Sakinaw lake	Freshwater	2263495108	IMG/m id	MTRHQSYGKASKQEKKRNVLKRFERIEVLRDMGKWKDGENKRVTGLPKTLIKK
	249	Sakinaw lake	Freshwater	2263599129	IMG/m id	MSQHTSLRGSGKITAKRNVLKRFERVNVLKERGQWKAGDRGLGLRKTKPSV
	250	Sakinaw lake	Freshwater	2263603253	IMG/m id	MTMDKSLRVRKGSASTRGVLTRAERITKLKEQERWQDGRSPLGLPKVRVFKMAMKKKKKAKEEGADAATPAAAGKAAAPAAK
	251	Sakinaw lake	Freshwater	2263739694	IMG/m id	MSRHSSFGKASKAEQKRNVLKRFERVKVLKKLGKWTEGSNVKVTGLPKTPVK
	252	Sakinaw lake	Freshwater	2263954371	IMG/m id	MSRHPSFGKASKGETKRNVLKRFERVKMLKKLGKWQEGVNTTVTNLPKTPAK
	253	Sakinaw lake	Freshwater	2264164215	IMG/m id	MSRHPSFGKASKGETKRNVLKRFERVKALKQLGKWQEGVNTIVTNLPKTLVK
	254	Sakinaw lake	Freshwater	2088090031 Predicted orf	local	MSIHPSFKSGKNKQQRSVLKRLARIILLRKENKWNEEKDSIFGLPK
	255	Sakinaw lake	Freshwater	2088090031 Predicted orf	local	MSIDPSLKIKGALSRHRNVLKRAERIELLKEEDRWTEEDSLLGLPKVAHRKSHAGRK
	256	Soil	Soil	2124908006 Predicted orf	local	MSQHRSLRAVATMGGKRNVLKRFERVGLLKKRGQWKEGDRITGLRKTKPDA
	257	Soil	Soil	2124908006 Predicted orf	local	MSLDRSLKVSGKLAGKRSVMTRTERIAKLSTDKKFDGAKGSALNLPKT
	258	Soil	Soil	2124908006 Predicted orf	local	MSQHRSLKGASTIVAKRNVLKRFERVELLKKRGQWKEGRKVIGLPKTKPDA
	259	Soil	Soil	2124908006 Predicted orf	local	MSQHRSLRAASTTGGKRNVLKRFERVALLKKRSQWKEGDRVTGLRKTKPEA
	260	Soil	Soil	2124908006 Predicted orf	local	MSIHSSLRGVDTLKGERSVFTRVERLAVLKKAGKFDEKAGSVYGLPKVRTR
	261	Soil	Soil	2124908007 Predicted orf	local	MSQHRSLRAAATLGGKRNVLKRFERLEVLQKARPMERGRAHHRPATK
	262	Soil	Soil	2124908007 Predicted orf	local	MSQHRSLRAIATVGGKRNVLKRFERVGLLKKRGHWKEGDRITGLRKTKPDA
	263	Soil	Soil	2124908007 Predicted orf	local	MSQHRSLRATATLGAKRNVWKRFERVELLKKRGQWKEGDPVTSLRKTKPPI
	264	Soil	Soil	2124908008 Predicted orf	local	MSQHKSLQGSSGLVVKRNVLKRFERVDILKKRGQWKAGDRVQGLRKTKPDV
	265	Soil	Soil	2124908008 Predicted orf	local	MTIDKSLKVKRGMSRVRNVLTRAERLGKLKDVERWKEGDPVLGLPKVRVM
	266	Soil	Soil	2124908008 Predicted orf	local	MSLDRSLKSANALIRHRNVLTRPERLAQLTDEGKWDDKKSVYGLPKVGHRKAAVVK
	267	Soil	Soil	2124908009 Predicted orf	local	MSQHRSLKGASTITAKRNVLKRFERVELLKKRGQFKDGQKVLGLPKTKPDV
	268	Soil	Soil	2124908009 Predicted orf	local	MSQHRSLKGVSTIAARRNVLKRFERVELLKKRGQWKKGNKVIGLPKTKPDV
	269	Soil	Soil	2124908009 Predicted orf	local	MSIDKSLKKASSLARARNVLTRPERLAALQEDDRWTPEKGVYNLPKTKYRRL
	270	Soil	Soil	2124908009 Predicted orf	local	MSQHRSLKGASTIITKRNVLKRFERVELLKKRGLFKPGDKVIGLPKTKPDA
	271	Soil	Soil	2124908009 Predicted orf	local	MSLDGSLKGKSTLERHRNVLRRGERIDVLEDAEKWTETNGSVFGLVKVG
	272	Soil	Soil	2124908009 Predicted orf	local	MSQHKSLQGSSGIVVKRNVLKRFERVDLLKNAANGKPAIASRASARPSPTFEH
	273	Switchgrass rhizosphere	Soil	2021593004 Predicted orf	local	MSQHRSLKGASTIAAKRNVLKRFERVDLLKKRGQWKETSKVIGLPKTKPDV
	274	Switchgrass rhizosphere	Soil	2124908021 Predicted orf	local	MSQHRSLKGASTIAAKRNVLKRFERVELLKKRSQWKDGQKVIGLPKTKPDV
	275	Switchgrass rhizosphere	Soil	2124908021 Predicted orf	local	MSQHRSLRGASTIAAKRNVLKRFERVDLLKKRGQWKETSKKIGLPKTKPDV
	276	Switchgrass rhizosphere	Soil	2162886007 Predicted orf	local	VSINKTLKIKNTMVRARNVLKRVERIAGMKENESWKDGQAPIHLPKFRVIKSAL
	277	Switchgrass rhizosphere	Soil	2162886013 Predicted orf	local	MSQHRSLKGASTIAAKRNVLKRFERVELLKKRGQWKERQKVLGLPKTKPD
	278	Switchgrass rhizosphere	Soil	2162886013 Predicted orf	local	MTIDKSLRVKAGAIKNRNVLSRAERIAKLLETDRWKDGDKVLGLPKVRVQ
	279	Terephthalate Degrading Bioreactor	Engineered	orf_111250/28070694	SIMAP name/ID	MTMDKSLRIRRALVKVRSVLTRAERIKRLQDLDRWGEESSPIGLPKVRVQKISMKKKKKKKKEEDTAEQGKGK
	280	Waseca Soil	Soil	2001200001 Predicted orf	local	MSLDRSLKVAGGLAAKRSVMTRPERIAKMTEDKKFDPKTGKVLGLPKTLVSK
	281	WWTP	Engineered	2022004001 Predicted orf	local	MSQHRSLRSAGSTAAKRNVMKRFERVALLKTRGQWKDGDRVTGLRKTKP
	282	Yellowstone Hot Spring	Thermalspring	2015219002 Predicted orf	local	MTMDKSLRVRKALVRNRSVLTRAERIQRLLVMDRWQEGDSPLGLPKVRVQKITM
	283	Yellowstone Hot Spring	Thermalspring	2015219002 Predicted orf	local	SLRAASVAGRKRNVLKRFERIALLKKRGLWKPGDRITGLPKTKPDAL
	284	Yellowstone Hot Spring	Thermalspring	2016842003 Predicted orf	local	MSRHPSFGKVGLAKKRNVLKRFQRIDQLKKQGRFKEGDSVYGLPKTKPQE
	285	Yellowstone Hot Spring	Thermalspring	2016842008 Predicted orf	local	MSVHRSLKTKNALERHRNVLSRAERLKILKDQERWDETKSVFGLPKVAHR

Chapter IV

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

Published in ISME journal

2014

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

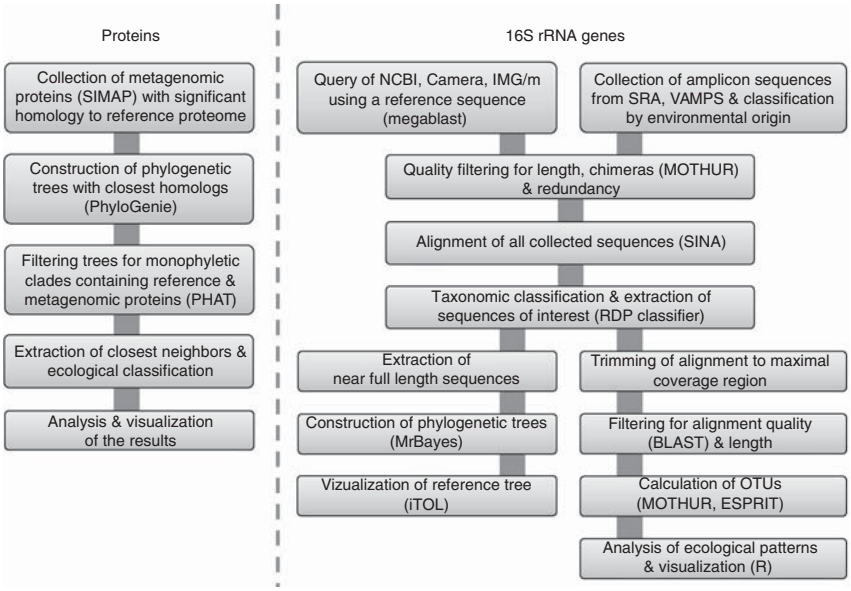

Figure 1 Flow chart illustrating the main steps in the analysis of metagenomic and amplicon sequence data for inferring diversity and ecology of defined microbial taxa. In this study, this approach was used for investigating the phylum *Chlamydiae*. A detailed description of each step is provided in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

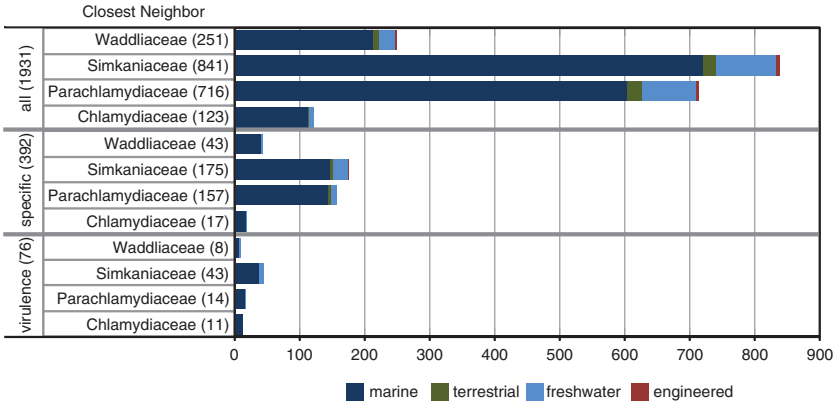

Figure 2 Ecological and taxonomic classification of putative chlamydial proteins in metagenomic sequence data. Proteins were classified based on their respective closest neighbor in maximum likelihood trees. Environmental origins grouped in four general categories are color coded. 'All' refers to all 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.

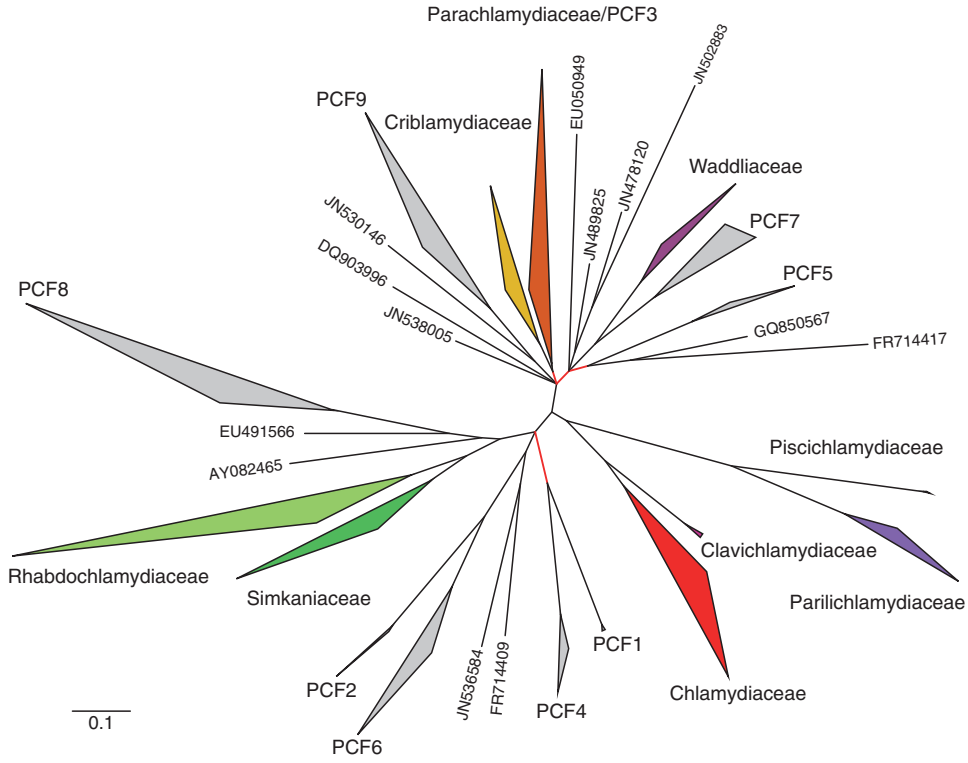

Figure 3 Relationships of described and predicted families in the phylum *Chlamydiae* based on near full-length 16S rRNA gene sequences (>1100 nt). The phylogenetic tree was calculated using Bayesian inference (MrBayes; (Huelsenbeck and Ronquist, 2001)). Branches with a posterior probability lower than 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

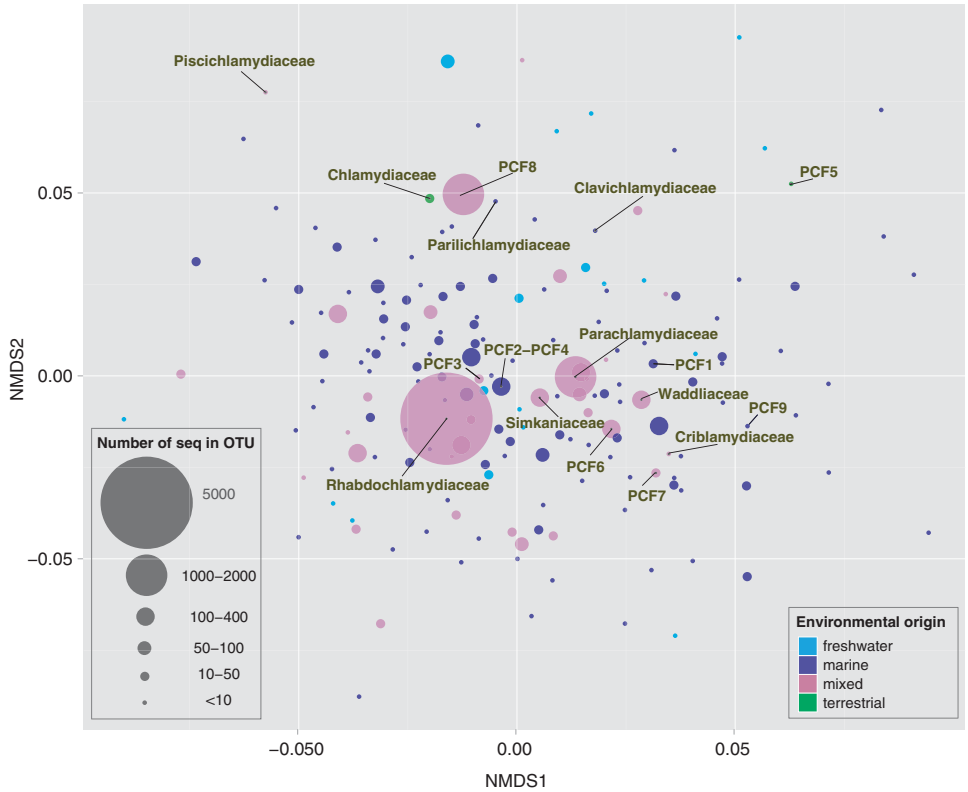

Figure 4 Diversity and ecology of chlamydial families based on NMDS of OTU distances. Filled circles represent family-level OTUs, with the size corresponding to the number of sequences included. The distance between circles indicates the relative distance between OTUs. Colors represent the environment 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 (Moon-van 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)

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

Supplementary information

Supplementary methods

Identification 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. All predicted proteins from environmental metagenomic datasets in SIMAP (October 2009) that showed highest similarity to a protein from sequenced chlamydial organisms were extracted using an E-value cutoff of 10^{-20} and an alignment coverage of at least 50% for both query and subject. To exclude incidental homologs that show a comparable homology to chlamydial and to other bacterial proteins, we included only those proteins that showed an E-value for the chlamydial homolog at least 10 times lower than the E-value for the closest non-chlamydial homolog. Subsequently, maximum likelihood phylogenetic trees [using RAxML (Stamatakis, 2006) and a JTT substitution model, discrete gamma distribution, 100 bootstrap resamplings] were calculated for all proteins in the quality-filtered, reduced set of metagenomic proteins. SIMAP was used to extract the closest homologs for each metagenomic protein. Alignment and tree calculations were performed using PhyloGenie (Frickey and Lupas, 2004). The tool PHAT included in the PhyloGenie package was then used to select all those trees in which the metagenomic protein formed a monophyletic group with known chlamydial homologs (bootstrap support >70%). Only those metagenomic proteins that were monophyletic with known chlamydial homologs in this analysis were considered of putative chlamydial origin. Metagenomic proteins were classified as chlamydia specific if the respective phylogenetic tree only contained chlamydial homologs as determined by PHAT. The published list of virulence associated chlamydial genes from the study of Collingro *et al.* (Collingro et al, 2011) was used as reference for the detection of virulence associated metagenomic chlamydia-like proteins using an in-house Perl script. Putative chlamydial metagenomic proteins that were shared between more than one tree were grouped together and considered orthologs potentially serving similar functions.

To identify the closest neighbors of putative chlamydial metagenomic proteins in phylogenetic trees, an in-house Perl script was used to analyze the tree topology. Briefly, a path of nodes was created connecting the leaf representing the metagenomic protein to the root of the tree (determined by midpoint rooting).

Starting from the leaf, the first internal node comprising a leaf represented by a non-metagenomic protein was determined. Then the distances, as sum of branch lengths, of all non metagenomic leaves branching from this internal node were calculated, and the leaf with the shortest distance was returned as the closest neighbor.

Identification of chlamydial 16S rRNA gene sequences

To search for chlamydial 16S rRNA genes in different sequence repositories, the full-length 16S rRNA gene sequence of *Simkania negevensis* (NR_029194) was used as a representative of the phylum *Chlamydiae* (other known chlamydial 16S rRNA gene sequences were also used and generated the same results). A very relaxed initial filtering step including all sequences with at least 60% nucleotide sequence identity ensured that no chlamydial 16S rRNA gene was missed, and the initial dataset consequently also comprised a large, redundant fraction of clearly non-chlamydial 16S rRNA genes. The NCBI databases nr and env_nr (Wheeler et al, 2008) and the CAMERA nucleotide databases Sanger and 454 (Sun et al, 2011) were queried using the blastn service provided by CAMERA. For IMG/m (Markowitz et al, 2012) metagenome sequences were downloaded and queried locally with the same settings. Since there is a redundancy in the metagenomic datasets in CAMERA and IMG/m, only those samples unique for IMG/m were analyzed. Other NCBI databases (EST, GSS, WGS, and TSA) were queried using megablast (Wheeler et al, 2008) at the NCBI website. All searches were performed in September 2011.

In addition, all amplicon 16S rRNA gene sequences spanning through variable regions V4 and V6 in the VAMPS database (http://vamps.mbl.edu/; (Huber et al, 2007)) and amplicon 16S rRNA gene sequences obtained from environmental sources using the 454 Titanium technology in the SRA database (Kodama et al, 2012) were downloaded in April 2012. The low quality ends of the SRA reads were trimmed to remove ambiguous nucleotides (N) at or close to the end of the sequence reads using an in-house Perl script.

Sequences shorter than 300 nucleotides and sequences containing one or more ambiguous positions (Ns) over 300 nucleotides were removed from all datasets. Finally, redundant sequences (identical or substrings) in each set of sequences collected from the different databases were removed using a Perl script. The final sets of sequences were then submitted to a local installation of the Ribosomal Protein Project (RDP) classifier (version 2.5 training set 9) (Wang et al, 2007), and those that were assigned to the phylum *Chlamydiae* with a confidence value of at least 80% were further analyzed. The final dataset also included twelve 16S rRNA gene sequences obtained in this study from PCR products of DNA extracted from water of Ace Lake in Antarctica (see *Detection of chlamydial 16S rRNA genes in a water sample from Ace Lake* below).

Check for chimeric sequences and calculation of operational taxonomic units (OTUs)

The complete set of the chlamydial 16S rRNA gene fragments was aligned using the SINA aligner (Pruesse et al, 2012). The coverage for each position of the multiple sequence alignment was calculated using a Perl script and plotted. This analysis showed that the region with the highest coverage spanned across the variable regions V4 and V6 (Figure S1), mainly due to the large proportion of sequences

originating from VAMPS that cover this region. Two conserved positions flanking this region were selected as anchors for the alignment (*E. coli* pos 573-1045), and all sequences were trimmed around these positions, resulting in an alignment consisting of 885 columns and including around 450 nucleotides.

The filtering step using the RPD classifier should have eliminated the majority of chimeric sequences in the sequence data set as chimeric sequences generated from 16S rRNA genes of distantly related microbes would not have been assigned to the *Chlamydiae* with the high confidence threshold used. However, the trimmed dataset was also analyzed for the presence of chimeric sequences using the respective programs UCHIME and Chimera Slayer (Edgar et al, 2011; Haas et al, 2011) that are available in the MOTHUR software package (Schloss et al, 2009), but none were detected using default settings. As an additional check for chimeras and other sequencing artifacts, all sequences were analyzed using Megablast against a small local database containing 16S rRNA gene sequences from all characterized species in the phylum *Chlamydiae* (n=22). Based on the assumption that any sequence fragment representing a non-chimeric, high-quality chlamydial 16S rRNA gene sequence should align perfectly with the reference *Chlamydiae* 16S rRNA gene sequences, all sequences were trimmed at the observed Megablast alignment positions. The resulting set of sequences was again filtered for size, and any sequences shorter than 400 nucleotides were omitted; all redundant sequences that may have arisen due to refining the analysis to a shorter region, were removed.

OTUs were calculated based on the multiple sequence alignment by MOTHUR (Schloss et al, 2009) using average linkage clustering. As an alternative approach we used the software ESPRIT (Sun et al, 2009), which does not rely on multiple sequence alignments but calculates OTUs based on a similarity matrix obtained from pairwise alignments. Default settings were used for alignments, and complete linkage and average linkage were used for clustering. Data were extracted for OTUs at 97%, 95%, 90%, 85% and 80% similarity levels corresponding roughly to the species, genus, family, class, and phylum level, respectively, used for taxonomic classification of members of the *Chlamydiae* (Everett et al, 1999).

Phylogenetic analysis

All near full-length 16S rRNA sequences (> 1100 nucleotides) classified as chlamydial were checked for chimeras with UCHIME and Chimera Slayer (Edgar et al, 2011; Haas et al, 2011) available in MOTHUR (Schloss et al, 2009), and only non-chimeric sequences were used for phylogenetic analysis. Sequences were aligned with SINA (Pruesse et al, 2012), and the obtained alignment was further refined manually. The tree calculation was performed with MrBayes (version 3.2) (Huelsenbeck and Ronquist, 2001) using 10 million generations, sampling every 1 000 generations and 25% “burn in” per sampling keeping all other default options . The Newick formatted tree with the highest score was extracted and visualized with iTol (Letunic and Bork, 2007). Nodes with posterior probability values less than 50% were collapsed.

Ecological classification of sequences and OTUs

To analyze general habitat patterns, metadata describing the sample origin was kept for each individual sequence throughout the analysis. For simplification, diverse environmental origins were unified into four main categories (terrestrial, marine, freshwater, and engineered) based on their principal characteristics. For example, all samples that were not from water saturated environments like farm soil, land animal samples, or any sample collected from dry land were considered "terrestrial". Water saturated environments were divided based on salinity. Therefore marine sediments, marine water column, saline lakes, hypersaline mats, samples of marine animals and other saline environments were considered "marine", while river and lake water and their sediments, drinking water and other general water supply samples were classified as "freshwater". Samples from chemostats, waste water treatment plants and other engineered systems that couldn't be otherwise classified were considered as "engineered".

After clustering OTUs were assigned an environmental classification based on the origin of the sequences they contained. In order for an OTU to be classified in any of the four main categories a unanimous agreement of all its sequences' environmental classification was required. OTUs were classified as "mixed" if they contained sequences from different environmental categories.

Statistical analysis

Non-metric multi dimension scaling (NMDS) was used to visualize environmental origin of, and similarities between, the observed chlamydial OTUs. Based on ESPRIT average linkage OTUs, a matrix containing distances between all OTUs was constructed by calculating the average sequence dissimilarities of their members. This distance matrix was processed with metaMDS (package vegan) in R to create the NMDS matrix (default options; http://cran.r-project.org/web/packages/vegan/index.html). The NMDS matrix together plus a list containing the environmental classification as a color code and a list containing the size of each OTU, served as input parameters for visualization of similarity, diversity and ecology, using ggplot (package ggplot2) in R.

Detection of chlamydial 16S rRNA genes in a water sample from Ace Lake

16S rRNA gene targeted PCR was performed on DNA obtained from the upper aerobic layer (5m) of the water column of Ace lake (Lauro et al, 2011) using the chlamydia-specific primer SigF2 (5'CRGCGTGGATGAGGCAT) (Haider et al, 2008) and the universal primer 1492R (5'-GGYTACCTTGTTACGACTT) (Loy et al, 2005). Reaction conditions were 5 min of initial denaturation at 95°C followed by 40 cycles of 30 sec denaturation at 95°C, 30 sec annealing at 54°C and 1.5 min of elongation at 72°C. A final step of 4 min at 72°C was added for the final elongation of the incomplete products. PCR products were purified and cloned with a commercial kit following the manufacturer's instructions (TopoXL cloning Kit, Invitrogen Life Technologies). Clones were screened by restriction fragment length polymorphism using *Msp*I and *Hae*III, and 25 unique RFLP patterns were selected for sequencing. The cloned inserts were reamplified using M13 primers, and the products were sequenced on an ABI 3130 XL genetic analyzer. Newly recovered 16S rRNA gene sequences were deposited in Genbank/EMBL/DDBJ (accession numbers KC902441-KC902452).

Supplementary Figures

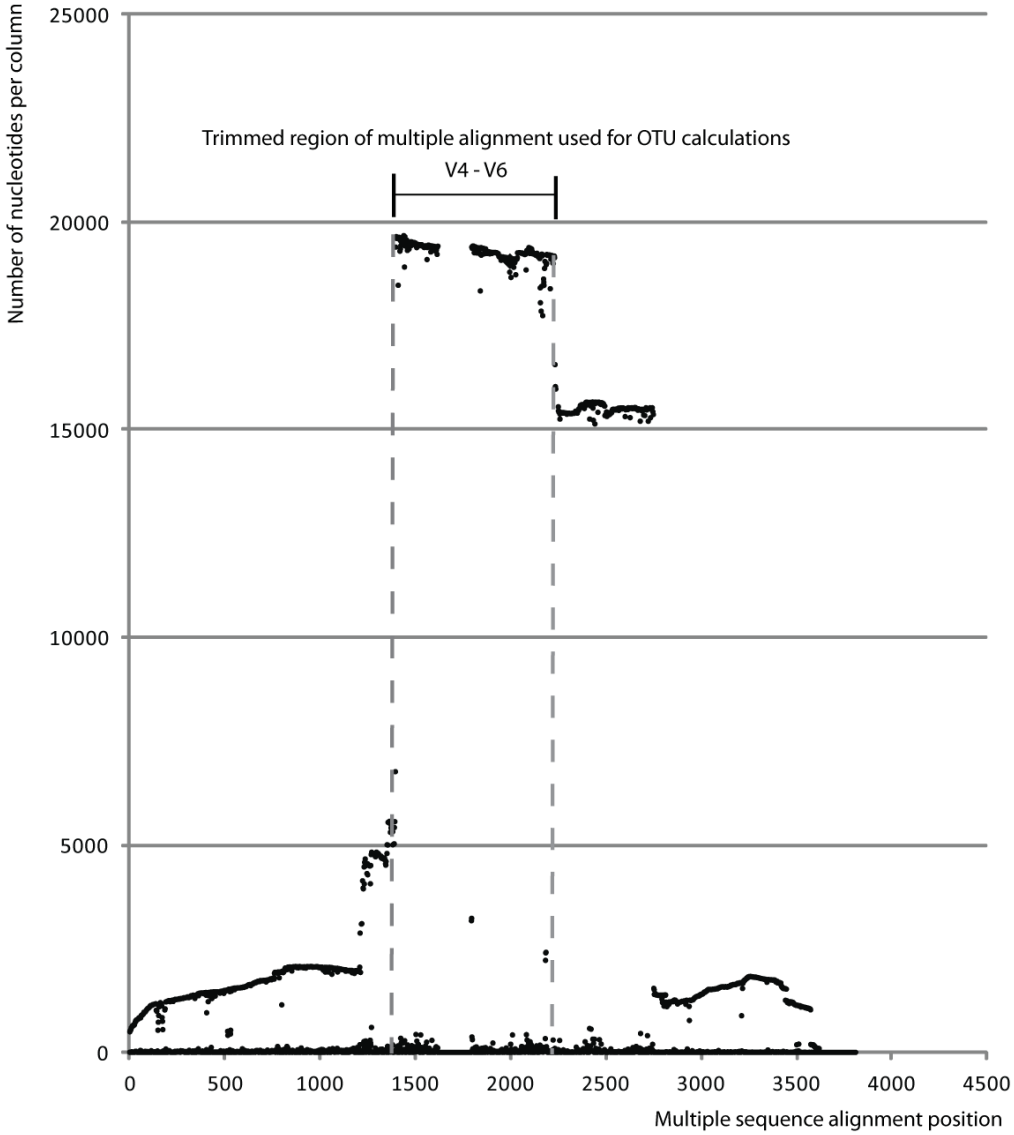

Figure S1. Number of nucleotides per alignment position in the multiple sequence alignment of all collected chlamydial 16S rRNA sequences. The region with the highest coverage that was used for OTU calculation spans through variable regions V4 to V6 and corresponds to *E. coli* positions 573-1045 (~ 450 nt).

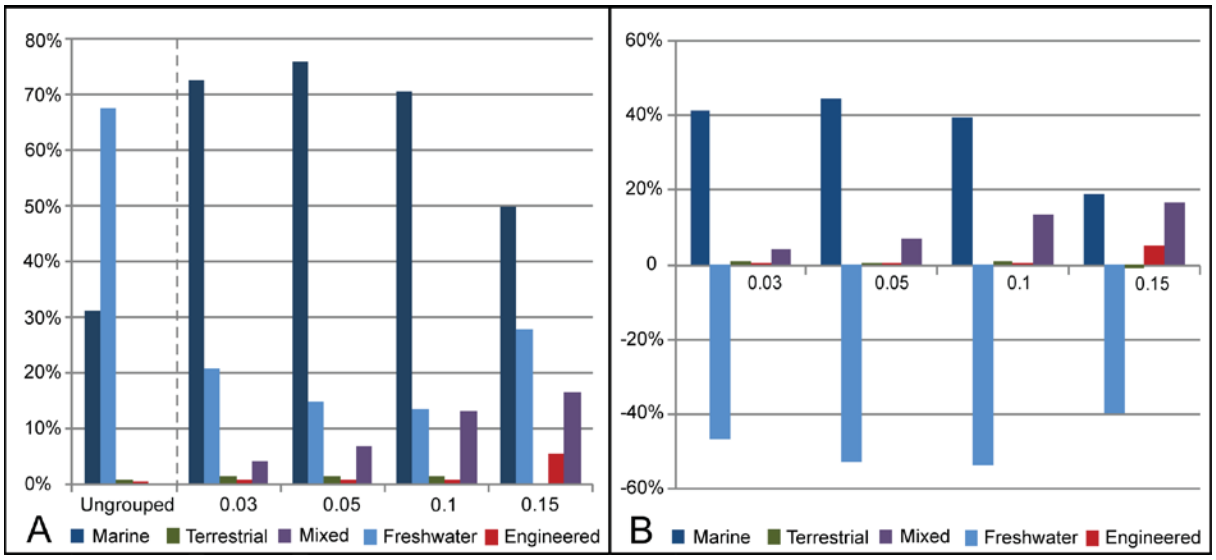

Figure S2. Ecological classification of chlamydial 16S rRNA sequences. A: Relative abundance of ungrouped sequences and OTUs based on their environmental classification. Bars represent the contribution of each environment (see color code) to the total amount of sequences or OTUs. "ungrouped" refers to the final set of sequences that was used for OTU calculation; 0.03 to 0.15 represent the different cut-off values used for OTU formation. **B:** Over/under representation of environmental categories at different OTU levels. The panel is based on the same data shown in panel A; bars indicate the difference between the relative abundance of OTUs compared to the relative abundance of (ungrouped) sequences for each environmental category. Despite the higher number of freshwater sequences (A, "ungrouped") they form fewer OTUs than the marine sequences (B), indicating a higher diversity of marine sequences.

Table S1: Metagenomic proteins that group monophyletically with known chlamydial virulence-associated proteins. Proteins were taxonomically classified at the family level using their closest neighbor in maximum likelihood trees. The described function of chlamydial homologs and the environmental origin of the metagenomic proteins are indicated. See Excel file.

Table S2: Number of unique chlamydial 16S rRNA sequences in different databases. Only sequences longer than 300 nt and classified as *Chlamydiae* by the RDP classifier (ref) with confidence higher than 80% were considered. Sequence redundancy among datasets collected from different databases was not assessed. See Excel file.

Tables S3. Clustering of full length 16S rRNA sequences and their V4-V6 region, respectively, into OTUs by two different methods. See Excel file.

Table S4. Full length 16S rRNA sequences used for phylogenetic analysis and their clustering into OTUs. See Excel file.

References

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.

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 27: 2194-2200.

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.

Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward DV, Giannoukos G *et al* (2011). Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Research* 21: 494-504.

Haider S, Collingro A, Walochnik J, Wagner M, Horn M (2008). Chlamydia-like bacteria in respiratory samples of community-acquired pneumonia patients. *Fems Microbiology Letters* 281: 198-202.

Huber JA, Mark Welch D, Morrison HG, Huse SM, Neal PR, Butterfield DA *et al* (2007). Microbial population structures in the deep marine biosphere. *Science* 318: 97-100.

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

Kodama Y, Shumway M, Leinonen R (2012). The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res* 40: D54-56.

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.

Loy A, Schulz C, Lucker S, Schopfer-Wendels A, Stoecker K, Baranyi C *et al* (2005). 16S rRNA gene-based oligonucleotide microarray for environmental monitoring of the betaproteobacterial order "Rhodocyclales". *Appl Environ Microb* 71: 1373-1386.

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.

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

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.

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 Microb* 75: 7537-7541.

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

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.

Wang Q, Garrity GM, Tiedje JM, Cole JR (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microb* 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.**

Table S1: Metagenomic proteins that group monophyletically with known chlamydial virulence-asscociated proteins.

Metagenomic read id (SIMAP)	Acc. no. of closest chlamydial neighbor	Discription of closest homolog	General description of closest homolog	Gene name	Organism name	Taxonomy	Metagenomic sample	Environmental type
129216449	YP_005663465.1	serine protease, MucD	diverse functions	htrA	Chlamydophila psittaci 6BC	Chlamydiaceae	Mediterranean Gutless Worm Metagenome.	Gutless Worm
179713205	YP_004672136.1	scc1-B gene product	chaperones of the type III secretion system	scc1	Simkania negevensis Z	Simkaniaceae	Isolation source: Guerrero Negro Hypersalir	Hypersaline Mat
3989145199	YP_004672362.1	sctD gene product	structural components of the type III secretion system	sctD, cdsD	Simkania negevensis Z	Simkaniaceae	Isolation source: Lake Washington Methane	lake
3989150060	YP_004672366.1	sctG-B gene product	chaperones of the type III secretion system	sctG, cdsG	Simkania negevensis Z	Simkaniaceae	Isolation source: Lake Washington Methane	lake
4203161593	YP_004672362.1	sctD gene product	structural components of the type III secretion system	sctD, cdsD	Simkania negevensis Z	Simkaniaceae	Lake Washington Methane enrichment	lake
4214159626	YP_004672373.1	sctQ-B gene product	structural components of the type III secretion system	sctQ, cdsQ	Simkania negevensis Z	Simkaniaceae	Oak Ridge Pristine Groundwater FRC FW30	Groundwater
421419510	YP_004671076.1	metal-dependent hydrolase, beta- lactamase su	potentially type III secreted effectors	metal dependent hydrolase CT380	Simkania negevensis Z	Simkaniaceae	Oak Ridge Pristine Groundwater FRC FW30	Groundwater
4214212582	YP_004672519.1	hypothetical protein	potentially type III secreted effectors	CT550	Simkania negevensis Z	Simkaniaceae	Oak Ridge Pristine Groundwater FRC FW30	Groundwater
421422761	YP_003708406.1	Virulence plasmid protein pGP6-D-related prote	manipulation of host cell immune response	pGP6-D	Waddlia chondrophila WSU 86-1044	Waddliaceae	Oak Ridge Pristine Groundwater FRC FW30	Groundwater
421422762	YP_003708406.1	Virulence plasmid protein pGP6-D-related prote	manipulation of host cell immune response	pGP6-D	Waddlia chondrophila WSU 86-1044	Waddliaceae	Oak Ridge Pristine Groundwater FRC FW30	Groundwater
4214314301	CCB92083.1	putative uncharacterized protein	chaperones of the type III secretion system	sctO, cdsO	Waddlia chondrophila 2032/99	Waddliaceae	Oak Ridge Pristine Groundwater FRC FW30	Groundwater
421437409	YP_008389.1	hypothetical protein	chaperones of the type III secretion system	sycE	Protochlamydia amoebophila UWE25	Parachlamydiaceae	Oak Ridge Pristine Groundwater FRC FW30	Groundwater
4214378732	YP_004672366.1	sctG-B gene product	chaperones of the type III secretion system	sctG, cdsG	Simkania negevensis Z	Simkaniaceae	Oak Ridge Pristine Groundwater FRC FW30	Groundwater
4463115550	YP_004671454.1	scc2 gene product	chaperones of the type III secretion system	scc2	Simkania negevensis Z	Simkaniaceae	Bacterioplankton: All Metagenomic 454 Rea	marine
4463796965	YP_004672363.1	sctE gene product	chaperones of the type III secretion system	sctE, cdsE	Simkania negevensis Z	Simkaniaceae	Bacterioplankton: All Metagenomic 454 Rea	marine
4467538653	YP_004672366.1	sctG-B gene product	chaperones of the type III secretion system	sctG, cdsG	Simkania negevensis Z	Simkaniaceae	BotanyBay: All Metagenomic Shotgun Reads	marine
4467634367	CAA06182.1	Omp1 protein, partial	adhesion and host cell entry	OMP	Chlamydophila abortus	Chlamydiaceae	BotanyBay: All Metagenomic Shotgun Reads	marine
4473420937	YP_004672519.1	hypothetical protein	potentially type III secreted effectors	CT550	Simkania negevensis Z	Simkaniaceae	Sapelo2008: All Metagenomic 454 Reads (N	marine
4519438385	YP_004672696.1	hypothetical protein	potentially type III secreted effectors	CT648	Simkania negevensis Z	Simkaniaceae	GutlessWorm: All Metagenomic Sequence R	Gutless Worm
4530104265	YP_004672136.1	scc1-B gene product	chaperones of the type III secretion system	scc1	Simkania negevensis Z	Simkaniaceae	AntarcticaAquatic: All Metagenomic Shotgun	marine
4530104266	YP_004672136.1	scc1-B gene product	chaperones of the type III secretion system	scc1	Simkania negevensis Z	Simkaniaceae	AntarcticaAquatic: All Metagenomic Shotgun	marine
4530104267	YP_007745.1	scc1 gene product	chaperones of the type III secretion system	scc1	Protochlamydia amoebophila UWE25	Parachlamydiaceae	AntarcticaAquatic: All Metagenomic Shotgun	marine
4530109133	YP_004670658.1	sctR gene product	structural components of the type III secretion system	sctR, cdsR	Simkania negevensis Z	Simkaniaceae	AntarcticaAquatic: All Metagenomic Shotgun	marine
4530140532	YP_003709171.1	rRNA methylase	effectors secreted by the type III secretion system	rRNA methylase	Waddlia chondrophila WSU 86-1044	Waddliaceae	AntarcticaAquatic: All Metagenomic Shotgun	marine
4530156080	ZP_06299396.1	putative SctF chaperone SctG	chaperones of the type III secretion system	sctG, cdsG	Parachlamydia acanthamoebae str. Hall'	Parachlamydiaceae	AntarcticaAquatic: All Metagenomic Shotgun	marine
4530156081	YP_004672366.1	sctG-B gene product	chaperones of the type III secretion system	sctG, cdsG	Simkania negevensis Z	Simkaniaceae	AntarcticaAquatic: All Metagenomic Shotgun	marine
4530166648	NP_828912.1	sctQ gene product	structural components of the type III secretion system	sctQ, cdsQ	Chlamydophila caviae GPIC	Chlamydiaceae	AntarcticaAquatic: All Metagenomic Shotgun	marine
4530166651	YP_515884.1	type III secretion system protein	structural components of the type III secretion system	sctQ, cdsQ	Chlamydophila felis Fe/C-56	Chlamydiaceae	AntarcticaAquatic: All Metagenomic Shotgun	marine
4530195807	YP_004670659.1	sctL gene product	structural components of the type III secretion system	sctL, cdsL	Simkania negevensis Z	Simkaniaceae	AntarcticaAquatic: All Metagenomic Shotgun	marine
4530217607	YP_004662993.1	virulence plasmid protein pGP6-D	manipulation of host cell immune response	pGP6-D	Simkania negevensis Z	Simkaniaceae	AntarcticaAquatic: All Metagenomic Shotgun	marine
4530217608	NP_038282.1	virulence protein pGP6-D	manipulation of host cell immune response	pGP6-D	Chlamydia muridarum Nigg	Chlamydiaceae	AntarcticaAquatic: All Metagenomic Shotgun	marine
4530217609	YP_009012.1	hypothetical protein	manipulation of host cell immune response	pGP6-D	Protochlamydia amoebophila UWE25	Parachlamydiaceae	AntarcticaAquatic: All Metagenomic Shotgun	marine
4530217610	YP_003708406.1	Virulence protein pGP6-D-related protein	manipulation of host cell immune response	pGP6-D	Waddlia chondrophila WSU 86-1044	Waddliaceae	AntarcticaAquatic: All Metagenomic Shotgun	marine
4530217612	NP_038282.1	virulence protein pGP6-D	manipulation of host cell immune response	pGP6-D	Chlamydia muridarum Nigg	Chlamydiaceae	AntarcticaAquatic: All Metagenomic Shotgun	marine
4530272674	YP_004672519.1	hypothetical protein	potentially type III secreted effectors	CT550	Simkania negevensis Z	Simkaniaceae	AntarcticaAquatic: All Metagenomic Shotgun	marine
4530291354	YP_004670660.1	hypothetical protein	effectors secreted by the type III secretion system	CT560	Simkania negevensis Z	Simkaniaceae	AntarcticaAquatic: All Metagenomic Shotgun	marine
4530291355	YP_004650948.1	hypothetical protein	effectors secreted by the type III secretion system	CT560	Parachlamydia acanthamoebae UV-7	Parachlamydiaceae	AntarcticaAquatic: All Metagenomic Shotgun	marine
4530350342	YP_004650955.1	hypothetical protein	potentially type III secreted effectors	Multicopper polyphenol oxidase CT	Parachlamydia acanthamoebae UV-7	Parachlamydiaceae	AntarcticaAquatic: All Metagenomic Shotgun	marine
4530350358	YP_008867.1	hypothetical protein	potentially type III secreted effectors	Multicopper polyphenol oxidase CT	Protochlamydia amoebophila UWE25	Parachlamydiaceae	AntarcticaAquatic: All Metagenomic Shotgun	marine
4530350416	YP_004670624.1	hypothetical protein	potentially type III secreted effectors	Multicopper polyphenol oxidase CT	Simkania negevensis Z	Simkaniaceae	AntarcticaAquatic: All Metagenomic Shotgun	marine
4530406099	YP_003709968.1	hypothetical protein wcw_1617	chaperones of the type III secretion system	sctO, cdsO	Waddlia chondrophila WSU 86-1044	Waddliaceae	AntarcticaAquatic: All Metagenomic Shotgun	marine
4530406100	YP_004377045.1	hypothetical protein G5S_0336	chaperones of the type III secretion system	sctO, cdsO	Chlamydophila pecorum E58	Chlamydiaceae	AntarcticaAquatic: All Metagenomic Shotgun	marine
4530481984	YP_007382.1	mip gene product	effectors secreted by the type III secretion system	mip	Protochlamydia amoebophila UWE25	Parachlamydiaceae	AntarcticaAquatic: All Metagenomic Shotgun	marine
453052427	YP_003709598.1	metal-dependent hydrolase, beta- lactamase su	potentially type III secreted effectors	metal dependent hydrolase CT380	Waddlia chondrophila WSU 86-1044	Waddliaceae	AntarcticaAquatic: All Metagenomic Shotgun	marine
4530623670	NP_829854.1	hypothetical protein	potentially type III secreted effectors	CT648	Chlamydophila caviae GPIC	Chlamydiaceae	AntarcticaAquatic: All Metagenomic Shotgun	marine
4530628466	YP_007382.1	mip gene product	effectors secreted by the type III secretion system	mip	Protochlamydia amoebophila UWE25	Parachlamydiaceae	AntarcticaAquatic: All Metagenomic Shotgun	marine
4530773846	YP_008400.1	pkn gene product	effectors secreted by the type III secretion system	pkn5	Protochlamydia amoebophila UWE25	Parachlamydiaceae	AntarcticaAquatic: All Metagenomic Shotgun	marine
4530932289	YP_003709994.1	FKBP-type peptidyl-prolyl cis-trans isomerase	effectors secreted by the type III secretion system	mip	Waddlia chondrophila WSU 86-1044	Waddliaceae	AntarcticaAquatic: All Metagenomic Shotgun	marine
453094726	YP_004376934.1	YopT tranlocation T	structural components of the type III secretion system	sctT, cdsT	Chlamydophila pecorum E58	Chlamydiaceae	AntarcticaAquatic: All Metagenomic Shotgun	marine
4543303730	YP_004672363.1	sctE gene product	chaperones of the type III secretion system	sctE, cdsE	Simkania negevensis Z	Simkaniaceae	WesternChannelOMM: All Metagenomic 454	marine
4543306273	YP_004672519.1	hypothetical protein	potentially type III secreted effectors	CT550	Simkania negevensis Z	Simkaniaceae	WesternChannelOMM: All Metagenomic 454	marine
EBA77306.1	YP_007915.1	protease-like activity factor	manipulation of host cell immune response	CPAF	Protochlamydia amoebophila UWE25	Parachlamydiaceae	GOS: All 454	marine
EBE43690.1	YP_004672132.1	sctU gene product	structural components of the type III secretion system	sctU, cdsU	Simkania negevensis Z	Simkaniaceae	GOS: All 454	marine
EBF12630.1	YP_001654130.1	hypothetical protein CTL0039	chaperones of the type III secretion system	sctO, cdsO	Chlamydia trachomatis 434/Bu	Chlamydiaceae	GOS: All 454	marine
EBM43873.1	ZP_06298342.1	hypothetical protein pah_c004o195	effectors secreted by the type III secretion system	rRNA methylase	Parachlamydia acanthamoebae str. Hall'	Parachlamydiaceae	GOS: All 454	marine
EBN20060.1	YP_005661967.1	hypothetical protein	potentially type III secreted effectors	Multicopper polyphenol oxidase CT	Chlamydophila pneumoniae LPCoLN	Chlamydiaceae	GOS: All 454	marine
EBO34119.1	YP_004671423.1	sctC-A gene product	structural components of the type III secretion system	sctC, cdsC	Simkania negevensis Z	Simkaniaceae	GOS: All 454	marine
EBO61181.1	YP_004672360.1	sycE gene product	chaperones of the type III secretion system	sycE	Simkania negevensis Z	Simkaniaceae	GOS: All 454	marine
EBY98170.1	YP_004670658.1	sctR gene product	structural components of the type III secretion system	sctR, cdsR	Simkania negevensis Z	Simkaniaceae	GOS: All 454	marine
EBY98171.1	YP_004670656.1	sctT gene product	structural components of the type III secretion system	sctT, cdsT	Simkania negevensis Z	Simkaniaceae	GOS: All 454	marine
ECD83630.1	YP_004672366.1	sctG-B gene product	chaperones of the type III secretion system	sctG, cdsG	Simkania negevensis Z	Simkaniaceae	GOS: All 454	marine
ECG22147.1	YP_004670417.1	pGP6-D gene product	manipulation of host cell immune response	pGP6-D	Simkania negevensis Z	Simkaniaceae	GOS: All 454	marine
ECN06433.1	YP_004670661.1	sctJ gene product	structural components of the type III secretion system	sctJ, cdsJ	Simkania negevensis Z	Simkaniaceae	GOS: All 454	marine
ECO84253.1	YP_004670660.1	hypothetical protein	effectors secreted by the type III secretion system	CT560	Simkania negevensis Z	Simkaniaceae	GOS: All 454	marine
ECO90885.1	YP_004672366.1	sctG-B gene product	chaperones of the type III secretion system	sctG, cdsG	Simkania negevensis Z	Simkaniaceae	GOS: All 454	marine
ECR93226.1	YP_004670417.1	pGP6-D gene product	manipulation of host cell immune response	pGP6-D	Simkania negevensis Z	Simkaniaceae	GOS: All 454	marine
ECT32411.1	YP_004670656.1	sctT gene product	structural components of the type III secretion system	sctT, cdsT	Simkania negevensis Z	Simkaniaceae	GOS: All 454	marine
ECT59526.1	ZP_06300478.1	putative type III secreted protein SctW	effectors secreted by the type III secretion system	sctW, copN	Parachlamydia acanthamoebae str. Hall'	Parachlamydiaceae	GOS: All 454	marine
ECT61183.1	YP_004672373.1	sctQ-B gene product	structural components of the type III secretion system	sctQ, cdsQ	Simkania negevensis Z	Simkaniaceae	GOS: All 454	marine
EDF40215.1	YP_004672557.1	prc gene product	manipulation of host cell immune response	tsp	Simkania negevensis Z	Simkaniaceae	GOS: All 454	marine
EDG27191.1	YP_004672373.1	sctQ-B gene product	structural components of the type III secretion system	sctQ, cdsQ	Simkania negevensis Z	Simkaniaceae	GOS: All 454	marine
EDG58047.1	YP_004670699.1	tsp gene product	manipulation of host cell immune response	tsp	Simkania negevensis Z	Simkaniaceae	GOS: All 454	marine
EDJ19034.1	YP_004651821.1	omcB gene product	manipulation of host cell immune response	omcB	Parachlamydia acanthamoebae UV-7	Parachlamydiaceae	GOS: All 454	marine
EDJ42278.1	YP_004672135.1	sctW gene product	effectors secreted by the type III secretion system	sctW, copN	Simkania negevensis Z	Simkaniaceae	GOS: All 454	marine
EDJ42279.1	YP_004672136.1	scc1-B gene product	chaperones of the type III secretion system	scc1	Simkania negevensis Z	Simkaniaceae	GOS: All 454	marine
EDJ57261.1	YP_004670699.1	tsp gene product	manipulation of host cell immune response	tsp	Simkania negevensis Z	Simkaniaceae	GOS: All 454	marine

Table S2: Number of unique chlamydial 16S rRNA sequences in different databases

	No. of sequences	Average seq. length	Max seq. length
NCBI nr	519	935	1538
NCBI env	368	728	1517
NCBI est	3	604	632
NCBI GSS	2	861	866
NCBI WGS	2	1014	1308
IMG/m	13	674	1527
CAMERA sanger	25	991	1145
CAMERA 454	243	305	555
VAMPS	14340	481	517
SRA	6543	399	827
Ace Lake clones	12	1473	1491
Total	**22070**	**471**	

Tables S3. Clustering of full length 16S rRNA sequences and their V4-V6 region, respectively, into OTUs by two different meth

		>1100 bp (271 seq)		>400 bp no filter (217 seq)		>400 bp filtered (210 seq)	
Tax. Order	**Cutoff**	ESPRIT	MOTHUR	ESPRIT	MOTHUR	ESPRIT	MOTHUR
Species	0.03	108	112	107	108	102	104
Genera	0.05	82	90	85	86	79	80
Families	**0.1**	**28**	**40**	**38**	**43**	**27**	**31**
Classes	0.15	3	8	9	12	4	8
Phyla	0.2	1	1	1	2	1	1

Table S4. Full length 16S rRNA sequences used for phylogenetic analysis and their clustering into OTUs

OTU number/ accession number	Family name/ sequence name	Environmental origin, host
Group_1		**Marine**
FR714417.1	clone IPI 854-950-1046	marine derived lake, Lago di Paola, Italy
Group_2		**Terrestrial**
AY082465.1	clone 44a-B1-34	subsurface acid mine drainage system
Group_3		**Marine**
FR714409.1	clone IPI 817-913-1009	marine derived lake, Lago di Paola, Italy
Group_4		
NR_026266.1	Verrucomicrobium spinosum strain DSM 4136	
Group_5		**Freshwater**
DQ444409.1	clone LT100PlH9	tropical freshwater lake Tanganyika
Group_6		**Marine**
JN502883.1	clone SBZA 4032	Guerrero Negro hypersaline microbial mat
Group_7		**Marine**
JN530146.1	clone SBZO 1546	Guerrero Negro hypersaline microbial mat
Group_8		**Marine**
JN536584.1	clone SBZP 2308	Guerrero Negro hypersaline microbial mat
Group_9		**Marine**
EU491566.1	clone EPR4059-B2-Bc66	seafloor lavas, Pacific Ocean
Group_10		**Marine**
DQ903996.1	clone PRPR83	marine sponge, Hawaii ,USA
Group_11		**Terrestrial**
JN607058.1	clone GAS254O1bO5	lava tube, Azores, Portugal
Group_12		
AY390429.1	Lentisphaera araneosa	
Group_13		
EF012750.1	Rhodopirellula baltica strain OJF6	
Group_14		**Marine**
EU050949.1	clone SS1 B 01 04	arctic sediment from the Kings Bay, Svalbard
Group_15	**Piscichlamydiaceae**	**Mixed**
AY462244.1	Piscichlamydia salmonis clone C093-1	Sea farmed Atlantic salmon, Norway
JQ065096.1	Piscichlamydia salmonis clone P28	Sea farmed Atlantic salmon, Norway
EU326495.1	Piscichlamydia salmonis clone Pch-D261006	Sea farmed Atlantic salmon, Norway
EF153480.1	Piscichlamydia salmonis clone BR281005-25	Fresh water salmon, Norway
Group_16	**PCF1**	**Marine**
FR714401.1	clone IPI 769-865-961	marine derived lake, Lago di Paola, Italy
FR714402.1	clone IPI 772-868-964	marine derived lake, Lago di Paola, Italy
Group_17	**PCF2**	**Marine**
FR714416.1	clone IPI 853-949-1045	marine derived lake, Lago di Paola, Italy
FR714404.1	clone IPI 779-875-971	marine derived lake, Lago di Paola, Italy
FR714420.1	clone IPI 884-980	marine derived lake, Lago di Paola, Italy
FR714410.1	clone IPI 821-9017	marine derived lake, Lago di Paola, Italy
FR714408.1	clone IPI 807-903-999	marine derived lake, Lago di Paola, Italy
FR714415.1	clone IPI 844-940-1036	marine derived lake, Lago di Paola, Italy
FR714414.1	clone IPI 842-938-1034	marine derived lake, Lago di Paola, Italy
FR714407.1	clone IPI 800-896-992	marine derived lake, Lago di Paola, Italy
FR714418.1	clone IPI 862-958-1054	marine derived lake, Lago di Paola, Italy
Group_18	**PCF3 (ECLI)**	**Engineered**
AF364575.1	clone P-9	Waste water treatment plant, Germany
AB504658.1	clone K29C2-29	methane oxidizing DHS reactor
AF364569.1	clone P-7	Waste water treatment plant, Germany

AF364564.1	clone P-4	Waste water treatment plant, Germany
Group_19	**PCF4**	**Marine**
FR714406.1	clone IPI 796-892	marine derived lake, Lago di Paola, Italy
FR714405.1	clone IPI 789-885-981	marine derived lake, Lago di Paola, Italy
FR714413.1	clone IPI 831-927-1023	marine derived lake, Lago di Paola, Italy
FR714419.1	clone IPI 864-960-1056	marine derived lake, Lago di Paola, Italy
FR714403.1	clone IPI 778-874-970	marine derived lake, Lago di Paola, Italy
FR714412.1	clone IPI 830-926-1022	marine derived lake, Lago di Paola, Italy
Group_20	**PCF5**	**Mixed**
JF706724.1	clone cvE71	fresh water systems, France
EU488135.1	clone CK 2C2 23	sediment from Thalassia sea grass bed, Florida, USA
Group_21	**Parilichlamydiaceae**	**Mixed**
JQ480303.1	clone ON26	gills of Oreochromis niloticus (Nile tilapia), Lake Victoria, Uganda
JQ480299.1	clone CF3	gills of Clarias gariepinus (African catfish), Lake Victoria, Uganda
JQ673516.1	Parilichlamydia carangidicola clone 25YTK11	gills of Seriola lalandi (Yellowtail Kingfish), Australia
JQ480302.1	clone ON3	gills of Oreochromis niloticus (Nile tilapia), Lake Victoria, Uganda
Group_22	**Waddliaceae**	**Mixed**
EU090708.1	clone CN761	clinical respiratory samples, Germany
AF346001.1	Waddlia chondrophila strain 2032/99	septic stillborn calf, Germany
NR_074886.1	Waddlia chondrophila strain WSU 86-1044	aborted bovine foetus
JF706723.1	clone cvE65	fresh water systems, France
FJ479189.1	clone p22m06ok	undisturbed tall grass prairie, Oklahoma, USA
NR_028697.1	Waddlia chondrophila WSU 86-1044	aborted bovine foetus
AY184804.1	Waddlia malaysiensis strain G817	urine of Eonycteris spelaea (fruit bat)
Group_23	**Simkaniaceae**	**Mixed**
AY140911.1	Fritschea eriococci strain Elm	Eriococcus spurious (scale insect), California, USA
AF400484.1	Fritschea bemisiae strain Jatropha	Bemisia tabaci biotype Jatropha, Arizona, USA
FJ976094.1	clone cvE38	fresh water systems, France
FJ976095.1	clone cvE419	fresh water systems, France
AF448723.3	clone cvE9	fresh water systems, France
NR_074932.1	Simkania negevensis strain Z	
JQ009299.1	Fritschea bemisiae strain Elm	Bemisia tabaci biotype Q, GuangZhou, China
NR_029194.1	Simkania negevensis strain Z	
EF177461.1	clone UUZM	Xenoturbella westbladi UUZM 57848, Cost of Sweden
EU326493.1	clone D261006	gills of Salmo salar (Atlantic salmon), Norway
JN606076.1	clone NS16	human nasal sample
Group_24	**PCF6**	**Marine**
KC902441.1	clone AAL4c3	marine derived lake, Ace, Antarctica
KC902442.1	clone AAL4c6	marine derived lake, Ace, Antarctica
KC902443.1	clone AAL4c8	marine derived lake, Ace, Antarctica
KC902444.1	clone AAL4c11	marine derived lake, Ace, Antarctica
KC902445.1	clone AAL4c23	marine derived lake, Ace, Antarctica
KC902446.1	clone AAL4c19	marine derived lake, Ace, Antarctica
KC902447.1	clone AAL4c20	marine derived lake, Ace, Antarctica
Group_25	**Chlamydiaceae**	**Terrestrial**
DQ019308.1	Chlamydia trachomatis strain L1/440	
NR_027576.1	Chlamydophila pecorum strain E58	
NR_036835.1	Chlamydia muridarum strain MoPn/Wiess-Nigg	
NR_036876.1	Chlamydophila felis strain Fe/Pn-1	
AB001774.1	Chlamydophila pecorum strain B0-1485	
AB001775.1	Chlamydophila pecorum strain B0-Maeda	
AB001776.1	Chlamydophila pecorum strain B0-Yokohama	
AB001777.1	Chlamydia pecorum strain SPV789	
AB001779.1	Chlamydophila psittaci strain Bud-1	
AB001784.1	Chlamydophila psittaci strain Cal-10	
AB001787.1	Chlamydia psittaci strain Itoh	
AB001791.1	Chlamydophila psittaci strain Ohmiya	
AB001795.1	Chlamydophila psittaci strain P1307	
AB001801.1	Chlamydophila psittaci strain P1888	
AB001802.1	Chlamydophila psittaci strain PgAu46	
AB001804.1	Chlamydia psittaci strain PCM27	
AB001806.1	Chlamydophila psittaci strain PCM44	
AB001813.1	Chlamydophila psittaci strain Sugawara	
AB001815.1	Chlamydophila psittaci strain T4	
AB285329.1	Chlamydophila psittaci strain CPX0308	fecal sample from an oriental white stork

AY661794.1	Chlamydia suis strain 13VII	piglet colon
AY661797.1	Chlamydia suis strain 32XII	piglet colon
D85702.1	Chlamydophila felis strain Fe/145	Felis catus (cat)
D85703.1	Chlamydophila felis strain Fe/B166	Felis catus (cat)
D85704.1	Chlamydophila felis strain Fe/C164	Felis catus (cat)
D85705.1	Chlamydophila felis strain Fe/C454	Felis catus (cat)
D85706.1	Chlamydophila felis strain Fe/Cello	Felis catus (cat)
D85707.1	Chlamydophila felis strain Fe/C429	Felis catus (cat)
D85708.1	Chlamydophila caviae strain Gp/Ic	Cavia porcellus (guinea pig)
D85709.1	Chlamydophila abortus Ov/B577	Ovis aries (sheep)
D85711.1	Chlamydophila psittaci strain Hu/Borg	Homo sapiens (human)
D85712.1	Chlamydophila psittaci strain Frt-Hu/Ca110	Homo sapiens (human) & Mustela putorius (ferret)
D85713.1	Chlamydophila psittaci strain Prt/GCP-1	parot
D85714.1	Chlamydophila pecorum strain Bo/Shizuoka	Bos taurus (cattle)
D85715.1	Chlamydophila pecorum strain Bo/Maeda	Bos taurus (cattle)
D85716.1	Chlamydophila pecorum strain Ov/IPA	Ovis aries (sheep)
D85717.1	Chlamydophila pecorum strain Koala type II	Phascolarctos cinereus (koala)
D85719.1	Chlamydia trachomatis strain B/TW-5/OT	Homo sapiens (human)
D85720.1	Chlamydia trachomatis strain C/TW-3/OT	Homo sapiens (human)
D85722.1	Chlamydia trachomatis strain E/UW-5/Cx	Homo sapiens (human)
DQ019291.1	Chlamydia trachomatis strain A/Har-1	Homo sapiens (human)
DQ019293.1	Chlamydia trachomatis strain B/Tunis-864	Homo sapiens (human)
DQ019299.1	Chlamydia trachomatis strain D/IC-CAL8	Homo sapiens (human)
DQ019310.1	Chlamydia trachomatis strain L4/404	Homo sapiens (human)
DQ444323.1	Chlamydophila pneumoniae strain WBB	Perameles bougainville (marl)
EF486854.1	Chlamydophila abortus strain FAG	Capra hircus (goat)
EF486857.1	Chlamydophila abortus strain POS	Ovis aries (sheep)
U73782.1	Chlamydia pecorum strain BE	Phascolarctos cinereus (koala)
U73783.1	Chlamydia pneumoniae strain P1	Homo sapiens (human)
L06108.1	Chlamydia pneumoniae strain TW183	Homo sapiens (human)
GQ398026.1	strain 08-1274 Flock3	Gallus gallus (chicken)
GQ398030.1	strain 08-1274 Flock22	Gallus gallus (chicken)
NR_029196.1	Chlamydia suis strain S45	Sus scrofa (pig)
GU068510.1	clone 122	Larus glaucescens (gall)
AY334528.1	Chlamydophila psittaci clone cvCps4	human bronchial aspirate
AY334530.1	Chlamydophila psittaci clone cvCps2	human sputum
AY334532.1	Chlamydophila felis clone cvCfe1	human bronchoalveolar lavage
AY334533.1	Chlamydophila felis clone cvCfe2	human bronchial aspirate
AY334534.1	Chlamydophila felis clone cvCfe3	human nasal wash
JF756077.1	clone 09-489/LP23	Columba livia (pigeon)
HQ662953.1	Chlamydophila psittaci strain 10-1398/28	Threskiornis aethiopicus (Sacred Ibis)
HQ662955.1	clone 10-1398/6	Threskiornis aethiopicus (Sacred Ibis)
JN426966.1	Chlamydophila psittaci strain HB1043	Sus scrofa (pig)
JN606072.1	Chlamydia psittaci strain NS7	human nasal sample
JN606073.1	Chlamydophila felis strain NS9	human nasal sample
JN392919.1	Amphibiichlamydia salamandrae strain AMCS11/1	Neurergus crocatus (salamander)
JN392920.1	Amphibiichlamydia salamandrae strain AMCS11/2	Neurergus crocatus (salamander)
JN402380.1	Amphibiichlamydia ranarum strain AMCS11/3	Lithobates catesbeianus (bullfrog)
HE660094.1	clone 10-1957/2	Gallus gallus (chicken)
NR_074946.1	Chlamydophila caviae strain GPIC	
NR_074947.1	Chlamydophila felis strain Fe/C-56	
U68420.2	Chlamydia suis strain R22	Sus scrofa (pig)
AY661795.1	Chlamydia suis strain 14V	piglet colon
AY661796.1	Chlamydia suis strain 14VII	piglet colon
U68426.2	Chlamydophila pneumoniae strain N16	Equus caballus (horse)
Group_26	**PCF7**	**Mixed**
FJ976107.1	clone cvE60	fresh water systems, France
AY114316.1	clone LD1-PA25	anoxic marine sediment, UK
AY114327.1	clone LD1-PA42	anoxic marine sediment, UK
GQ850567.1	clone d98	bottom water in the northern Bering Sea, China
Group_27	**Clavichlamydiaceae**	**Mixed**
AF364568.1	clone P-6	Waste water treatment plant, Germany
EF577392.1	Clavochlamydia salmonicola	gills of Salmo trutta (brown trout), Norway
DQ011662.1	Clavochlamydia salmonicola strain CH301104	gills of Salmo salar (Atlantic salmon), Norway
JN123362.1	Clavochlamydia salmonicola isolate Br25	gills of Salmo trutta (brown trout), Norway
Group_28	**PCF8**	**Mixed**
FJ976097.1	clone cvE21	fresh water systems, France
AF448722.3	clone cvE6	fresh water systems, France
DQ903997.1	clone PRPR85	marine sponge, Hawaii ,USA

EU363464.1	clone CRIB 32	water network biofilm, Spain
FJ976096.1	clone cvE16	fresh water systems, France
FJ976098.1	clone cvE18	fresh water systems, France
HM063023.1	clone KK135A0008	lava tube, Hawaii, USA
HM444977.1	clone EP912A0005	lava tube, Hawaii, USA
HM444986.1	clone EP912A0051	lava tube, Hawaii, USA
FR714411.1	clone IPI 825-921-1017	marine derived lake, Lago di Paola, Italy
JN701140.1	clone MD2O4O1hO5	lava tube, Azores ,Portugal
JN606074.1	clone NS11e	human nasal sample
JN606075.1	clone NS13	human nasal sample
JN615791.1	clone GTM2313b10	lava tube, Azores, Portugal
JN616122.1	clone GP278O7gO4	lava tube, Azores, Portugal
JN616169.1	clone GP278O8gO8	lava tube, Azores, Portugal
JQ675408.1	clone CC01f45b05	cave water, New mexico, USA
JX279901.1	clone W-Pla-28	wastewater, China
JX317585.1	clone 49-m13 f.1.ab1	freshwater aquarium, USA
DQ444442.1	clone LT110PlE2	tropical freshwater lake Tanganyika
Group_29	**Rhabdochlamydiaceae**	**Mixed**
JN051145.1	isolate NS3	human nasal sample
AF364560.1	clone P-1	Waste water treatment plant, Germany
AF364561.1	clone P-2	Waste water treatment plant, Germany
AF364562.1	clone P-13	Waste water treatment plant, Germany
AF364566.1	clone P-14	Waste water treatment plant, Germany
AF364567.1	clone P-15	Waste water treatment plant, Germany
AF364570.1	clone P-16	Waste water treatment plant, Germany
AF364571.1	clone P-17	Waste water treatment plant, Germany
AF364572.1	clone P-18	Waste water treatment plant, Germany
AF364573.1	clone P-19	Waste water treatment plant, Germany
AF364574.1	clone P-8	Waste water treatment plant, Germany
AF364576.1	clone P-10	Waste water treatment plant, Germany
AF364577.1	clone P-11	Waste water treatment plant, Germany
AF364578.1	clone P-12	Waste water treatment plant, Germany
AY223862.1	Rhabdochlamydia porcellionis	Porcellio scaber (rough woodlouse)
AY928092.1	Rhabdochlamydia crassificans	Blatta orientalis (oriental cockroach)
DQ903988.1	clone PRPR10	marine sponge, Hawaii ,USA
EU090707.1	clone CN554	human respiratory samples , Germany
EU090709.1	clone CN808	human respiratory samples , Germany
EU133918.1	clone FFCH17845	soil, Oklahoma, USA
EU363465.1	clone CRIB 34	river water, Spain
EU683887.1	clone CRIB33	Waste water treatment plant, Spain
FJ976099.1	clone cvE58	fresh water systems, France
FJ976100.1	clone cvE55	fresh water systems, France
EF445478.1	clone KF-9	Soil, Himalaya, India
GQ287585.1	clone P1s-222	Soil, Himalaya, India
HM445488.1	clone GP27685gO2	lava tube, Azores, Portugal
JF513056.1	clone cvE88	fresh water systems, France
JF513057.1	clone cvE99	fresh water systems, France
JN167597.1	Renichlamydia lutjani strain ELO	Lutjanus kasmira (bluestripe snapper)
JN616113.1	clone GP278O7eO9	lava tube, Azores, Portugal
JN616229.1	clone GP71172e05	lava tube, Azores, Portugal
JN850423.1	clone GP27685cO3	lava tube, Azores, Portugal
Group_30	**Parachlamydiaceae/Criblamydiaceae**	**Mixed**
AF083614.1	Parachlamydia acanthamoebae strain UWE1	amoeba symbiont
AB359005.1	clone AnDHS-P22	annamox reactor
AB506677.1	isolate S13	amoeba symbiont
AB506678.1	isolate S40	amoeba symbiont
AB506679.1	isolate R18	amoeba symbiont
AF083615.1	Protochlamydia amoebophila strain UWE25	amoeba symbiont
AF083616.1	isolate UWC22	amoeba symbiont
AF098330.1	isolate TUME1	amoeba symbiont
AF177275.1	Neochlamydia hartmannellae strain A1Hsp	amoeba symbiont
AF308693.1	clone corvenA4	human corneal sample
AF364563.1	clone P-3	Waste water treatment plant, Germany
AF364565.1	clone P-5	Waste water treatment plant, Germany
AF366365.1	Parachlamydia acanthamoebae strain Hall's coccus	amoeba symbiont
AF478463.2	clone cvC7	human nasal sample
AF478473.2	clone cvC15	human sputum
AJ715410.1	Parachlamydia acanthamoebae strain UV-7	amoeba symbiont
AM408788.1	isolate EI1	amoeba symbiont
AM408789.1	isolate EI2	amoeba symbiont

AM408793.1	isolate EI6	amoeba symbiont
AM941720.1	isolate Berg17	amoeba symbiont
AY220545.2	clone cvE5	freshwater pond, Italy
AY326517.1	clone 3-1	soil, Amazon
AY326519.1	clone 530-2	soil, Amazon
DQ124300.1	Criblamydia sequanensis strain Seine	river water, France
DQ309029.1	Parachlamydia acanthamoebae strain Seine	river water, France
DQ632609.1	Protochlamydia naegleriophila strain KNic	amoeba symbiont
EU074225.1	Estrella lausannensis strain CRIB 30	river water, Spain
EU090706.1	clone CN823	human respiratory samples , Germany
EU363463.1	clone CRIB 31	river water, Spain
EU384664.1	Protochlamydia naegleriophila strain CRIB 36	Waste water treatment plant, Spain
EU683885.1	clone CRIB37	Waste water treatment plant, Spain
EU683886.1	clone CRIB38	river water, Spain
EU707854.1	Protochlamydia naegleriophila strain CRIB35	Waste water treatment plant, Spain
FJ529996.1	clone MABRDTU43	nitrifying biofilm reactor
FJ532290.1	clone CRIB44	river water, Spain
FJ532291.1	clone CRIB43	river water, Spain
FJ532292.1	clone CRIB39	Waste water treatment plant, Spain
FJ532293.1	clone CRIB40	Waste water treatment plant, Spain
FJ532294.1	clone CRIB41	Waste water treatment plant, Spain
FJ532295.1	Protochlamydia naegleriophila strain CRIB42	Waste water treatment plant, Spain
FJ976092.1	clone cvE12	river water, France
FJ976093.1	clone cvE14	fresh water, France
FJ976101.1	Protochlamydia naegleriophila strain cvE27	fresh water, France
FJ976104.1	clone cvE22	soil, France
FJ976105.1	Parachlamydia acanthamoebae strain cvE20	fresh water, France
FJ415740.1	clone SOY123	soil, China
NR_026357.1	Parachlamydia acanthamoebae strain Bn9	amoeba symbiont
AB504586.1	clone K26G1-12	methane oxidizing DHS reactor
AB504644.1	clone K29C2-11	methane oxidizing DHS reactor
GQ221847.1	Metachlamydia lacustris strain CHSL	amoeba symbiont
JF706725.1	clone cvE70	fresh water, France
JN051144.1	Parachlamydia acanthamoebae strain NS2	human respiratory samples , France
JN093034.1	clone cvE4b	fresh water, France
JN112799.1	Mesochlamydia elodeae strain KV	amoeba symbiont
JN478120.1	clone SBYT 1825	Guerrero Negro hypersaline microbial mat
JN489825.1	clone SBYX 4984	Guerrero Negro hypersaline microbial mat
JN538005.1	clone SBZP 4742	Guerrero Negro hypersaline microbial mat
JQ346728.1	Protochlamydia amoebophila strain UWE25	amoeba symbiont
JX846629.1	Protochlamydia naegleriophila strain Pcb1	amoeba symbiont
NR_074271.1	Protochlamydia amoebophila strain UWE25	amoeba symbiont
orpgwFw301_C1167	orf C1167	freshwater metagenome
Group_31	**PCF9**	**Marine**
EU491150.1	clone P9X2b3G08	seafloor lavas, Pacific Ocean
JQ013360.1	clone W5-15b	deep-sea sediment

Chapter V

Improved axenization method reveals complexity of symbiotic associations between bacteria and acanthamoebae

Published in Environmental Microbiology Reports

2014

environmental microbiology reports

Environmental Microbiology Reports (2014) doi:10.1111/1758-2229.12162

Improved axenization method reveals complexity of symbiotic associations between bacteria and acanthamoebae

Ilias Lagkouvardos,[1] Jie Shen[2] and Matthias Horn[1]*

[1]*Department of Microbiology and Ecosystem Science, University of Vienna, Vienna, Austria.*

[2]*School of Public Health, Department of Epidemiology, Fudan University, Shanghai, China.*

Summary

Bacteria associated with free-living amoebae have attracted considerable attention because of their role in human disease and as models for studying endosymbiosis. However, the identification and analysis of such novel associations are hindered by the limitations of methods for isolation and axenization of amoebae. Here, we replaced the heat-inactivated *Escherichia coli*, which is typically used as food source during axenization, with a live *E. coli tolC* knockout mutant strain hypersensitive to antibiotics. Together with the addition of otherwise sublethal amounts of ampicillin, this approach tripled the success rate and reduced the time required for axenization by at least 3 days. Using this method for two environmental samples, 10 *Acanthamoeba* strains were isolated, seven of which contained bacterial symbionts. In three cases, amoebae harbouring two phylogenetically distinct symbionts were recovered, supporting a more widespread occurrence of multi-partner symbiotic associations among free-living amoebae.

Introduction

Amoebozoa (amoebae) represent an ubiquitous and diverse group of unicellular eukaryotes with a predatory lifestyle that serve an important ecological role by controlling microbial communities and linking trophic levels in food webs (Rodriguez-Zaragoza, 1994; Rosenberg *et al*., 2009). However, some bacteria, including pathogens of humans and animals, have developed mechanisms to survive phagocytosis and to persist or even proliferate intracellularly in these protists (Barker and Brown, 1994; Greub and Raoult, 2004). Amoebae have thus been considered as training grounds and vectors of pathogenic bacteria (Molmeret *et al*., 2005; Corsaro and Greub, 2006; Thomas *et al*., 2006; 2010; Greub, 2009; Anacarso *et al*., 2012). They also serve as models for the study of host–pathogen interactions (Swanson and Hammer, 2000; Sandstrom *et al*., 2011) and for the analysis of the evolution of endosymbiosis and the intracellular lifestyle (Horn, 2008).

Acanthamoeba species, found typically in soil and freshwater, are frequently associated with obligate intracellular symbionts (Horn and Wagner, 2004). These bacteria belong to either of four distinct evolutionary lineages, the *Alphaproteobacteria* (Horn *et al*., 1999; Birtles *et al*., 2000), the *Betaproteobacteria* (Horn *et al*., 2002), the *Bacteroidetes* (Horn *et al*., 2001) or the *Chlamydiae* (Amann *et al*., 1997; Collingro *et al*., 2005b). *Acanthamoeba* isolates containing endosymbionts have been repeatedly recovered from geographically distant samples, with their symbionts often displaying high genetic similarity (Schmitz-Esser *et al*., 2008; Matsuo *et al*., 2010). In most cases, amoebae were associated with a single symbiont phylotype, but there have been a few reports in which two phylogenetically distinct symbionts were found co-occurring in a single amoeba host, raising questions about extent and implications of such multi-partner associations (Heinz *et al*., 2007; Matsuo *et al*., 2010).

There are two main methods for the discovery and identification of bacterial symbionts of amoeba. Either amoebae are directly isolated from environmental samples, or the samples are co-cultivated with amoeba lab strains (Horn *et al*., 1999; Collingro *et al*., 2005a; Thomas *et al*., 2006; Schmitz-Esser *et al*., 2008; Corsaro *et al*., 2009). Only the former method is able to recover bacterial symbionts together with their natural amoeba hosts. The isolation of amoebae was first described by Neff (1958). In this approach, environmental samples are applied to non-nutrient agar (NNA) plates covered with live *Escherichia coli*, and amoebae are then isolated as they graze and move away from the inoculation site. To facilitate a more detailed analysis of bacterial symbionts, the availability of axenic cultures, i.e. amoeba cultures in a nutrient-rich medium without bacteria as food source, is essential. The use of antibiotics to eliminate live *E. coli*

Received 13 December, 2013; accepted 30 January, 2014. *For correspondence. E-mail horn@microbial-ecology.net; Tel. (+43) 1 4277 76608; Fax (+43) 1 4277 876608.

during axenization is, however, not recommended as antibiotics may also inhibit intracellular symbionts. Therefore, prior to adaptation to the axenic medium, amoeba are generally first transferred to and passaged multiple times on NNA plates seeded with heat-inactivated *E. coli* to avoid bacterial growth in the medium. Because amoebae often grow poorly on heat-inactivated *E. coli* (de Moraes and Alfieri, 2008), this step is time-consuming, and amoebae often fail axenization.

In this study, we show that the use of *E. coli tolC* knockout mutants eliminates the need for the adaptation of amoebae to heat-inactivated *E. coli* and increases the success of axenization. Using this method, 10 amoeba isolates were obtained from two environmental samples. The prevalence and diversity of bacterial symbionts in these isolates suggests that associations of amoeba with symbiotic bacteria are more complex and dynamic than currently recognized.

Results and discussion

Using hypersensitive E. coli *for axenization of amoebae*

In order to bypass the limiting step of amoebal passage on heat-inactivated *E. coli*, we tested the use of *E. coli* strain JW5503-1 ΔtolC732::kan, one of the single-gene knockout mutants constructed in the Keio collection (Baba *et al.*, 2006) (*E. coli* Genetic Stock Center). In this strain, the gene coding for the outer membrane transporter TolC is replaced with a kanamycin resistance cassette. The deletion of *tolC* renders *E. coli* hypersensitive to ampicillin (minimum inhibitory concentration, MIC = 2 μg ml^{-1} compared with 6 μg ml^{-1} for the wild type) and gentamicin (MIC = 0.3 μg ml^{-1} compared with 0.8 μg ml^{-1}) without otherwise negative growth effect (Tamae *et al.*, 2008). This facilitates the selective elimination of *E. coli* using otherwise sublethal amounts of antibiotics and allows immediate transfer of amoebae from NNA plates seeded with the *E. coli tolC* knockout strain to culture flasks containing nutrient-rich axenic medium. In addition, it should be possible to supplement the axenic medium with live *E. coli* during the initial stages of axenization in order to facilitate adaptation to axenic growth conditions. In this set-up, growth of the hypersensitive *E. coli tolC* knockout strain in the axenic medium is controlled with low concentrations of ampicillin that are subinhibitory for intracellular symbionts.

Improved axenization of amoebae

We isolated amoebae from two samples collected at the Danube river bank (DRB) and from Danube river sediment (DRS) near the city of Vienna, Austria and compared axenization protocols with heat-inactivated *E. coli* K-12 and hypersensitive *E. coli* ΔtolC732::kan respectively. The environmental samples were placed on NNA plates covered with live hypersensitive *E. coli*. Plates were incubated at room temperature, and five agar pieces containing amoebae were excised for each sample and transferred to new NNA plates. The isolates (named DRB 1-5 and DRS 1-5 respectively) were passaged several times on NNA plates covered with live hypersensitive *E. coli*. Subsequently, amoebae were transferred either directly into axenic media amended with 10 μg ml^{-1} of ampicillin or to NNA plates covered with heat-inactivated *E. coli* prior to the transfer to axenic media (see Supporting Information Appendix S1 for additional details).

Nine out of the 10 isolates transferred directly to axenic culture media could be successfully axenized. In contrast, only three isolates could be adapted to axenic growth when first transferred to heat-inactivated *E. coli* (Fig. 1). In more detail, two isolates (DRS2, DRB4) displayed immediate growth inhibition and eventual encystation on NNA with heat-inactivated *E. coli*. Five isolates grew on heat-inactivated *E. coli* but could then not be adapted to axenic conditions. Using hypersensitive compared with heat-inactivated *E. coli* thus increased the number of successfully axenized isolates by a factor of three.

Transfer to and growth on heat-inactivated *E. coli* required an average of 3 days to achieve a distance from the transfer site (approximately 2 cm) to ensure no transmission of live *E. coli* to the axenic media. In addition, for one isolate (DRS4), adaptation to axenic media via heat-inactivated *E. coli* took 5 days longer than when directly transferred to media supplemented with ampicillin (Fig. 1). Thus, omitting the passage on heat-inactivated *E. coli* and using hypersensitive *E. coli* and ampicillin saved between 3 and 8 days for axenization. Taken together, our novel approach minimizes axenization time and significantly increases the success rate.

Diversity and co-occurrence of bacterial symbionts

All of the nine amoeba isolates where axenization was successful could be assigned to the genus *Acanthamoeba* based on morphological characteristics. Sequencing of 18S rRNA genes revealed that all but one are highly similar to each other and most closely related to *A. castellanii* Neff (GenBank Acc. U07416) (GenBank Acc. KF924599, KF92601, KF92605). The 18S rRNA gene of one isolate (DRS2, GenBank Acc. KF92600) was most similar to *A. polyphaga* Nagington (GenBank Acc. AF019062) (Supporting Information Appendix S1). Seven isolates contained bacterial symbionts, which could be readily detected by staining with 4,6-diamidino-2-phenylindole dihydrochloride and subsequent fluorescence *in situ* hybridization (FISH) (Daims *et al.*, 2005) (Supporting Information Appendix S1). Sequencing of bacterial 16S rRNA genes and phylogenetic analysis

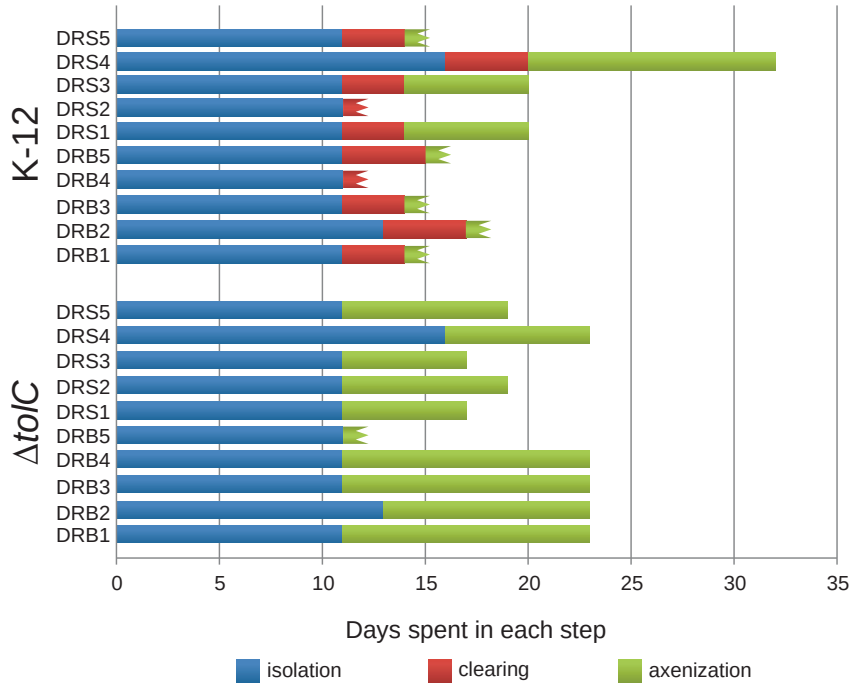

Fig. 1. Comparison of traditional and *E. coli* Δ*tolC*-based methods for the isolation and axenization of free-living amoebae. The course of isolation and axenization is shown for 10 amoeba strains obtained from two different samples (DRS1-5, DRB1-5) using either a traditional protocol and heat-inactivated *E. coli* or the improved protocol with live hypersensitive *E. coli* cells as food source. The time required for each step is colour-coded. Blue is the time required for growth on NNA plates (isolation phase); red is the time required for adaptation and growth on NNA plates covered with heat-inactivated *E. coli* (cleaning phase, only necessary in the traditional approach); green is the time needed for adaptation to growth in liquid axenic media (axenization phase). A broken end indicates failure (encystation) at the respective stage of isolation/axenization.

demonstrated that the bacterial symbionts show a remarkable identity (> 99%) to known bacterial symbionts of *Acanthamoeba* species recovered previously from diverse samples and locations (Table 1, Supporting Information Appendix S1). In total, six phylogenetically distinct symbionts were identified, including representatives from all major taxonomic groups of *Acanthamoeba* symbionts (Table 1). Three out of seven amoeba isolates contained two phylogenetically different bacterial symbionts, with a different combination of symbionts for each case. The intracellular location of all symbionts was confirmed by FISH using specific probes for each taxonomic group (Fig. 2).

The amoeba isolates recovered in this study are remarkable in two respects. First, the large diversity of symbionts in different isolates obtained from the same environmental sample, or from two samples that were taken in close spatial proximity, is in sharp contrast with previous studies that generally reported single symbiont-containing amoeba isolates per sample (Collingro *et al.*,

Table 1. Taxonomic affiliation of bacterial symbionts.

Amoeba isolate	Symbiont taxonomy	Acc. no.	Closest neighbour (Acc. no.)	% identity
DRB1	*Paraceadibacter*	KF924589	*Paracaedibacter* sp. UWC9 (AF132137) (Horn *et al.*, 1999)	99.5
DRB2	*Neochlamydia*	KF924590	Endosymbiont of *Acanthamoeba* sp. S13 (AB506677) (Matsuo *et al.*, 2010)	99.6
DRS1	*Protochlamydia*	KF924591	*Protochlamydia* sp. CRIB40 (FJ532293) (Thomas *et al.*, 2006)	99.5
	Procabacter	KF924592	*Procabacter* sp. UWE2 (AF177424) (Horn *et al.*, 2002)	98.6
DRS2	*Neochlamydia*	KF924593	Endosymbiont of *Acanthamoeba* sp. S13 (AB506677) (Matsuo *et al.*, 2010)	99.5
DRS3	*Rickettsiales*	KF924595	Endosymbiont of *Acanthamoeba* sp. UWC36 (AF069962) (Fritsche *et al.*, 1999)	99.1
	Amoebophilus	KF924594	*Amoebophilus asiaticus* US1 (HM159369) (Schmitz-Esser *et al.*, 2010)	99.5
DRS4	*Protochlamydia*	KF924597	*Protochlamydia* sp. CRIB40 (FJ532293) (Thomas *et al.*, 2006)	99.7
	Amoebophilus	KF924596	*Amoebophilus asiaticus* US1 (HM159369) (Schmitz-Esser *et al.*, 2010)	99.6
DRS5	*Protochlamydia*	KF924598	*Protochlamydia* sp. CRIB40 (FJ532293) (Thomas *et al.*, 2006)	99.6

The closest neighbours in 16S rRNA gene-based phylogenetic trees were used to determine the taxonomic affiliation of bacterial symbionts found in seven axenized amoeba isolates.

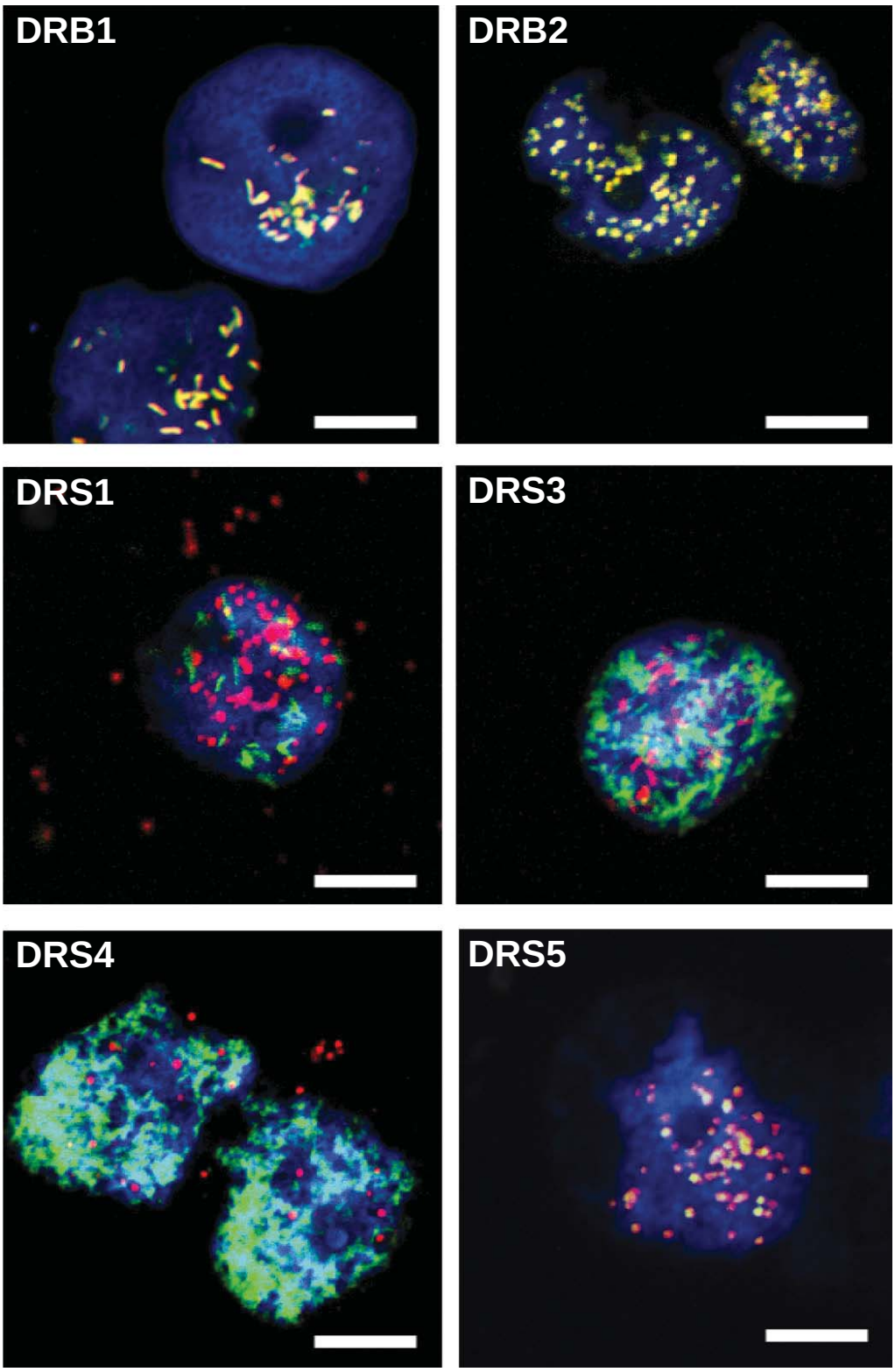

Fig. 2. Identification of bacterial symbionts by fluorescence *in situ* hybridization. The bacterial symbionts of the recovered *Acanthamoeba* isolates were identified by 16S rRNA gene sequencing (Table 1) and FISH using general probes for eukaryotes (EUK516, blue) and bacteria (EUB338I-III, green in DRB1, DRB2, DRS5), and group-specific probes for the symbionts: *Caedibacter* (CC23a, red in DRB1), *Chlamydiales* (Chls523, red in DRB2, DRS1, DRS4, DRS5), Procabacter (Proca438, green in DRS1), *Amoebophilus* (Aph1180, green in DRS3, DRS4) and rickettsiae (AcRic90, red in DRS3). Additional information on the probes is available at the oligonucleotide database probeBase (http://www.microbial-ecology.net/probebase; (Loy *et al.*, 2007). The overlap of red and green fluorescence signals appears yellow. Bar, 10 μm.

2005b; Schmitz-Esser *et al*., 2008). Second, the co-occurrence of two phylogenetically different symbionts in three out of seven *Acanthamoeba* isolates is unexpected, as the majority of amoeba isolates investigated so far contain only a singly symbiont phylotype (Heinz *et al*., 2007; Corsaro *et al*., 2010; Matsuo *et al*., 2010).

In order to investigate whether the co-occurrence of phylogenetically different symbionts represents a natural phenomenon or an artefact of the isolation and axenization method, single cysts were picked using a micromanipulator from the first NNA plates from which three isolates originated (DRS1, DRS3 and DRS5), and clonal cultures were established (Supporting Information Appendix S1). In total, nine clonal cultures were obtained, and the analysis of these isolates by FISH confirmed the presence and co-occurrence of the bacterial symbionts detected earlier (data not shown). This suggests that co-infections of amoebae are likely not artefacts of the isolation method but that amoebae containing phylogenetically different symbionts occurred naturally in the investigated sample.

In summary, the recovery of different amoeba isolates from two related environmental samples using the improved isolation and axenization protocol revealed a surprising diversity of bacterial symbionts and an unexpected large fraction of isolates containing more than one symbiont phylotype.

Conclusions

The occurrence of nearly identical bacterial symbionts in *Acanthamoeba* isolates obtained from geographically distinct regions has been noted earlier and suggested a global distribution of the major lineages of *Acanthamoeba* symbionts (Fritsche *et al*., 2000; Molmeret *et al*., 2005; Schmitz-Esser *et al*., 2008). Here, we could show that a similar diversity of symbionts may exist at a much smaller scale, within two adjacent environmental samples. Moreover, 40% of the amoeba isolates recovered here contained not only a single but two different symbiont phylotypes. This suggests that bacterial symbionts of acanthamoebae are more promiscuous than previously recognized and that the interaction between these amoebae and their diverse symbionts is much more complex under natural conditions. The role of the bacterial symbionts for the biology and ecology of acanthamoebae is, however, currently unknown. While many of them seem to tap their host's metabolism (Trentmann *et al*., 2007; Haferkamp *et al*., 2013), which is indicative for a parasitic lifestyle, a potential benefit for the host cannot be excluded and might explain their prevalence, even at small spatial scales.

The recovery of multiple amoeba isolates and their diverse symbionts in this study was facilitated by an improved isolation and axenization protocol. The use of a live hypersensitive *E. coli* strain reduced the time and labour required for axenization and importantly increased the success rate. The method introduced here thus allows for a deeper screening of environmental samples for free-living amoebae. Its application in future studies and the combination of limited dilution or flow cytometry-based sorting with the hypersensitive *E. coli* strain as food source will help to further illuminate the diversity and complexity of symbiotic associations between bacteria and amoeba in nature.

Acknowledgements

This work was funded by Austrian Science Fund (FWF) grant Y277-B03 and the University of Vienna (Graduate School 'Symbiotic Interactions'). Jie Shen was supported by a grant from the Austrian Agency for International Mobility and Cooperation in Education, Science and Research (OeAD). Matthias Horn acknowledges support from the European Research Council (ERC StG 'EvoChlamy').

References

Amann, R., Springer, N., Schonhuber, W., Ludwig, W., Schmid, E.N., Muller, K.D., and Michel, R. (1997) Obligate intracellular bacterial parasites of *Acanthamoebae* related to *Chlamydia* spp. *Appl Environ Microbiol* **63:** 115–121.

Anacarso, I., de Niederhausern, S., Messi, P., Guerrieri, E., Iseppi, R., Sabia, C., and Bondi, M. (2012) *Acanthamoeba polyphaga*, a potential environmental vector for the transmission of food-borne and opportunistic pathogens. *J Basic Microbiol* **52:** 261–268.

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., *et al*. (2006) Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol* **2:** 2006–2008.

Barker, J., and Brown, M.R.W. (1994) Trojan-horses of the microbial world – protozoa and the survival of bacterial pathogens in the environment. *Microbiology* **140:** 1253–1259.

Birtles, R.J., Rowbotham, T.J., Michel, R., Pitcher, D.G., Lascola, B., Alexiou-Daniels, S., and Raoult, D. (2000) 'Candidatus *Odyssella thessalonicensis*' gen. nov., sp nov., an obligate intracellular parasite of *Acanthamoeba* species. *Int J Syst Evol Microbiol* **50:** 63–72.

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*. *Microbiology* **151:** 301–309.

Collingro, A., Toenshoff, E.R., Taylor, M.W., Fritsche, T.R., Wagner, M., and Horn, M. (2005b) 'Candidatus *Protochlamydia amoebophila*', an endosymbiont of *Acanthamoeba* spp. *Int J Syst Evol Microbiol* **55:** 1863–1866.

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

Corsaro, D., Michel, R., Walochnik, J., Muller, K.D., and Greub, G. (2010) *Saccamoeba lacustris* sp nov (Amoebozoa: Lobosea: Hartmannellidae), a new lobose amoeba, parasitized by the novel chlamydia 'Candidatus *Metachlamydia lacustris*' (Chlamydiae: Parachlamydiaceae). *Eur J Protistol* **46:** 86–95.

Daims, H., Stoecker, K., and Wagner, M. (2005) Fluorescence *in situ* hybridization for the detection of prokaryotes. In *Molecular Microbial Ecology*. Osborn, A.M., and Smith, C.J. (eds). Abingdon, UK: Bios-Garland, pp. 213–239.

Fritsche, T.R., Horn, M., Seyedirashti, S., Gautom, R.K., Schleifer, K.H., and Wagner, M. (1999) *In situ* detection of novel bacterial endosymbionts of *Acanthamoeba* spp. phylogenetically related to members of the order Rickettsiales. *Appl Environ Microbiol* **65:** 206–212.

Fritsche, T.R., Horn, M., Wagner, M., Herwig, R.P., Schleifer, K.H., and Gautom, R.K. (2000) Phylogenetic diversity among geographically dispersed Chlamydiales endosymbionts recovered from clinical and environmental isolates of *Acanthamoeba* spp. *Appl Environ Microbiol* **66:** 2613–2619.

Greub, G. (2009) *Parachlamydia acanthamoebae*, an emerging agent of pneumonia. *Clin Microbiol Infect* **15:** 18–28.

Greub, G., and Raoult, D. (2004) Microorganisms resistant to free-living amoebae. *Clin Microbiol Rev* **17:** 413–433.

Haferkamp, I., Penz, T., Geier, M., Ast, M., Mushak, T., Horn, M., and Schmitz-Esser, S. (2013) The endosymbiont *Amoebophilus asiaticus* encodes an S-adenosylmethionine carrier that compensates for its missing methylation cycle. *J Bacteriol* **195:** 3183–3192.

Heinz, E., Kolarov, I., Kastner, C., Toenshoff, E.R., Wagner, M., and Horn, M. (2007) An *Acanthamoeba* sp containing two phylogenetically different bacterial endosymbionts. *Environ Microbiol* **9:** 1604–1609.

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

Horn, M., and Wagner, M. (2004) Bacterial endosymbionts of free-living amoebae. *J Eukaryot Microbiol* **51:** 509–514.

Horn, M., Fritsche, T.R., Gautom, R.K., Schleifer, K.H., and Wagner, M. (1999) Novel bacterial endosymbionts of *Acanthamoeba* spp. related to the *Paramecium caudatum* symbiont *Caedibacter caryophilus*. *Environ Microbiol* **1:** 357–367.

Horn, M., Harzenetter, M.D., Linner, T., Schmid, E.N., Muller, K.D., Michel, R., and Wagner, M. (2001) Members of the *Cytophaga-Flavobacterium-Bacteroides* phylum as intracellular bacteria of *Acanthamoebae*: proposal of 'Candidatus *Amoebophilus asiaticus*'. *Environ Microbiol* **3:** 440–449.

Horn, M., Fritsche, T.R., Linner, T., Gautom, R.K., Harzenetter, M.D., and Wagner, M. (2002) Obligate bacterial endosymbionts of *Acanthamoeba* spp. related to the beta-*Proteobacteria*: proposal of 'Candidatus *Procabacter acanthamoebae*' gen. nov., sp. nov. *Int J Syst Evol Microbiol* **52:** 599–605.

Loy, A., Maixner, F., Wagner, M., and Horn, M. (2007) probeBase – an online resource for rRNA-targeted oligonucleotide probes: new features 2007. *Nucleic Acids Res* **35:** D800–D804.

Matsuo, J., Kawaguchi, K., Nakamura, S., Hayashi, Y., Yoshida, M., Takahashi, K., *et al*. (2010) Survival and transfer ability of phylogenetically diverse bacterial endosymbionts in environmental *Acanthamoeba* isolates. *Environ Microbiol Rep* **2:** 524–533.

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

de Moraes, J., and Alfieri, S.C. (2008) Growth, encystment and survival of *Acanthamoeba castellanii* grazing on different bacteria. *FEMS Microbiol Ecol* **66:** 221–229.

Neff, R.J. (1958) Mechanisms of purifying amoebae by migration on agar surfaces. *J Protozool* **5:** 226–231.

Rodriguez-Zaragoza, S. (1994) Ecology of free-living amoebae. *Crit Rev Microbiol* **20:** 225–241.

Rosenberg, K., Bertaux, J., Krome, K., Hartmann, A., Scheu, S., and Bonkowski, M. (2009) Soil amoebae rapidly change bacterial community composition in the rhizosphere of *Arabidopsis thaliana*. *ISME J* **3:** 675–684.

Sandstrom, G., Saeed, A., and Abd, H. (2011) *Acanthamoeba*-bacteria: a model to study host interaction with human pathogens. *Curr Drug Targets* **12:** 936–941.

Schmitz-Esser, S., Toenshoff, E.R., Haider, S., Heinz, E., Hoenninger, V.M., Wagner, M., and Horn, M. (2008) Diversity of bacterial endosymbionts of environmental *Acanthamoeba* isolates. *Appl Environ Microbiol* **74:** 5822–5831.

Schmitz-Esser, S., Tischler, P., Arnold, R., Montanaro, J., Wagner, M., Rattei, T., and Horn, M. (2010) The genome of the amoeba symbiont 'Candidatus *Amoebophilus asiaticus*' reveals common mechanisms for host cell interaction among amoeba-associated bacteria. *J Bacteriol* **192:** 1045–1057.

Swanson, M.S., and Hammer, B.K. (2000) *Legionella pneumophila* pathogenesis: a fateful journey from amoebae to macrophages. *Annu Rev Microbiol* **54:** 567–613.

Tamae, C., Liu, A., Kim, K., Sitz, D., Hong, J., Becket, E., *et al*. (2008) Determination of antibiotic hypersensitivity among 4000 single-gene-knockout mutants of *Escherichia coli*. *J Bacteriol* **190:** 5981–5988.

Thomas, V., Herrera-Rimann, K., Blanc, D.S., and Greub, G. (2006) Biodiversity of amoebae and amoeba-resisting bacteria in a hospital water network. *Appl Environ Microbiol* **72:** 2428–2438.

Thomas, V., McDonnell, G., Denyer, S.P., and Maillard, J.Y. (2010) Free-living amoebae and their intracellular pathogenic microorganisms: risks for water quality. *FEMS Microbiol Rev* **34:** 231–259.

Trentmann, O., Horn, M., van Scheltinga, A.C.T., Neuhaus, H.E., and Haferkamp, I. (2007) Enlightening energy parasitism by analysis of an ATP/ADP transporter from chlamydiae. *PLoS Biol* **5:** 1938–1951.

Supporting information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Appendix S1. Supplementary experimental procedures.

Improved axenization method reveals complexity of symbiotic associations between bacteria and acanthamoebae

Ilias Lagkouvardos, Jie Shen, Matthias Horn

Supplementary Experimental Procedures

Media recipes

Page Amoeba Saline (PAS) (10 x) (Page, 1967)	
NaCl	1,2 g
$MgSO_4$*7 H_2O	0,04 g
$CaCl_2$*2 H_2O	0,04 g
Na_2HPO_4	1,42 g
KH_2PO_4	1,36 g
Distilled water	1 L
Non Nutrient Agar (NNA)	
PAS (10 x)	100 ml
Agar	15-20 g
Distilled water	900 ml
PYG (pH 6.5)	
Peptone	20 g
Glucose	18 g
Yeast Extract	2 g
Sodium citrate	1 g
$MgSO_4$*7 H_2O	980 mg
Na_2HPO_4*7 H_2O	355 mg
KH_2PO_4	340 mg
$Fe(NH_4)_2(SO_4)_2$* 6H_2O	20 mg
Distilled water	1L

PYNFH (pH 6.4) (De Jonckheere, 1993)	
Peptone, Bacto	10 g
Yeast Extract	10 g
Yeast nucleic acid	1 g
Folic acid	15 mg
Hemin	1 mg
Buffer Solution	20 ml
Distilled water	880 ml
*add 100ml FBS after autoclave	
Buffer Solution (pH 6.5)	
KH_2PO_4	18.1 g
Na_2HPO_4	25.0 g
Distilled water	1L
LB (pH 7)	
Tryptone	10 g
Yeast Extract	5 g
NaCl	5 g
Distilled water	1L
TSY (pH 7.3)	
Trypticase Soy Broth	30 g
Yeast Extract	10 g
Distilled water	1L

PYNFH and PYG were autoclaved at 110 °C; PAS, NNA, and TSY at 120 °C

Preparation of *E. coli* cells

Escherichia coli strain JW5503-1 ΔtolC732::kan (Baba et al., 2006) was obtained from the culture collection of the *E. coli* Genetic Stock Center at Yale University and streaked on Lysogeny Broth agar (LB) plates. Single colonies were picked and tested for their resistance to kanamycin (50 µg/ml) and susceptibility to ampicillin (10 µg/ml) (Tamae et al., 2008). Finally a single colony with the correct phenotype was selected and a glycerol stock was made. For preparation of *E. coli* cultures for plating on non-nutrient agar (NNA) plates, a flask containing 100 ml of LB medium was inoculated and incubated overnight at 37 °C. Bacteria were harvested by centrifugation (8000 rpm), washed once and resuspended in 10 ml Page's Amoebic Saline (PAS). For heat inactivation, *E. coli* K-12 cells resuspended in PAS were placed in a water bath for 1h at 95°C. An aliquot of this preparation was tested for growth on LB plates to ensure complete inactivation. Live and heat-inactivated *E. coli* cells were stored at 4 °C until usage.

Isolation and axenization of amoebae

The isolation of amoebae was based on the "walk out" method described by Neff (Neff, 1958). Aliquots of the suspension containing live *E. coli* cells (200 µl) were spread on NNA plates. Environmental samples were added in the middle of the plate and incubated at room temperature. Amoebae that have migrated away from the environmental sample were identified using an inverted light microscope and excised as agar pieces (around 0.5 cm^2), which were placed upside down on fresh NNA plates covered with live *E. coli*. The plates were incubated until the amoebae almost reached the rim of the plates at which time point they were transferred again. This procedure was carried out three times in total to ensure that contaminants from the samples would not be carried over.

For axenization six agar pieces containing amoebae with live *E. coli* cells were excised per plate and transferred to a 6-well cell culture plate containing 10 ml of growth medium per well (2 x TSY, 2 x PYG and 2 x PYNFH) supplemented with 10 µl of the live *E. coli* cell suspension and a final concentration of 10 µg/ml ampicillin.

Alternatively, agar pieces containing amoebae were transferred to fresh NNA plates covered with heat-inactivated *E. coli*. The plates were incubated, and amoebae were allowed to migrate before the transfer to 6-well cell culture plates containing the same growth media but supplemented with heat-inactivated *E. coli* and without antibiotics.

The cell culture plates were examined by microscopy every two days, and when an increase in the number of amoebae was observed they were transferred to 25 cm^2 cell culture flasks.

If live *E. coli* was used the axenic medium was replaced after one day and supplemented with ampicillin to prevent growth of residual live *E. coli*. No ampicillin was added during subsequent medium exchanges. A culture was considered axenic if amoebae continued to grow in the absence of extracellular bacteria.

Establishment of clonal amoeba cultures

A piece of agar was removed from NNA plates and the amoeba trophozoites or cysts were washed into the agar hole applying 1 ml of PAS repeatedly on the surface of the plate. The plate was then placed under an inverted microscope (Axio Observer.D1 Zeiss) with a micromanipulator (TransferMan NK2 & CellTram vario, Eppendorf). Single amoeba cysts or trophozoites were picked using glass capillaries (TransferTip, Eppendorf) and placed individually on new NNA plates seeded with live *E. coli*.

Identification of amoeba hosts and bacterial symbionts

Amoeba isolates were fixed on microscope slides and initially screened for the presence of bacterial symbionts using the DNA stain 4',6-diamidine-2'-phenylindole dihydrochloride (DAPI). Cultures containing symbionts were harvested from 25 m^2 culture flasks by centrifugation at 7500 rpm, and total DNA was isolated using the DNeasy Blood & Tissue Kit (Qiagen). PCR was performed using primers targeting the 16S ribosomal RNA gene (616v, 5'-AGAGTTTGATYMTGGCTC (Juretschko et al., 1998), and 1492R, 5'-GGYTACCTTGTTACGACTT (Loy et al., 2005)) and an annealing temperature of 52 °C. Since the forward primer 616v does not cover members of the phylum *Chlamydiae* well, the alternative forward primer SigF2 (5'CRGCGTGGATGAGGCAT, (Haider et al., 2008) was used in addition. For 18S rRNA of the hosts, the primer JDP1 (5′-GGCCCAGATCGTTTACCGTGAA) and the reverse primer JDP2 (5′-TCTCACAAGCTGCTAGGGAGTCA) (Schroeder et al., 2001) were used with an annealing temperature of 62 °C. All PCR products were purified and cloned in TopoXL vectors (TopoXL cloning Kit, Invitrogen Life Technologies). Cloned 16S and 18S rRNA gene fragments were then re-amplified using M13 primers and sequenced on an ABI 3130 XL Genetic Analyzer.

Obtained 16S rRNA gene sequences were used to search against the NCBI nt database using BLAST (Altschul et al., 1990). The 20 best hits for each sequence were collected, aligned by the SILVA aligner (Quast et al., 2013) and used for the calculation of phylogenetic trees with the maximum likelihood method (GTR model, 16 gamma categories, 100 bootstraps) implemented in MEGA5 (Tamura et al., 2011). The closest neighbors of the newly obtained sequences were determined based on the tree topology.

Fluorescence *in situ* hybridization

Axenized amoebae grown in culture flasks were harvested in 50 ml tubes and centrifuged at 5000 rpm for 5 min. After discarding the supernatant and re-suspension of the amoeba pellet in 10 ml PAS, the amoeba cells were centrifuged again at 5000 rpm for 5min. The final pellet was re-suspended in 5 ml PAS, 15 µl were placed on a glass slide, and amoebae were allowed to attach to the glass surface. After 30 min buffer was removed and amoeba trophozoites were fixed using 4% paraformaldehyde (PFA) for 15 min at room temperature. After fixation PFA was removed and the cells were washed using double distilled water. Hybridization was carried out

as described elsewhere. Slides were examined by epifluorescence and confocal laser scanning microscopy (Axioplan 2 Zeiss, CLSM LSM 510 Meta Zeiss).

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic Local Alignment Search Tool. *Journal of Molecular Biology* **215**: 403-410.

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M. et al. (2006) Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol* **2**: 2006 0008.

De Jonckheere, J.F. (1993) A group I intron in the SSUrDNA of some Naegleria spp. demonstrated by polymerase chain reaction amplification. *J Eukaryot Microbiol* **40**: 179-187.

Haider, S., Collingro, A., Walochnik, J., Wagner, M., and Horn, M. (2008) *Chlamydia*-like bacteria in respiratory samples of community-acquired pneumonia patients. *FEMS Microbiol Lett* **281**: 198-202.

Juretschko, S., Timmermann, G., Schmid, M., Schleifer, K.H., Pommerening-Roser, A., Koops, H.P., and Wagner, M. (1998) Combined molecular and conventional analyses of nitrifying bacterium diversity in activated sludge: *Nitrosococcus mobilis* and *Nitrospira*-like bacteria as dominant populations. *Appl Environ Microbiol* **64**: 3042-3051.

Loy, A., Schulz, C., Lucker, S., Schopfer-Wendels, A., Stoecker, K., Baranyi, C. et al. (2005) 16S rRNA gene-based oligonucleotide microarray for environmental monitoring of the betaproteobacterial order "*Rhodocyclales*". *Appl Environ Microbiol* **71**: 1373-1386.

Neff, R.J. (1958) Mechanisms of purifying amoebae by migration on agar surfaces. *J Protozool* **5**: 226-231.

Page, F.C. (1967) Taxonomic Criteria for Limax Amoebae with Descriptions of 3 New Species of Hartmannella and 3 of Vahlkampfia. *Journal of Protozoology* **14**: 499-&.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P. et al. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41**: D590-596.

Schroeder, J.M., Booton, G.C., Hay, J., Niszl, I.A., Seal, D.V., Markus, M.B. et al. (2001) Use of subgenic 18S ribosomal DNA PCR and sequencing for genus and genotype identification of acanthamoebae from humans with keratitis and from sewage sludge. *Journal of Clinical Microbiology* **39**: 1903-1911.

Tamae, C., Liu, A., Kim, K., Sitz, D., Hong, J., Becket, E. et al. (2008) Determination of antibiotic hypersensitivity among 4,000 single-gene-knockout mutants of *Escherichia coli*. *J Bacteriol* **190**: 5981-5988.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011) MEGA5: Molecular evolutionary genetics analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony methods. *Mol Biol Evol* **28**: 2731-2739.

Chapter VI

Genome of Acanthamoeba castellanii highlights extensive lateral gene transfer and early evolution of tyrosine kinase signaling

Published in Genome Biology 2013

Supplementary online materials are available at: http://genomebiology.com/2013/14/2/R11/additional

Genome Biology

RESEARCH Open Access

Genome of *Acanthamoeba castellanii* highlights extensive lateral gene transfer and early evolution of tyrosine kinase signaling

Michael Clarke[1†], Amanda J Lohan[1†], Bernard Liu[2], Ilias Lagkouvardos[3], Scott Roy[4], Nikhat Zafar[5], Claire Bertelli[6], Christina Schilde[7], Arash Kianianmomeni[8], Thomas R Bürglin[9], Christian Frech[10], Bernard Turcotte[11], Klaus O Kopec[12], John M Synnott[1], Caleb Choo[10], Ivan Paponov[13], Aliza Finkler[14], Chris Soon Heng Tan[15], Andrew P Hutchins[16], Thomas Weinmeier[17], Thomas Rattei[17], Jeffery SC Chu[18], Gregory Gimenez[19], Manuel Irimia[20], Daniel J Rigden[21], David A Fitzpatrick[22], Jacob Lorenzo-Morales[23], Alex Bateman[24], Cheng-Hsun Chiu[25], Petrus Tang[26], Peter Hegemann[8], Hillel Fromm[14], Didier Raoult[19], Gilbert Greub[6], Diego Miranda-Saavedra[16], Nansheng Chen[10], Piers Nash[27], Michael L Ginger[28], Matthias Horn[3], Pauline Schaap[7], Lis Caler[5] and Brendan J Loftus[1*]

Abstract

Background: The Amoebozoa constitute one of the primary divisions of eukaryotes, encompassing taxa of both biomedical and evolutionary importance, yet its genomic diversity remains largely unsampled. Here we present an analysis of a whole genome assembly of *Acanthamoeba castellanii* (*Ac*) the first representative from a solitary free-living amoebozoan.

Results: *Ac* encodes 15,455 compact intron-rich genes, a significant number of which are predicted to have arisen through inter-kingdom lateral gene transfer (LGT). A majority of the LGT candidates have undergone a substantial degree of intronization and *Ac* appears to have incorporated them into established transcriptional programs. *Ac* manifests a complex signaling and cell communication repertoire, including a complete tyrosine kinase signaling toolkit and a comparable diversity of predicted extracellular receptors to that found in the facultatively multicellular dictyostelids. An important environmental host of a diverse range of bacteria and viruses, *Ac* utilizes a diverse repertoire of predicted pattern recognition receptors, many with predicted orthologous functions in the innate immune systems of higher organisms.

Conclusions: Our analysis highlights the important role of LGT in the biology of *Ac* and in the diversification of microbial eukaryotes. The early evolution of a key signaling facility implicated in the evolution of metazoan multicellularity strongly argues for its emergence early in the Unikont lineage. Overall, the availability of an *Ac* genome should aid in deciphering the biology of the Amoebozoa and facilitate functional genomic studies in this important model organism and environmental host.

Background

Acanthamoeba castellanii (*Ac*) is one of the predominant soil organisms in terms of population size and distribution, where it acts both as a predator and an environmental reservoir for a number of bacterial, fungal and viral species [1]. Selective grazing by *Ac* in the rhizosphere alters microbial community structure and is an important contributor to the development of root architecture and nutrient uptake by plants [2]. *Ac* can also be isolated from almost any body of water and manifests in a wide variety of man-made water systems, including potable water sources, swimming pools, hot tubs, showers and hospital air conditioning units [3,4]. Acanthamoebae are frequently associated with a diverse range of bacterial

* Correspondence: brendan.loftus@ucd.ie
† Contributed equally
[1]Conway Institute, University College Dublin, Belfield, Dublin 4, Ireland
Full list of author information is available at the end of the article

symbionts [5,6]. A subset of the microbes that serve as prey for *Ac* have evolved virulence stratagems to use *Ac* as both a replicative niche and as a vector for dispersal and are important human intracellular pathogens [7,8]. These pathogens utilize analogous strategies to infect and persist within mammalian macrophages, illustrating the role of environmental hosts such as *Ac* in the evolution and maintenance of virulence [9,10]. Commonalities at the level of host response between amoebae and macrophages to such pathogens have led to the use of both *Dictyostelium discoideum* (*Dd*) and *Ac* as model systems to study pathogenesis [11,12].

Published Amoebozoa genomes from both the obligate parasite *Entamoeba histolytica* (*Eh*) and the facultatively multicellular *Dd* have both highlighted unexpected complexities at the level of cell motility and signaling [13,14]. As the only solitary free-living representative, the genome of *Ac* establishes a unique reference point for comparisons for the interpretation of other amoebozoan genomes. Experimentally, *Ac* has been a more thoroughly studied organism than most other free living amoebae, acting as a model organism for studies on the cytoskeleton, cell movement, and aspects of gene regulation, with a large body of literature supporting its molecular interactions [15-18].

Results and discussion

Lateral gene transfer

Lateral gene transfer (LGT) is considered a key process of genome evolution and several studies have indicated that phagotrophs manifest an increased rate of LGT compared to non-phagotrophic organisms [19]. As a geographically dispersed bacteriovorous amoebae with a penchant for harboring endosymbionts, *Ac* encounters a rich and diverse supply of foreign DNA, providing ample opportunity for LGT. Homology-based searches of the proteome illustrate the potential for diverse contributions to the genome (Figure 1).

We therefore undertook a phylogenomic analysis to determine cases of predicted inter-domain LGT in the *Ac* genome (Section 2 of Additional file 1). Our analysis identified 450 genes, or 2.9% of the proteome, predicted to have arisen through LGT (Figure 2; Section 2 of Additional file 1). To determine the fate and ultimate utility of the LGT candidates within the *Ac* genome, we examined their expression levels across a number of experimental conditions using RNA.seq (Table S1.6.1 in Additional file 1). Our results show that most of the LGT candidates are expressed in at least some of the conditions tested (Additional file 2).

Genetic exchange is also thought to occur between phylogenetically disparate organisms that reside within the same amoebal host cell [20,21]. *Ac* contains three copies of a miniature transposable element (ISSoc2) of the IS607 family of insertion sequences related to those present in genomes of thermophilic cyanobacteria [22] and several giant nucleocytoplasmic large DNA viruses (NCLDVs). In the Mimivirus genome the IS elements are found within islands of genes of bacterial origin, some of which appear to have been contributed by a cyanobacterial donor. This data underscores the complex intermediary role that *Ac*, as host to both NCLDVs and cyanobacteria [17] may play in facilitating genetic transfer between sympatric species.

Comparison of predicted LGT across amoeboid genomes

In order to compare the impact and scale of LGT across *Ac* and other amoeba, we applied the same phylogenomic approach used to identify LGT in the *Ac* genome to published genomes of other amoeboid protists, including *Dd*, *Eh*, *Entamoeba dispar* (*Ed*) and *Naegleria gruberi* (*Ng*). Our findings predict that *Ac* and the excavate *Ng* encode a notably higher number of laterally acquired bacterial genes than either of the more closely related parasitic *Entamoeba* or the social *Dd* amoebozoans (Figure 2a). The taxonomic distribution of putative LGT donors is broadly similar for both *Entamoeba* species, but surprisingly also between *Ac* and *Ng* (Figure 2b,c; Section 2 of Additional file 1). The genomes of both *Eh* and *Ed* are predicted to have experienced a proportionately higher influx from anaerobic and host-associated microbes than their free-living counterparts *Ac* and *Ng* (Figure 2c; Additional file 2), likely reflecting the composition of microbes within their habitats. Many of the LGT candidates across all of the amoebae have predicted metabolic functions, suggesting that LGT in amoebae is reflective of trophic strategy and driven by the selective pressure of new ecological niches. Our data illustrating LGT as a contributing factor in shaping the biology of a diversity of amoeboid genomes provide further evidence supporting an underappreciated role for LGT in the diversification of microbial eukaryotes [23].

Introns

Intron-exon structures exhibit complex phylogenetic patterns with orders-of-magnitude differences across eukaryotic lineages, which imply frequent transformations during eukaryotic evolution [24]. Some researchers have argued that intron gain is episodic with long periods of stasis [25] punctuated by periods of rapid gain while others argue for generally higher rates [26]. Strikingly, *Ac* genes have an average of 6.2 introns per gene, among the highest known in eukaryotes [27]. Genes predicted to have arisen through LGT have slightly lower but broadly comparable intron densities, offering an opportunity to study the evidence for proposed mechanisms underpinning post-LGT intron gain [28]. An analysis of LGT introns, however, did not provide support for any of the proposed

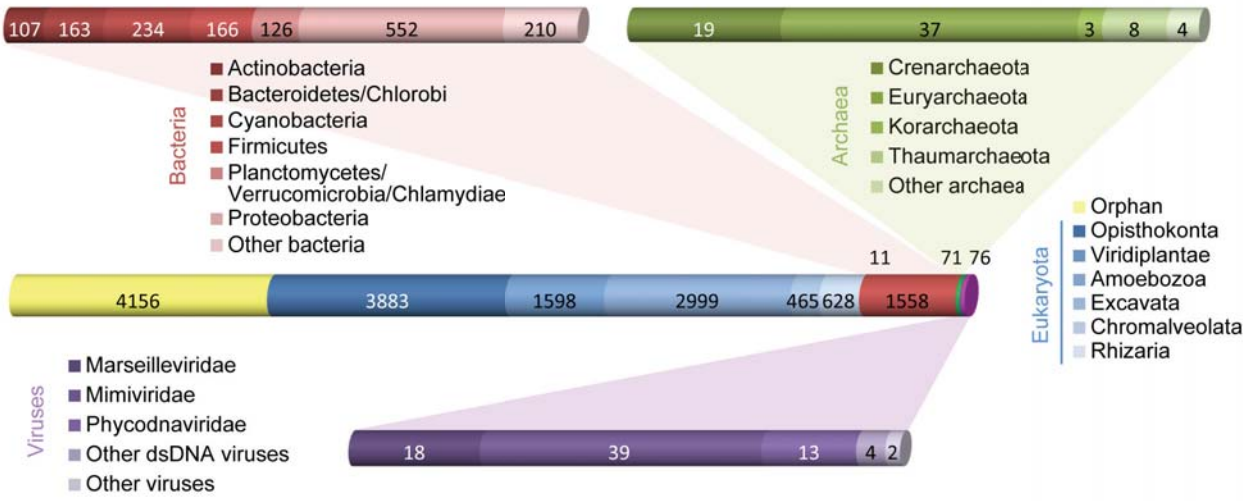

Figure 1 Measures of the composition of the *Ac* genome based on sequence similarity. For each protein, the best BLASTP hit to the non-redundant database, that is, the match with the lowest e-value, was recovered and the classification of the corresponding organism was extracted according to NCBI taxonomy. The central bar represents the full complement of annotated *Ac* genes exhibiting a best BLASTP hit respectively against the four kingdoms - Eukaryota (blue), Bacteria (red), Archaea (green) and viruses (purple) - with orphan genes depicted in yellow. Results for Eukaryota are subdivided according to the major taxonomic phyla in varying shades of blue. Subdivisions of phyla within the Bacteria (red shading), Archaea (green shading) and viruses (purple shading) are depicted in the expanded upper and lower sidebars. dsDNA, double-stranded DNA.

mechanisms of intron gain (Section 2 of Additional file 1). Thus, while the preponderance of introns in LGTs clearly indicates substantial intron gain at some point, it appears that, for *Ac*, these events have been very rare in recent times, consistent with a punctate model of intron gain.

Cell signaling

As a unicellular sister grouping to the multicellular Dictyostelids, *Ac* provides a unique point of comparison to gain insight into the molecular underpinnings of multicellular development in Amoebozoa. Cell-cell communication is a hallmark of multicellularity and we looked at putative receptors for extracellular signals and their downstream targets. G-protein-coupled receptors (GPCRs) represent one of the largest families of sensors for extracellular stimuli. Overall, *Ac* encodes 35 GPCRs (compared to 61 in *Dd*), representing 4 out of the 6 major families of GPCRs [29] while lacking metabotropic glutamate-like GPCRs or fungal pheromone receptors. We identified three predicted fungal-associated glucose-sensing Git3 GPCRs [30] and an expansion in the number of frizzled/smoothened receptors [31] (Figure S3.1.1 in Additional file 1). We identified seven G-protein alpha subunits and a single putative target, phospholipase C, for GPCR-mediated signaling. The number and diversity of receptors in *Ac* raises the question of what they are likely to be sensing. Nematodes employ many of their GPCRs in detecting molecules secreted by their bacterial food sources [32], and given the diversity of *Ac*'s feeding environments, many of the *Ac* GPCRs may fulfill a similar role.

Environmental sensing

We identified 48 sensor histidine kinases (SHKs), of which 17 harbor transmembrane domains and may function as receptors (Figure S3.2.1 in Additional file 1). Remarkably, there are also 67 nucleotidyl cyclases consisting of an extracellular receptor domain separated by a single transmembrane helix from an intracellular cyclase domain flanked by two serine/threonine kinase domains. This domain configuration is present in a number of the amoeba-infecting giant viruses but thus far appears unique for a cellular organism (Figure S3.3.1 in Additional file 1). *Ac* is able to survive under microaerophilic conditions such as those found in the deeper layers of underwater sediments or within the rhizosphere. The genome encodes a number of prolyl 4-hydroxylases that likely mediate oxygen response; however, *Ac* also contains a number of heme-nitric oxide/oxygen binding (H-NOX) proteins that, unlike those in other eukaryotes, are not found in conjunction with guanylyl cyclases [33]. The *Ac* H-NOX proteins lack a critical tyrosine residue in the non-polar distal heme pocket, making it likely that they are for nitric oxide (NO) rather than oxygen signaling [34]. Both *Dd* and *Ac* are responsive to light, although the photoreceptor that mediates phototaxis in *Dictyostelium* has yet to be identified [35]. We identified two rhodopsins both with

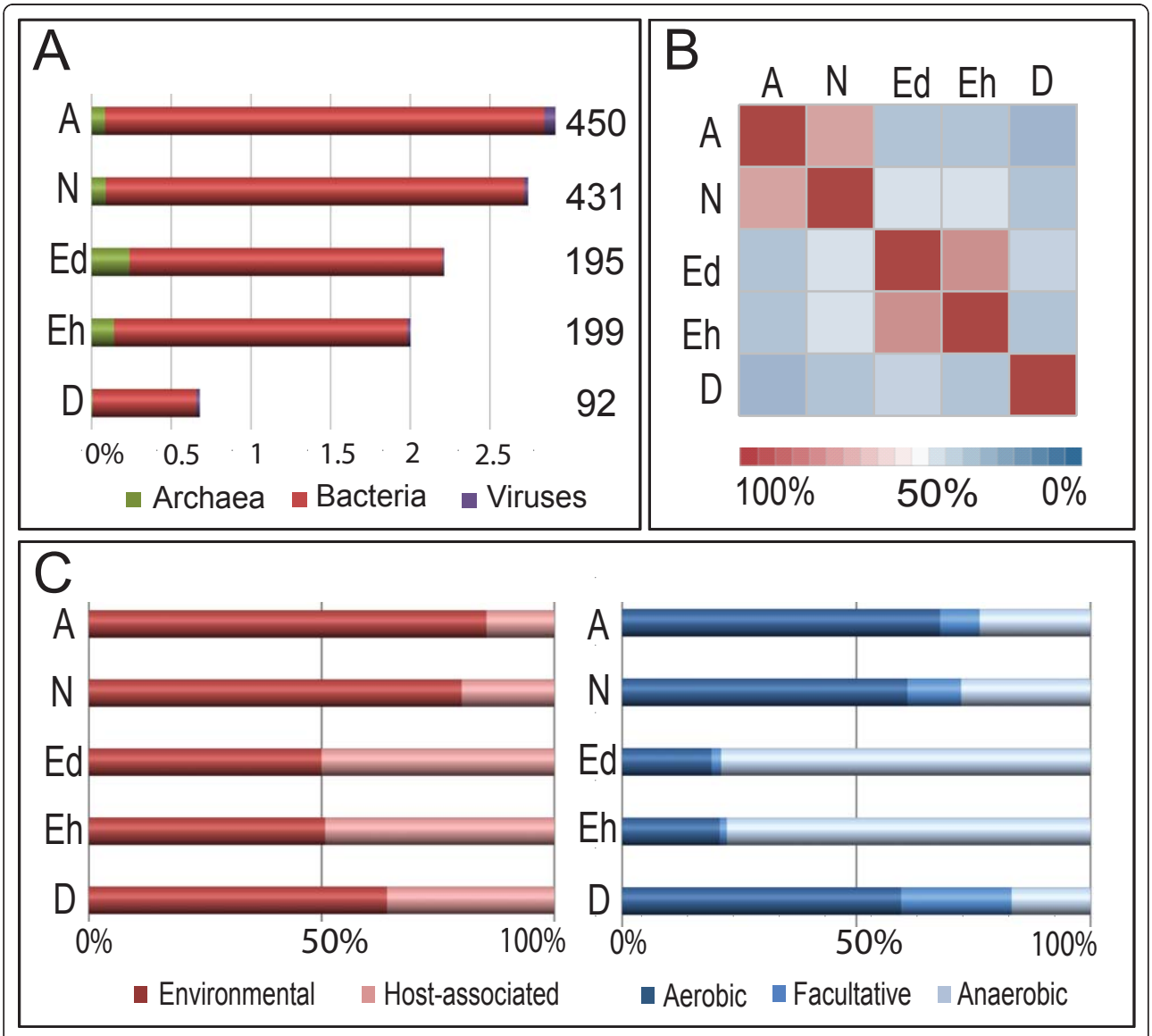

Figure 2 Predicted LGT-derived genes from Bacteria, Archaea and viruses encoded in the genomes of free-living and parasitic amoebae. LGT-derived genes were predicted using a phylogenomics approach consisting of an initial similarity-based screening using SIMAP [111], several filtering steps to extract amoebal proteins with prokaryotic best hits, followed by automatic calculation and manual inspection of phylogenetic trees using PhyloGenie and PHAT [112]. **(a)** Percentage of lineage-specific LGT candidates in each genome; the absolute number of LGT candidates per genome is indicated next to each bar. **(b)** Heat map illustrating the Bray-Curtis similarity of the taxonomic affiliation (at the level of classes within the domain Bacteria) of putative LGT donors. **(c)** Ecological classification of putative LGT donors with respect to their oxygen requirement and association with a host. The ecology of putative donors was extrapolated from the lifestyles of the respective closest extant relatives.

carboxy-terminal histidine kinase and response regulator domains with homology to the sensory rhodopsins of the green algae that represent candidates for light sensors in *Ac* (Figure 3).

Cellular response

Modulation of cellular response to environmental cues is enacted by a diversity of protein kinases and *Ac* is predicted to encode 377, the largest number predicted to date for any amoebozoan (Section 4 of Additional file 1). In *Ac*, the mitogen-activated protein kinase (MAPK) kinase pathway has been shown to be involved in encystment [36] and its genome encodes homologues of both of *Dd*'s two MAPK proteins, ErkA and ErkB [37]. Phosphotyrosine (pTyr) signaling mediated through tyrosine kinases was until recently thought to be generally absent from the amoebozoan lineage [38]. This signaling capacity has been associated with intercellular communication, the evolutionary step towards

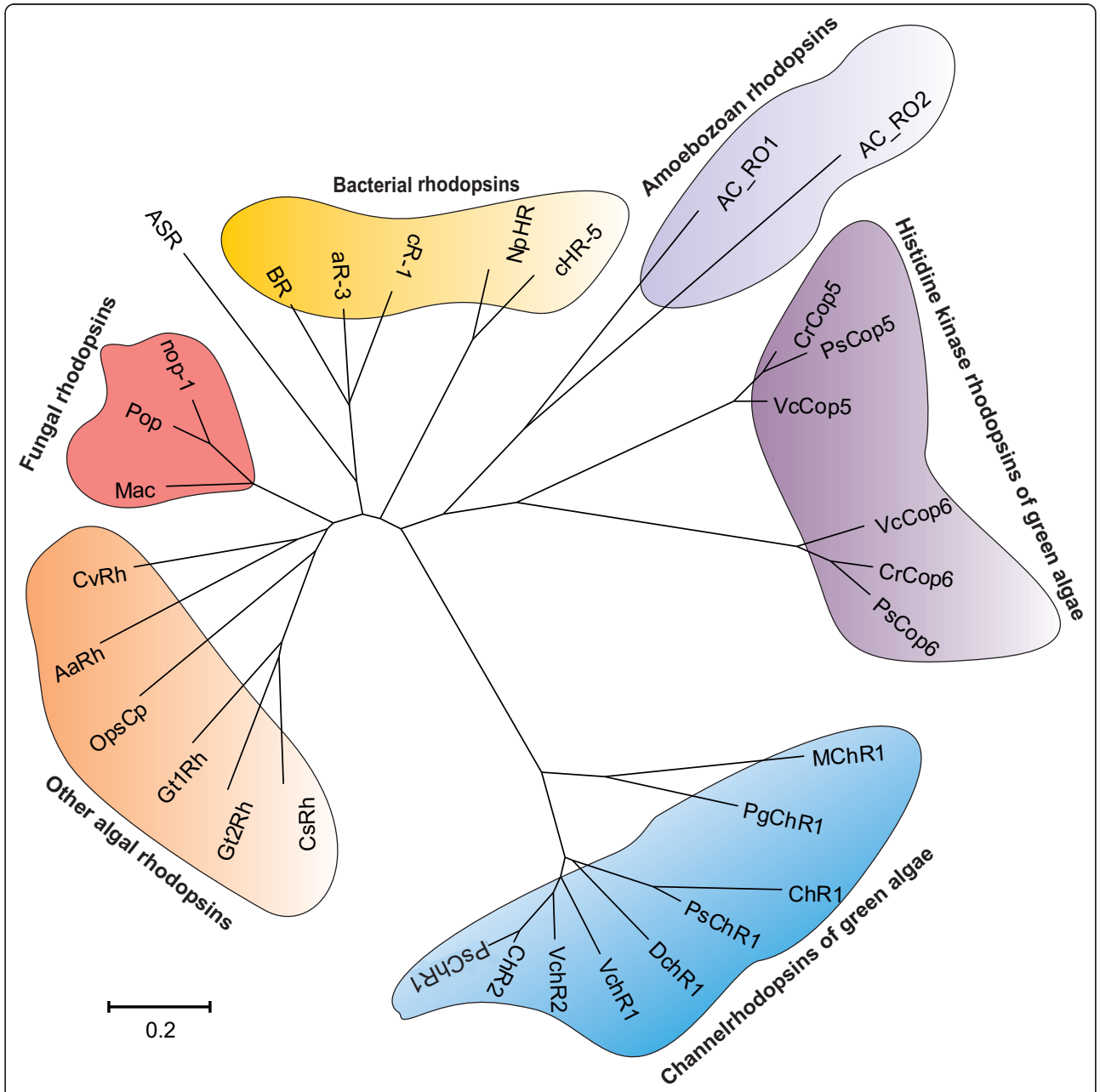

Figure 3 Phylogenetic tree of rhodopsins from Amoebozoa, algae, bacteria and fungi. The tree was constructed by the neighbor-joining method based on the amino acid sequence of the rhodopsin domain using MEGA version 5 [113]. The scale bar indicates the number of substitutions per site. Detailed rhodopsin information is listed in Table S3.6.1 in Additional file 1.

multicellularity and the expansion of organismal complexity in metazoans [39]. pTyr is thought to depend upon a triad of signaling molecules; tyrosine kinase 'writers' (PTKs), tyrosine phosphatase 'erasers' (PTPs) and Src homology 2 (SH2) 'reader' domains that connect the phosphorylated ligand-containing domains to specify downstream signaling events [39]. Remarkably, the genome of *Ac* encodes 22 PTKs, 12 PTPs, and 48 SH2 domain-containing proteins (Figure 4a), revealing a primordial yet elaborate pTyr signaling system in the amoebozoan lineage (Figure 4b).

The *Ac* PTK domains are highly conserved in key catalytic residues, resembling dedicated PTKs found in metazoans (Figure S4.2.1 in Additional file 1), and are distinct from *Dd* and *Eh* PTKs that are more tyrosine kinase like (TKL) (Figure S4.2.2 in Additional file 1). *Ac* PTK homologues are present in the apusomonad

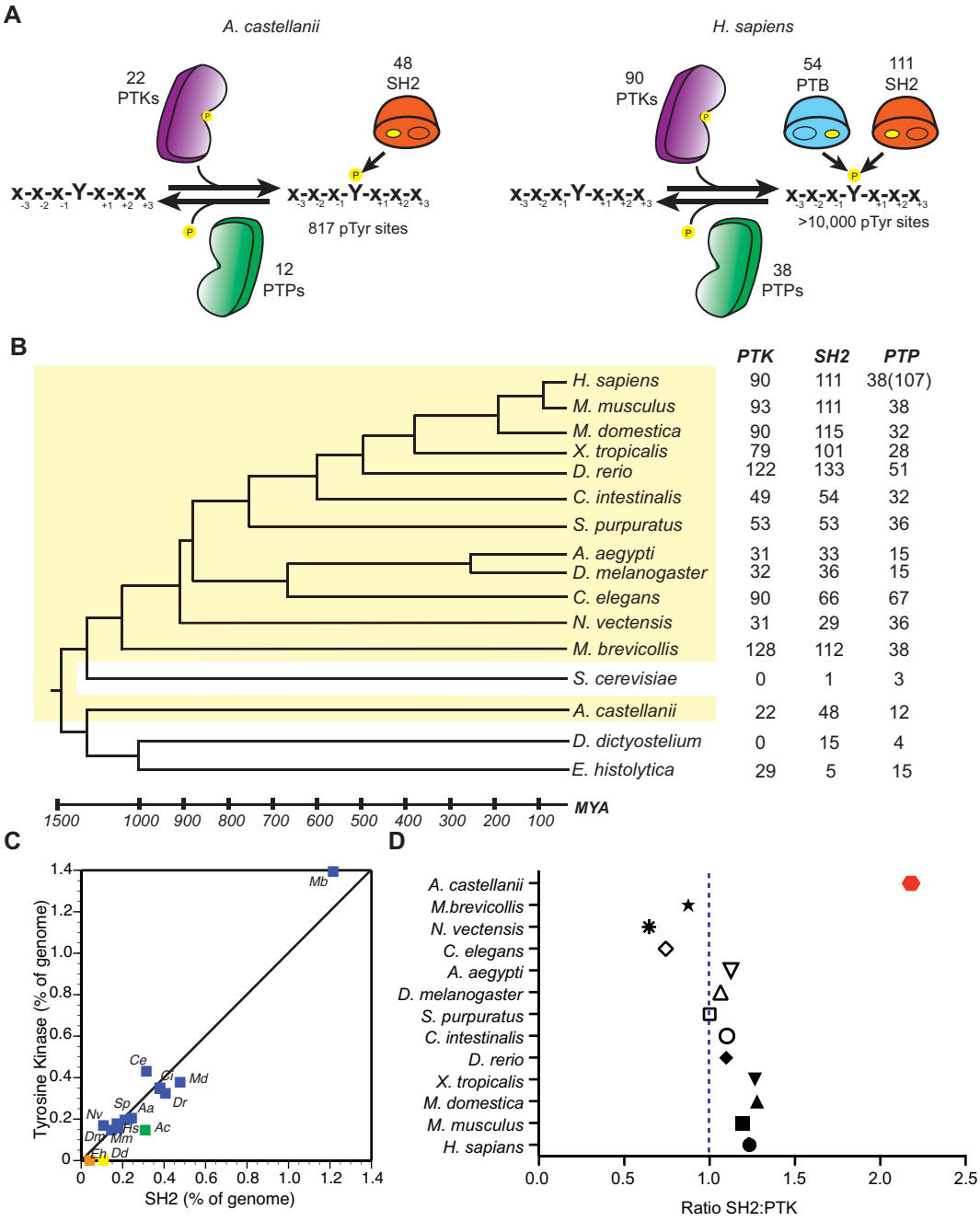

Figure 4 The phosphotyrosine signaling circuitry of *Ac*. **(a)** Phosphotyrosine signaling is modulated by the writers (PTKs), erasers (PTPs) and readers (Src homology 2 (SH2); phosphotyrosine binding (PTB)). **(b)** The total number of PTKs, classical PTPs (total PTPs) and SH2 encoded genes across multiple eukaryote genomes. Highlighted in yellow are branches that compose a complete phosphotyrosine signaling circuit. The branched divergence times and lengths in millions of years (mya) are indicated. **(c)** The percentage of the genome devoted to encoding PTKs and SH2 domains. **(d)** *Ac* displays the greatest ratio of SH2:PTKs compared to other eukaryotes.

Thecamonas trahens and have also recently been described in two filasterean species, *Capsaspora owczarzaki* and *Ministeria vibrans* [38]. One unusual feature of the pTyr machinery in *Ac* is the 2:1 ratio of SH2 to PTK domains as comparisons across opisthokonts show a strong correlation and co-expansion of these two domains with a ratio close to 1:1 (Figure 4c,d) [40]. This increased ratio in *Ac* indicates either an expansion to handle the cellular requirements of pTyr signaling or that aspects of PTK function are accomplished by TKL

or dual specificity kinases as appears to be the case in *Dd* [41]. We also found that *Ac* has fewer tyrosine residues in its proteome in comparison to *Dd*, which lacks PTKs (Figure S4.3.1 in Additional file 1). This result is in line with recent analysis of metazoan genomes, suggesting increased pressure for selection against disadvantageous phosphorylation of tyrosine residues in genomes with extensive pTyr signaling [42].

Domain organization and composition of pTyr components reveal the selective pressures for adapting pTyr signaling into various pathways. Seven PTKs have predicted transmembrane domains and may function as receptor tyrosine kinases hinting at their potential for intercellular communication. The majority of PTKs in *Ac*, however, show unique domain combinations; six PTKs contain a sterile alpha motif (SAM) domain, which is found in members of the ephrin receptor family (Figure S4.4.3 in Additional file 1). The *Ac* SH2 proteins are conserved within the pTyr binding pocket and resemble SH2 domains from the SOCS, RIN, CBL and RASA families (Figure S4.4.2 in Additional file 1); however, the domain composition within these proteins differs between those of *Monosiga brevicollis* and metazoans (Figure S4.4.3A in Additional file 1). Approximately half of the *Ac* SH2 proteins share domain architectures with *Dd*, including the STAT family of transcription factors (Figure S4.4.3B in Additional file 1). The presence of homologous SH2 proteins in *Dd* coupled with the complete facility in *Ac* predicts an emergence of the complete machinery for pTyr early in the Unikont lineage. This finding is in contrast with models that posit a complete pTyr signaling machinery emerging late in the Unikont lineage [39] and has important implications for understanding the relationship between pTyr signaling and the evolution of multicellularity. The lack of clear metazoan orthologues makes it difficult to trace the evolutionary paths of pTyr signaling networks [43] or to accurately predict the cellular functions and adaptations of pTyr in *Ac*. However, with phosphoproteomics and sequence analysis, insights into ancient pTyr signaling circuits may be revealed through future studies in *Ac* (Figure S4.5.1 in Additional file 1).

Cell adhesion

Ac is not known to participate in social activity yet must adhere to a diversity of surfaces within the soil and practice discrimination between self and prey during phagocytosis [44]. *Ac* shares some adhesion proteins with *Dd* (Table S5.1.1 in Additional file 1) but homologues of the calcium-dependent, integrin-like Sib cell-adhesion proteins are absent. Surprisingly, *Ac* contains a number of bacterial-like integrin and hemagglutinin domain adhesion proteins that may improve its ability to attach to bacterial cells or biofilms [45]. *Ac* encodes two MAM domain-containing proteins, a domain found in functionally diverse receptors with roles in cell-cell adhesion [46]. *Ac* has a copy of the laminin-binding protein (AhLBP) first identified in *Acanthamoeba healyi*, which has been shown to act as a non-integrin laminin binding receptor [47]. Remarkably, *Ac* also encodes proteins containing cell adhesion immunoglobulin domains (Section 5 of Additional file 1). Both show affinity to the I-set subfamily [48] and contain weakly predicted transmembrane domains (Figure S5.1.1 in Additional file 1).

Microbial recognition through pattern recognition receptors

Ac grazes on a variety of micro fauna, which requires the mobilization of a set of defense responses initiated upon microbial recognition. In vertebrates molecular signatures often termed microbe-associated molecular patterns (MAMPs) [49] are detected by pattern-recognition receptors (PRRs) that activate downstream transcriptional responses. As *Ac* practices selective feeding behavior we looked for the presence of predicted PRRs in the *Ac* genome (Figure 5). One of the best-studied MAMPs is lipopolysaccharide and discrimination mediated through lectin-mediated protein-carbohydrate interactions is an important innate immunity strategy in both vertebrates and invertebrates [50]. *Ac* contains six members of the bactericidal permeability-increasing protein (BPI)/lipopolysaccharide-binding protein (LBP) family and two peptidoglycan binding proteins (Figure 5; Section 6 of Additional file 1). *Ac* also encodes a membrane bound homologue of an MD-2-related protein that, in vertebrate immunity, has been implicated in opsonophagocytosis of Gram-negative bacteria through its interactions with lipopolysaccharide [51].

Receptor-mediated endocytosis of *Legionella pneumophila* in *Ac* is mediated by the c-type lectin mannose binding protein (MBP) [52]. MBP also represents the principal virulence factor in pathogenic Acanthamoebae [53]. In addition to MBP, the *Ac* genome encodes two paralogues of MBP with similarity to the amino-terminal region of the protein. Rhamnose-binding lectins serve a variety of functions in invertebrates, one of which is their role as germline-encoded PRRs in innate immunity [54]. They are absent from other Amoebozoa, although *Ac* encodes 11 D-galactoside/L-rhamnose binding (SUEL) lectin domain-containing proteins. Approximately half of the SUEL lectin domain proteins harbour epidermal growth factor domains, a combination reminiscent of the selectin family of adhesion proteins found exclusively in vertebrates [55]. An L-rhamnose synthesis pathway thought to contribute to biosynthesis of the lipopolysaccharide-like outer layer of the virus particle has recently been identified in Mimivirus that may facilitate its uptake by *Ac* [56,57]. *Ac* also encodes a protein where multiple

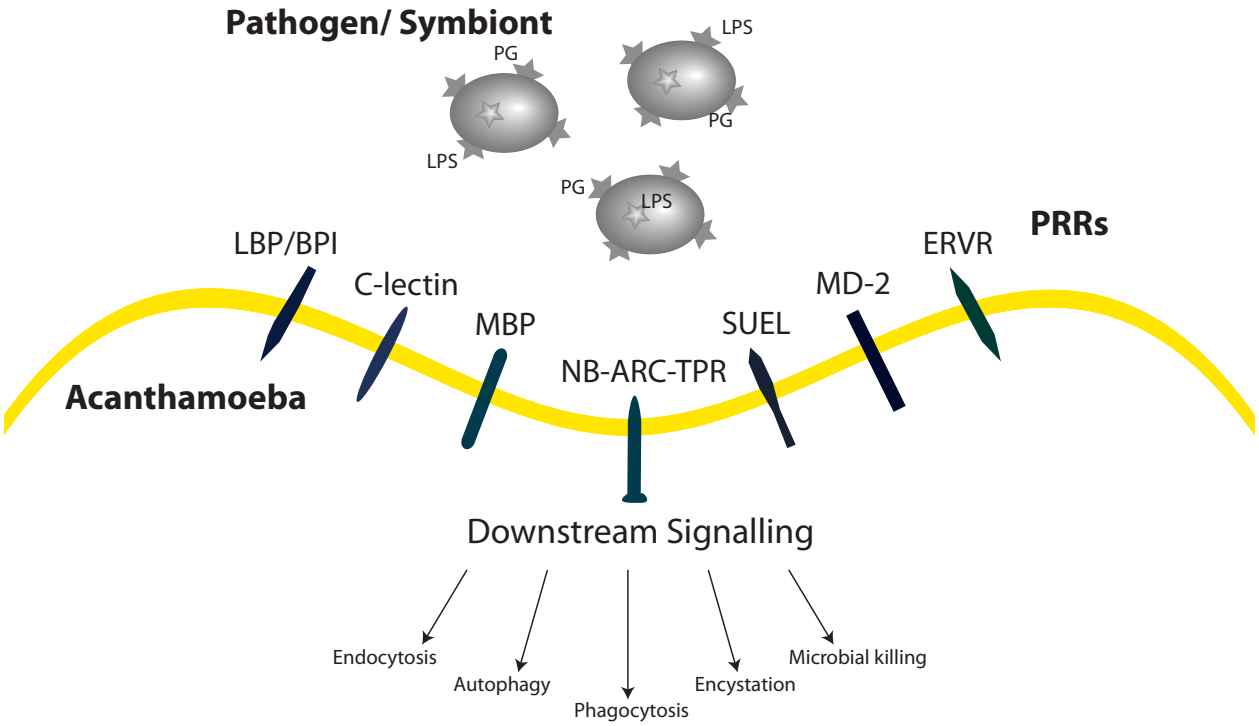

Figure 5 Potential PRRs in *Ac*: LBP/BPI, lipopolysaccharide binding protein/ bactericidal permeability-increasing protein; C-lectin, C-type lectin: MBP, mannose binding protein; SUEL, D-galactoside/L-rhamnose binding SUEL lectin domain containing; NB-ARC-TPR, NB-ARC tetratricopeptide repeat containing protein; ERVR, endogenous virus receptor.

copies of H-type lectin are joined with an inhibitor of apoptosis domain. The H-lectin domain is predicted to bind to N-acetylgalactosamine (GalNAc) and is found in *Dictyostelium* discoidin I & II [58] and other invertebrates where it plays a role in antibacterial defense [59]. In the brown algae *Ectocarpus* leucine-rich repeat (LRR) containing GTPases of the ROCO family and NB-ARC-TPR proteins have been proposed to represent PRRs that are involved in immune response [60]. *Ac* encodes a NB-ARC-TPR homologue with a disease resistance domain (IPR000767) and an LRR-ROCO GTPase.

Antimicrobial defense

Ac encodes proteins with potential roles in antiviral defense including homologues of NCLDV major capsid proteins [61] as well as homologues of Dicer and Piwi, both of which have been implicated in RNA-mediated antiviral silencing [62]. Our data also illustrate early evolution of a number of interferon-inducible innate immunity proteins absent from other sequenced Amoebozoa. These include a homologue of the interferon-γ-inducible lysosomal thiol reductase enzyme (GILT), an important host factor targeted by *Listeria monocytogenes* during infection in macrophages [63]. In addition, *Ac* encodes two interferon-inducible GTPase homologues, which in vertebrates promote cell-autonomous immunity to vacuolar bacteria, including *Mycobacteria* and *Legionella* species [64]. *Ac* also contains a natural resistance-associated macrophage protein (NRAMP) homologue, which has been implicated in protection against *L. pneumophila* and *Mycobacterium avium* infection in both macrophages and *Dd* [65].

Metabolism

Ac has traditionally been considered to be an obligate aerobe, although the recent identification of the oxygen-labile enzymes pyruvate:ferredoxin oxidoreductase and FeFe-hydrogenase perhaps pointed towards a cryptic capacity for anaerobic ATP production [66]. Predictions for nitrite and fumarate reduction, hydrogen fermentation, together with a likely mechanism for acetate synthesis, coupled to ATP production indicate a considerable capacity for anaerobic ATP generation. This clearly sets *Ac* apart from *Dd*, which hunts within the aerobic leaf litter, but provides parallels with *Ng*, the alga *Chlamydomonas reinhardtii* and other soil-dwelling protists that are likely to experience considerable variation in local oxygen tensions [67]. These protists achieve their flexible, facultative anaerobic metabolism, however, using different pathways (Figure S7.1 in Additional file 1). In addition, the classic anaerobic twists on glycolysis provided by pyrophosphate-dependent phosphofructokinase and pyruvate phosphate dikinase [68] are absent from *Ac*. This suggests that

although multiple pathways are available for oxidation of NADH to NAD+ in the absence of oxygen, including a capacity for anaerobic respiration in the presence of nitrite (NO_2^-), a shift to a more ATP-sparing form of glycolysis is not necessary under low oxygen-tension. Given genome-led predictions of facultative anaerobic ATP metabolism, as well as extensive use of receptors and signaling pathways classically associated with animal biology, we also considered the possibility of a hypoxia-inducible factor (HIF)-dependent system for oxygen sensing, similar to that seen across the animal kingdom, including the simple animal *Trichoplax adhaerens* [69,70]. However, despite conservation of a Skp1/HIFα-related prolyl hydroxylase in *Ac*, we found no genes encoding proteins with the typical domain architecture of animal HIFα or HIFβ. Currently, therefore, HIF-dependent oxygen sensing remains restricted to metazoan lineages.

Ac also retains biosynthetic pathways involved in anabolic metabolism that are absent in *Dd* (for example, the shikimic acid pathway and a classic type I pathway for fatty acid biosynthesis; Table S7.1 in Additional file 1), although investment in extensive polyketide biosynthesis [71] is not evident. An autophagy pathway, as defined by genetic studies of yeast, *Dd* and other organisms [72], is present in *Ac* with little paralogue expansion or loss of known autophagy-related (ATG) genes evident (Figure 7.2 in Additional file 1) and likely contributes to both intracellular re-modeling in response to environmental cues and the interaction with phagocytosed microbes.

Transcription factors

Ac shares a broadly comparable repertoire of transcription factors with *Dd* excepting a number of lineage-specific expansions (Table S8.1 in Additional file 1). *Ac* encodes 22 zinc cluster transcription factors compared to the 3 in *Dd* (Figure S8.2.1 in Additional file 1) [73]. It has almost double the number of predicted homeobox genes (25) compared to the 13 in *Dd* [74]. Two are of the MEIS and PBC class respectively, with an expansion in a homologue of Wariai, a regulator of anterior-posterior patterning in *Dictyostelium* [75] comprising most of the additional members (Figure S8.3.2 in Additional file 1). Strikingly, we also identified 22 Regulatory factor × (*RFX*) genes, the first identified in an Amoebozoan [76]. The *Ac RFX* repertoire is the earliest branching yet identified and forms an outgroup to other known *RFX* genes (Section 8 of Additional file 1). *Ac* has been proposed to affect plant root branching in the rhizosphere via its effects on auxin balance in plants [77]. It encodes a number of genes involved in auxin biosynthesis as well as those involved in free auxin (indole-3-acetic acid (IAA)) de-activation via formation of IAA conjugates (Table S9.1 in Additional file 1). These data suggest that *Ac* plays a role in altering the level of IAA in the rhizosphere through a strategy of alternative biosynthesis and sequestration. *Ac* may also respond transcriptionally to auxin as it encodes a member of the calmodulin-binding transcription activator (CAMTA) family (Figure S8.4.1 in Additional file 1), which in plants coordinate stress responses via effects on auxin signaling [78,79].

Conclusions

Comparative genomics of the Amoebozoa has until now been restricted to comparisons between the multicellular dictyostelids and the obligate parasite *Eh* [80,81]. *Ac*, while sharing many of their features, enriches the repertoire of amoebozoan genomes in a number of important areas, including signaling and pattern recognition. LGT has significantly contributed to both the genome and transcriptome of *Ac* whose accessory genome shares unexpected similarities with a phylogenetically distant amoeba. The presence of prokaryotic TEs in *Ac* illustrates its role in the evolution of some of the earth's most unusual organisms [82] as well a number of important human pathogens [7,8][83].

Ac has adopted bacterial-like adhesion proteins to facilitate adherence to biofilms and H-NOX based nitric oxide signaling which likely aids in their dispersal [84]. Overall the adaptive value conferred by LGT is highlighted by the expression of the large majority in *Ac* across multiple conditions, which points to their adoption into novel transcriptional networks. Given the feeding behavior of *Ac*, it seems plausible that eukaryote-to-eukaryote gene transfers may also have provided adaptive benefits [23]. Increased sampling will be necessary to establish the extent to which such gene transfers made their way into the *Ac* genome and whether 'you are what you eat' equally applies to a diet of eukaryotes [23].

Ac participates in a myriad of as yet unexplored interactions, as reflected in the diversity of genes devoted to sensory perception and signal transduction of extracellular stimuli. *Ac*'s survival in the rhizosphere is likely contingent on interactions not only with other microbes but also on a cross-talk with plant roots through manipulation of the levels of the plant hormone auxin. LGT may also have provided *Ac* with some of its recognition and environmental sensing components. An interesting parallel is the planktonic protozoan *Oxyrrhis marina*, which utilizes both MBP and LGT-derived sensory rhodopsins, to enable selective feeding behavior through prey detection and biorecognition [85]. We predict that host response of *Ac* to pathogens and symbionts is likely modulated via a diversity of predicted PRRs that act in an analogous manner to effectors of innate immunity in higher organisms. Given the close association of *Ac* with a number of important intracellular pathogens, it will be interesting to determine which host-pathogen interactions can trace their origins to encounters with primitive cells such as *Ac*.

Ac shares protein family expansions in signal transduction with other Amoebozoa while introducing new components based on novel domain architectures (nucleotidyl cyclases) [86]. The presence of the complete pTyr signaling toolkit especially when contrasted with its absence in the multicellular dictyostelids is a remarkable finding of the *Ac* genome analysis. However the role of tyrosine kinase signaling in both amoebozoan and mammalian phagocytosis [87-89] indicates that it likely represents an ancestral function. The most parsimonious interpretation predicts the supplanting of functions originally carried out by tyrosine kinases by other kinases in the Amoebozoa. This emphasizes the importance of representative sampling and in its absence the inherent difficulties in re-constructing ancestral signaling capacities.

Transcriptional response networks can be re-programmed either through expansion of transcription factors or their target genes [90]. *Ac* and *Dd* share a conserved core of transcription factors with any differences between them largely accounted for by lineage-specific amplifications. These may result in sub- or neo-functionalization contributing to the adaptive radiation of Acanthamoebae into new ecological niches.

Comparison of *Ac* with *Dd* highlights a broadly similar apparatus for environmental sensing and cell-cell communication and implies that the molecular elements underpinning the transition to a multicellular lifestyle may be widespread. Such transitions would likely have involved co-option of ancestral functions into multicellular programs and have occurred multiple times. Our analysis suggests that many signal processing and regulatory modules of higher animals and plants likely have deep origins and are balanced with subsequent losses in certain lineages including tyrosine kinases in fungi, plants and many protists.

The availability of an *Ac* genome offers the first opportunity to initiate functional genomics in this important constituent of a variety of ecosystems and should foster a better understanding of the amoebic lifestyle. Utilizing the genome as a basis for unraveling the molecular interactions between *Ac* and a variety of human pathogens will provide a platform for understanding the contributions of environmental hosts to the evolution of virulence.

Materials and methods

DNA isolation

Ac strain Neff (ATCC 30010) was grown at 30°C with moderate shaking to an OD_{550} of approximately 1.0. Total nucleic acid preparations were depleted of mitochondrial DNA contamination via differential centrifugation of cell extracts [91]. High molecular weight DNA was extracted from nuclear pellets either on Cesium chloride-Hoechst 33258 dye gradients as per [92] or by utilizing the Qiagen Genomic-tip 20/G kit (Qiagen, Hilden, Germany).

Genomic DNA library preparation and sequencing

All genomic DNA libraries were generated according to the Illumina protocol Genomic DNA Sample Prep Guide - Oligo Only Kit (1003492 A); sonication was substituted for the recommended nebulization as the method for DNA fragmentation utilising a Biorupter™ (Diagenode, Liége, Belgium). The library preparation methodology of end repair to create blunt ended fragments, addition of a 3'-A overhang for efficient adapter ligation, ligation of the adapters, and size selection of adapter ligated material was carried out using enzymes indicated in the protocol. Adapters and amplification primers were purchased from Illumina (Illumina, San Diego, CA, USA); both Single Read Adapters (FC-102-1003) and Paired End Adapters (catalogue number PE-102-1003) were used in library construction. All enzymes for library generation were purchased from New England Biolabs (Ipswitch, MA, USA). A limited 14-cycle amplification of size-selected libraries was carried out. To eliminate adapter-dimers, libraries were further sized selected on 2.5% TAE agarose gels. Purified libraries were quantified using a Qubit™ fluorometer (Invitrogen, Carlsbad, CA, USA) and a Quant-iT™ double-stranded DNA High-Sensitivity Assay Kit (Invitrogen). Clustering and sequencing of the material was carried out as per the manufacturer's instructions on the Illumina GAII platform in the UCD Conway Institute (UCD, Dublin, Ireland).

RNA extraction and RNA.seq library preparation and sequencing

For all tested conditions (Table S1.6.1 in Additional file 1) except the infection series, RNA was extracted from a minimum of 1×10^6 cells using TRIzol® (Invitrogen/Life Technologies, Paisley, UK). For infection material the detailed protocol is published in [93]. Strand-specific RNA.seq libraries were generated from total RNA using a modified version of [94] which is detailed in [93]. Briefly, total RNA was poly(A) selected, fragmented, reverse transcribed and second strand cDNA marked with the addition of dUTP. Standard Illumina methodology was followed - end-repair, A-addition, adapter ligation and library size selection - with the exception of the use of 'home-brew 6-nucleotide indexed' adapters as per Craig *et al*. [95]. Prior to limited amplification of the libraries, the dUTP marked second strand was removed via Uracil DNA-Glycosylase (Bioline, London, UK) digestion. Final libraries were quantified using the High Sensitivity DNA Quant-iT™ assay kit and Qubit™ Fluorometer (Invitrogen/Life Technologies). All sequencing was carried out in UCD Conway Institute on an Illumina GAII as per the manufacturer's instructions.

Sequencing and assembly

Genome assembly was carried out using a two-step process. Firstly, the Illumina reads were assembled using the Velvet [96] short read assembler to generate a series of contigs. These assembled contigs were used to generate a set of pseudo-reads 400 bp in length. These pseudo reads were then assembled in conjunction with the 454 FLX and Sanger sequences using version 2.3 of the GS De Novo Assembler using default parameters (Table S1.1.1 in Additional file 1). The assembly contained 45.1 Mb of scaffold sequence, of which 3.4 Mb (7.5%) represents gaps and 75% of the genome is contained in less than 100 scaffolds. For assembly statistics see Table S1.2.1 in Additional file 1. In order to determine the coverage of the transcriptome, we aligned our genome assembly to a publicly available EST dataset from GenBank (using the entrez query acanthamoeba EST) AND 'Acanthamoeba castellanii' [porgn:txid5755]). Of the 13,784 EST sequences downloaded, 12,975 (94%) map over 50% of their length with an average percent identity of 99.2% and 12,423 (90%) map over 70% of their length with an average percent identity of 99.26%.

Gene structure prediction

Gene finding was carried out on the largest 384 scaffolds of the *Ac* assembly using an iterative approach by firstly generating gene models directly from RNA.seq to train a gene-finding algorithm using a genome annotation pipeline followed by manual curation. Firstly, predicted transcripts were generated using RNA.seq data from a variety of conditions (Table S1.4.1 in Additional file 1) in conjunction with the G.Mo.R-Se algorithm (Gene Modelling using RNA.seq), an approach aimed at building gene models directly from RNA.seq data [97] running with default parameters. This algorithm generated 20,681 predicted transcripts. We then used these predicted transcripts to train the genefinder SNAP [98] using the MAKER genome annotation pipeline [99,100]. MAKER is used for the annotation of prokaryotic and eukaryotic genome projects. It identifies repeats, aligns ESTs (in this case the transcripts generated by the G.Mo.R-Se algorithm) and proteins from (nr) to a genome, produces *ab-initio* gene predictions and automatically synthesizes these data into gene annotations. The 17,013 gene predictions generated by MAKER were then manually annotated using the Apollo genome annotation curation tool [101,102]. Apollo allows the deletion of gene models, the creation of gene models from annotations and the editing of gene starts, stops, and 3' and 5' splice sites. Models were manually annotated examining a variety of evidence, including expressed sequence data and matches to protein databases (Section 1 of Additional file 1). Out of a total of 113,574 exons, 32,836 are exactly covered and 64,724 are partially covered by transcripts and 7,193 genes have at least 50% of their entire lengths covered by transcript data.

Functional annotation assignments

Functional annotation assignments were carried out using a combination of automated annotation as described previously [103] followed by manual annotation. Briefly, gene level searches were performed against protein, domain and profile databases, including JCVI in-house non-redundant protein databases, Uniref [104], Pfam [105], TIGRfam HMMs [106], Prosite [107], and InterPro [108]. After the working gene set had been assigned an informative name and a function, each name was manually curated and changed where it was felt a more accurate name could be applied. Predicted genes were classified using Gene Ontology (GO) [109]. GO assignments were attributed automatically, based on other assignments from closely related organisms using Pfam2GO, a tool that allows automatic mapping of Pfam hits to GO assignments.

Data access

This whole genome shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession AHJI00000000. The version described in this paper is the first version, AHJI01000000. The RNA.seq data are available under accessions SRA061350 and SRA061370-SRA061379.

Additional material

Additional file 1: Supplementary online material.

Additional file 2: Supplementary material supporting the LGT analysis.

Abbreviations

Ac: *Acanthamoeba castellanii*; bp: base pair; *Dd*: *Dictyostelium discoideum*; *Ed*: *Entamoeba dispar*; *Eh*: *Entamoeba histolytica*; EST: expressed sequence tag; GO: Gene Ontology; GPCR: G-protein-coupled receptor; HIF: hypoxia-inducible factor; H-NOX: heme-nitric oxide/oxygen binding; IAA: indole-3-acetic acid; LGT: lateral gene transfer; LRR: leucine-rich repeat; MAMP: microbe-associated molecular pattern; MAPK: mitogen-activated protein kinase; MBP: mannose binding protein; *Ng*: *Naegleria gruberi*; PRR: pattern-recognition receptor; PTK: tyrosine kinase 'writer'; PTPs: tyrosine phosphatase 'eraser'; pTyr: phosphotyrosine; RFX: Regulatory factor X; SH2: Src homology 2 'reader' domain; TKL: tyrosine kinase like.

Authors' contributions

Experiments were conceived and designed by MC, AJL, and BL. Analyses were carried out by all authors. Cell cultures of *A. castellanii* were grown and DNA isolated by AJL. DNA sequencing libraries were made and sequencing carried out by AJL. The manuscript was drafted by BL, with contributions from all authors. All authors read and approved the final manuscript for publication.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements
This work was funded by the Irish Science Foundation (SFI) grants 05/RP1/B908 and 05/RP1/908/EC07 awarded to BJL. The authors thank the Broad Institute and the investigators of 'the Origins of Multicellularity Sequencing Project, Broad Institute of Harvard and MIT' [110] for making data publicly available and the Liverpool Centre for Genomic Research for provision of 454 sequencing data. IL and MH were funded through grants from the Austrian Research Fund (Y277-B03), the European Research Council (Starting Grant EVOCHLAMY 281633), and the University of Vienna (Graduate School 'Symbiotic Interactions'). BT was supported by a grant from the Natural Sciences and Engineering Research Council of Canada (grant #184053). TRB is funded by the Centre for Biosciences. EST sequencing was supported by grants (CMRPD1A0581, CMRPG450131, and SMRPD160011) from Chang Gung Memorial Hospital to PT and CHC. JLM was supported by the Ramón y Cajal Subprogramme of the Spanish Ministry of Economy and Competitivity RYC-2011-08863.

Author details
[1]Conway Institute, University College Dublin, Belfield, Dublin 4, Ireland. [2]Samuel Lunenfeld Research Institute, Mount Sinai Hospital, 600 University Ave, Room 1081 Toronto, Ontario M5G 1X5, Canada. [3]Department für Mikrobielle Ökologie, Universität Wien, Althanstr. 14, A-1090 Wien, Austria. [4]Department of Biology, San Francisco State University, 1600 Holloway Ave, San Francisco, CA 94132, USA. [5]Bioinformatics Department, J Craig Venter Institute, Inc., 9704 Medical Center DriveRockville, MD 20850, USA. [6]Center for Research on Intracellular Bacteria, Institute of Microbiology, Institute of Microbiology, Rue du Bugnon 48, 1011 Lausanne, Switzerland. [7]College of Life Sciences, University of Dundee, Dow Street, Dundee DD1 5EH, UK. [8]Institute of Biology, Experimental Biophysics, Humboldt-Universität zu Berlin, Invalidenstrasse 42, D-10115 Berlin, Germany. [9]Department of Biosciences and Nutrition and Center for Biosciences, Karolinska Institutet, Hälsovägen 7, Novum, SE 141 83 Huddinge, Sweden. [10]Department of Molecular Biology and Biochemistry, Simon Fraser University, 8888 University Drive, Burnaby, BC V5A 1S6, Canada. [11]Department of Medicine, McGill University, McIntyre Medical Building, 3655 Sir William Osler, Montreal, Quebec H3G 1Y6, Canada. [12]Max Planck Institute for Developmental Biology, Spemannstr. 35 - 39, 72076 Tübingen, Germany. [13]Institut für Biologie II/Molecular Plant Physiology, Faculty of Biology, Albert-Ludwigs University of Freiburg, Freiburg, Germany. [14]Department of Plant Sciences, Britannia 04, Tel-Aviv University, Tel-Aviv 69978, Israel. [15]CeMM-Research Center for Molecular Medicine of the Austrian Academy of Sciences, Lazarettgasse 14, AKH BT 25.3, A-1090 Vienna, Austria. [16]World Premier International (WPI) Immunology Frontier Research Center (IFReC), Osaka University, 3-1 Yamadaoka, Suita, 565-0871 Osaka, Japan. [17]Department für Computational Systems Biology, Universität Wien, Althanstraße 14, 1090 Wien, Austria. [18]Department of Medical Genetics, Medical Genetics, C201 - 4500 Oak Street, Vancouver, BC, V6H 3N1, Canada. [19]Unité des rickettsies, IFR 48, CNRS-IRD UMR 6236, Faculté de médecine, Université de la Méditerranée, Marseille, France. [20]Banting and Best Department of Medical Research, Donnelly Centre, University of Toronto, 160 College Street, Room 230,Toronto, Ontario M5S 3E1, Canada. [21]Institute of Integrative Biology, Biosciences Building, University of Liverpool, Crown Street, Liverpool L69 7ZB, UK. [22]Department of Biology, NUI Maynooth, Co Kildare, Ireland. [23]University Institute of Tropical Diseases and Public Health of the Canary Islands, University of La Laguna, Avda. Astrofísico Fco. Sánchez, S/N 38203 La Laguna, Tenerife, Canary Islands, Spain. [24]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [25]Divisions of Pediatric Infectious Diseases, Department of Pediatrics, Chang Gung Children's Hospital and Chang Gung Memorial Hospital, Taoyuan, Taiwan. [26]Department of Parasitology, Chang Gung University, Taoyuan, Taiwan. [27]Ben May Department for Cancer Research and Committee on Cancer Biology, The University of Chicago, Chicago, IL 60637, USA. [28]Faculty of Health and Medicine, Division of Biomedical and Life Sciences, Lancaster University, Lancaster, LA1 4YQ, UK.

Received: 19 July 2012 Revised: 26 October 2012
Accepted: 1 February 2013 Published: 1 February 2013

References

1. De Jonckheere JF: **Ecology of Acanthamoeba.** *Rev Infect Dis* 1991, **13(Suppl 5)**:S385-387.
2. Rosenberg K, Bertaux J, Krome K, Hartmann A, Scheu S, Bonkowski M: **Soil amoebae rapidly change bacterial community composition in the rhizosphere of Arabidopsis thaliana.** *ISME J* 2009, **3**:675-684.
3. Nwachuku N, Gerba CP: **Health effects of Acanthamoeba spp. and its potential for waterborne transmission.** *Rev Environ Contam Toxicol* 2004, **180**:93-131.
4. Thomas V, McDonnell G, Denyer SP, Maillard JY: **Free-living amoebae and their intracellular pathogenic microorganisms: risks for water quality.** *FEMS Microbiol Rev* 2009.
5. Horn M, Wagner M: **Bacterial endosymbionts of free-living amoebae.** *J Eukaryot Microbiol* 2004, **51**:509-514.
6. Horn M: **Chlamydiae as symbionts in eukaryotes.** *Annu Rev Microbiol* 2008, **62**:113-131.
7. Greub G, Raoult D: **Microorganisms resistant to free-living amoebae.** *Clin Microbiol Rev* 2004, **17**:413-433.
8. Molmeret M, Horn M, Wagner M, Santic M, Abu Kwaik Y: **Amoebae as training grounds for intracellular bacterial pathogens.** *Appl Environ Microbiol* 2005, **71**:20-28.
9. Salah IB, Ghigo E, Drancourt M: **Free-living amoebae, a training field for macrophage resistance of mycobacteria.** *Clin Microbiol Infect* 2009, **15**:894-905.
10. Winiecka-Krusnell J, Linder E: **Free-living amoebae protecting Legionella in water: the tip of an iceberg?.** *Scand J Infect Dis* 1999, **31**:383-385.
11. Dallaire-Dufresne S, Paquet VE, Charette SJ: **[Dictyostelium discoideum: a model for the study of bacterial virulence].** *Can J Microbiol* 2011, **57**:699-707.
12. Sandström G, Saeed A, Abd H: **Acanthamoeba-bacteria: a model to study host interaction with human pathogens.** *Curr Drug Targets* 2011, **12**:936-941.
13. Eichinger L, Pachebat JA, Glockner G, Rajandream MA, Sucgang R, Berriman M, Song J, Olsen R, Szafranski K, Xu Q, Tunggal B, Kummerfeld S, Madera M, Konfortov BA, Rivero F, Bankier AT, Lehmann R, Hamlin N, Davies R, Gaudet P, Fey P, Pilcher K, Chen G, Saunders D, Sodergren E, Davis P, Kerhornou A, Nie X, Hall N, Anjard C, *et al*: **The genome of the social amoeba Dictyostelium discoideum.** *Nature* 2005, **435**:43-57.
14. Loftus B, Anderson I, Davies R, Alsmark UC, Samuelson J, Amedeo P, Roncaglia P, Berriman M, Hirt RP, Mann BJ, Nozaki T, Suh B, Pop M, Duchene M, Ackers J, Tannich E, Leippe M, Bruchhaus I, Willhoeft U, Bhattacharya A, Chillingworth T, Churcher C, Hance Z, Harris B, Harris D, Jagels K, Moule S, Mungall K, Ormond D, *et al*: **The genome of the protist parasite Entamoeba histolytica.** *Nature* 2005, **433**:865-868.
15. Bowers B, Korn ED: **The fine structure of Acanthamoeba castellanii (Neff strain). II. Encystment.** *J Cell Biol* 1969, **41**:786-805.
16. Marciano-Cabral F, Cabral G: **Acanthamoeba spp. as agents of disease in humans.** *Clin Microbiol Rev* 2003, **16**:273-307.
17. Pollard TD, Korn ED: **Acanthamoeba myosin. I. Isolation from Acanthamoeba castellanii of an enzyme similar to muscle myosin.** *J Biol Chem* 1973, **248**:4682-4690.
18. Ulsamer AG, Smith FR, Korn ED: **Lipids of Acanthamoeba castellanii. Composition and effects of phagocytosis on incorporation of radioactive precursors.** *J Cell Biol* 1969, **43**:105-114.
19. Keeling PJ, Palmer JD: **Horizontal gene transfer in eukaryotic evolution.** *Nat Rev Genet* 2008, **9**:605-618.
20. Merhej V, Notredame C, Royer-Carenzi M, Pontarotti P, Raoult D: **The rhizome of life: the sympatric Rickettsia felis paradigm demonstrates the random transfer of DNA sequences.** *Mol Biol Evol* 2011, **28**:3213-3223.
21. Thomas V, Greub G: **Amoeba/amoebal symbiont genetic transfers: lessons from giant virus neighbours.** *Intervirology* 2010, **53**:254-267.
22. Nelson WC, Bhaya D, Heidelberg JF: **Novel miniature transposable elements in thermophilic Synechococcus and their impact on an environmental population.** *J Bacteriol* 2012, **194**:3636-3642.
23. Andersson JO: **Gene transfer and diversification of microbial eukaryotes.** *Annu Rev Microbiol* 2009, **63**:177-193.
24. Lynch M, Conery JS: **The origins of genome complexity.** *Science* 2003, **302**:1401-1404.
25. Roy SW, Fedorov A, Gilbert W: **Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain.** *Proc Natl Acad Sci USA* 2003, **100**:7158-7162.
26. Li W, Tucker AE, Sung W, Thomas WK, Lynch M: **Extensive, recent intron gains in Daphnia populations.** *Science* 2009, **326**:1260-1262.
27. Roy SW: **Intron-rich ancestors.** *Trends Genet* 2006, **22**:468-471.

28. Roy SW, Irimia M, Penny D: **Very little intron gain in Entamoeba histolytica genes laterally transferred from prokaryotes.** *Mol Biol Evol* 2006, **23**:1824-1827.
29. Fredriksson R, Schiöth HB: **The repertoire of G-protein-coupled receptors in fully sequenced genomes.** *Mol Pharmacol* 2005, **67**:1414-1425.
30. Hoffman CS: **Glucose sensing via the protein kinase A pathway in Schizosaccharomyces pombe.** *Biochem Soc Trans* 2005, **33**:257-260.
31. Huang HC, Klein PS: **The Frizzled family: receptors for multiple signal transduction pathways.** *Genome Biol* 2004, **5**:234.
32. Gaillard I, Rouquier S, Giorgi D: **Olfactory receptors.** *Cell Mol Life Sci* 2004, **61**:456-469.
33. Iyer LM, Anantharaman V, Aravind L: **Ancient conserved domains shared by animal soluble guanylyl cyclases and bacterial signaling proteins.** *BMC Genomics* 2003, **4**:5.
34. Fitzpatrick DA, O'Halloran DM, Burnell AM: **Multiple lineage specific expansions within the guanylyl cyclase gene family.** *BMC Evol Biol* 2006, **6**:26.
35. Jekely G: **Evolution of phototaxis.** *Phil Trans R Soc Lond B Biol Sci* 2009, **364**:2795-2808.
36. Dudley R, Jarroll EL, Khan NA: **Carbohydrate analysis of Acanthamoeba castellanii.** *Exp Parasitol* 2009, **122**:338-343.
37. Segall JE, Kuspa A, Shaulsky G, Ecke M, Maeda M, Gaskins C, Firtel RA, Loomis WF: **A MAP kinase necessary for receptor-mediated activation of adenylyl cyclase in Dictyostelium.** *J Cell Biol* 1995, **128**:405-413.
38. Suga H, Dacre M, de Mendoza A, Shalchian-Tabrizi K, Manning G, Ruiz-Trillo I: **Genomic survey of premetazoans shows deep conservation of cytoplasmic tyrosine kinases and multiple radiations of receptor tyrosine kinases.** *Sci Signal* 2012, **5**:ra35.
39. Lim WA, Pawson T: **Phosphotyrosine signaling: evolving a new cellular communication system.** *Cell* 2010, **142**:661-667.
40. Liu BA, Shah E, Jablonowski K, Stergachis A, Engelmann B, Nash PD: **The SH2 domain-containing proteins in 21 species establish the provenance and scope of phosphotyrosine signaling in eukaryotes.** *Sci Signal* 2011, **4**: ra83.
41. Tan JL, Spudich JA: **Developmentally regulated protein-tyrosine kinase genes in Dictyostelium discoideum.** *Mol Cell Biol* 1990, **10**:3578-3583.
42. Tan CS, Pasculescu A, Lim WA, Pawson T, Bader GD, Linding R: **Positive selection of tyrosine loss in metazoan evolution.** *Science* 2009, **325**:1686-1688.
43. Li L, Tibiche C, Fu C, Kaneko T, Moran MF, Schiller MR, Li SS, Wang E: **The human phosphotyrosine signaling network: evolution and hotspots of hijacking in cancer.** *Genome Res* 2012, **22**:1222-1230.
44. Stuart LM, Ezekowitz RA: **Phagocytosis: elegant complexity.** *Immunity* 2005, **22**:539-550.
45. Watnick PI, Fullner KJ, Kolter R: **A role for the mannose-sensitive hemagglutinin in biofilm formation by Vibrio cholerae El Tor.** *J Bacteriol* 1999, **181**:3606-3609.
46. Beckmann G, Bork P: **An adhesive domain detected in functionally diverse receptors.** *Trends Biochem Sci* 1993, **18**:40-41.
47. Hong YC, Lee WM, Kong HH, Jeong HJ, Chung DI: **Molecular cloning and characterization of a cDNA encoding a laminin-binding protein (AhLBP) from Acanthamoeba healyi.** *Exp Parasitol* 2004, **106**:95-102.
48. Harpaz Y, Chothia C: **Many of the immunoglobulin superfamily domains in cell adhesion molecules and surface receptors belong to a new structural set which is close to that containing variable domains.** *J Mol Biol* 1994, **238**:528-539.
49. Ausubel FM: **Are innate immune signaling pathways in plants and animals conserved?.** *Nat Immunol* 2005, **6**:973-979.
50. Fujita T: **Evolution of the lectin-complement pathway and its role in innate immunity.** *Nat Rev Immunol* 2002, **2**:346-353.
51. Tissières P, Pugin J: **The role of MD-2 in the opsonophagocytosis of Gram-negative bacteria.** *Curr Opin Infect Dis* 2009, **22**:286-291.
52. Alsam S, Sissons J, Dudley R, Khan NA: **Mechanisms associated with Acanthamoeba castellanii (T4) phagocytosis.** *Parasitol Res* 2005, **96**:402-409.
53. Garate M, Cubillos I, Marchant J, Panjwani N: **Biochemical characterization and functional studies of Acanthamoeba mannose-binding protein.** *Infect Immun* 2005, **73**:5775-5781.
54. Watanabe Y, Tateno H, Nakamura-Tsuruta S, Kominami J, Hirabayashi J, Nakamura O, Watanabe T, Kamiya H, Naganuma T, Ogawa T, Naudé RJ, Muramoto K: **The function of rhamnose-binding lectin in innate immunity by restricted binding to Gb3.** *Dev Comp Immunol* 2009, **33**:187-197.
55. Hynes RO, Zhao Q: **The evolution of cell adhesion.** *J Cell Biol* 2000, **150**: F89-96.
56. Ghigo E, Kartenbeck J, Lien P, Pelkmans L, Capo C, Mege JL, Raoult D: **Ameobal pathogen Mimivirus infects macrophages through phagocytosis.** *PLoS Pathogens* 2008, **4**:e1000087.
57. Parakkottil Chothi M, Duncan GA, Armirotti A, Abergel C, Gurnon JR, Van Etten JL, Bernardi C, Damonte G, Tonetti M: **Identification of an L-rhamnose synthetic pathway in two nucleocytoplasmic large DNA viruses.** *J Virol* 2010, **84**:8829-8838.
58. Poole S, Firtel RA, Lamar E, Rowekamp W: **Sequence and expression of the discoidin I gene family in Dictyostelium discoideum.** *J Mol Biol* 1981, **153**:273-289.
59. Sanchez JF, Lescar J, Chazalet V, Audfray A, Gagnon J, Alvarez R, Breton C, Imberty A, Mitchell EP: **Biochemical and structural analysis of Helix pomatia agglutinin. A hexameric lectin with a novel fold.** *J Biol Chem* 2006, **281**:20171-20180.
60. Zambounis A, Elias M, Sterck L, Maumus F, Gachon CM: **Highly dynamic exon shuffling in candidate pathogen receptors... What if brown algae were capable of adaptive immunity?.** *Mol Biol Evol* 2012, **29**:1263-1276.
61. Koonin EV: **Taming of the shrewd: novel eukaryotic genes from RNA viruses.** *BMC Biol* 2010, **8**:2.
62. Aliyari R, Ding SW: **RNA-based viral immunity initiated by the Dicer family of host immune receptors.** *Immunol Rev* 2009, **227**:176-188.
63. Singh R, Jamieson A, Cresswell P: **GILT is a critical host factor for Listeria monocytogenes infection.** *Nature* 2008, **455**:1244-1247.
64. MacMicking JD: **Interferon-inducible effector mechanisms in cell-autonomous immunity.** *Nat Rev Immunol* 2012, **12**:367-382.
65. Peracino B, Wagner C, Balest A, Balbo A, Pergolizzi B, Noegel AA, Steinert M, Bozzaro S: **Function and mechanism of action of Dictyostelium Nramp1 (Slc11a1) in bacterial infection.** *Traffic* 2006, **7**:22-38.
66. Hug LA, Stechmann A, Roger AJ: **Phylogenetic distributions and histories of proteins involved in anaerobic pyruvate metabolism in eukaryotes.** *Mol Biol Evol* 2010, **27**:311-324.
67. Ginger ML, Fritz-Laylin LK, Fulton C, Cande WZ, Dawson SC: **Intermediary metabolism in protists: a sequence-based view of facultative anaerobic metabolism in evolutionarily diverse eukaryotes.** *Protist* 2010, **161**:642-671.
68. Slamovits CH, Keeling PJ: **Pyruvate-phosphate dikinase of oxymonads and parabasalia and the evolution of pyrophosphate-dependent glycolysis in anaerobic eukaryotes.** *Eukaryot Cell* 2006, **5**:148-154.
69. Loenarz C, Coleman ML, Boleininger A, Schierwater B, Holland PW, Ratcliffe PJ, Schofield CJ: **The hypoxia-inducible transcription factor pathway regulates oxygen sensing in the simplest animal, Trichoplax adhaerens.** *EMBO Rep* 2011, **12**:63-70.
70. Rytkönen KT, Storz JF: **Evolutionary origins of oxygen sensing in animals.** *EMBO Rep* 2011, **12**:3-4.
71. Sucgang R, Kuo A, Tian X, Salerno W, Parikh A, Feasley CL, Dalin E, Tu H, Huang E, Barry K, Lindquist E, Shapiro H, Bruce D, Schmutz J, Salamov A, Fey P, Gaudet P, Anjard C, Babu MM, Basu S, Bushmanova Y, van der Wel H, Katoh Kurasawa M, Dinh C, Coutinho PM, Saito T, Elias M, Schaap P, Kay RR, Henrissat B, *et al*: **Comparative genomics of the social amoebae Dictyostelium discoideum and Dictyostelium purpureum.** *Genome Biol* 2011, **12**:R20.
72. Duszenko M, Ginger ML, Brennand A, Gualdrón-López M, Colombo MI, Coombs GH, Coppens I, Jayabalasingham B, Langsley G, de Castro SL, Menna-Barreto R, Mottram JC, Navarro M, Rigden DJ, Romano PS, Stoka V, Turk B, Michels PA: **Autophagy in protists.** *Autophagy* 2011, **7**:127-158.
73. MacPherson S, Larochelle M, Turcotte B: **A fungal family of transcriptional regulators: the zinc cluster proteins.** *Microbiol Mol Biol Rev* 2006, **70**:583-604.
74. Bürglin TR: **Homeodomain subtypes and functional diversity.** *Sub-Cell Biochem* 2011, **52**:95-122.
75. Han Z, Firtel RA: **The homeobox-containing gene Wariai regulates anterior-posterior patterning and cell-type homeostasis in Dictyostelium.** *Development* 1998, **125**:313-325.
76. Piasecki BP, Burghoorn J, Swoboda P: **Regulatory Factor × (RFX)-mediated transcriptional rewiring of ciliary genes in animals.** *Proc Natl Acad Sci USA* 2010, **107**:12969-12974.
77. Krome K, Rosenberg K, Dickler C, Kreuzer K, Ludwig-Müller J, Ullrich-Eberius C, Scheu S, Bonkowski M: **Soil bacteria and protozoa affect root**

branching via effects on the auxin and cytokinin balance in plants. *Plant Soil* 2010, **328**:191-201.
78. Finkler A, Ashery-Padan R, Fromm H: **CAMTAs: calmodulin-binding transcription activators from plants to human.** *FEBS Lett* 2007, **581**:3893-3898.
79. Galon Y, Aloni R, Nachmias D, Snir O, Feldmesser E, Scrase-Field S, Boyce JM, Bouché N, Knight MR, Fromm H: **Calmodulin-binding transcription activator 1 mediates auxin signaling and responds to stresses in Arabidopsis.** *Planta* 2010, **232**:165-178.
80. Eichinger L, Noegel AA: **Comparative genomics of Dictyostelium discoideum and Entamoeba histolytica.** *Curr Opin Microbiol* 2005, **8**:606-611.
81. Song J, Xu Q, Olsen R, Loomis WF, Shaulsky G, Kuspa A, Sucgang R: **Comparing the Dictyostelium and Entamoeba genomes reveals an ancient split in the Conosa lineage.** *PLoS Comput Biol* 2005, **1**:e71.
82. Raoult D, Boyer M: **Amoebae as genitors and reservoirs of giant viruses.** *Intervirology* 2010, **53**:321-329.
83. Lurie-Weinberger MN, Gomez-Valero L, Merault N, Glöckner G, Buchrieser C, Gophna U: **The origins of eukaryotic-like proteins in Legionella pneumophila.** *Int J Med Microbiol* 2010, **300**:470-481.
84. Plate L, Marletta MA: **Nitric oxide modulates bacterial biofilm formation through a multicomponent cyclic-di-GMP signaling network.** *Mol Cell* 2012, **46**:449-460.
85. Martel CM: **Conceptual bases for prey biorecognition and feeding selectivity in the microplanktonic marine phagotroph Oxyrrhis marina.** *Microbial Ecol* 2009, **57**:589-597.
86. Anantharaman V, Iyer LM, Aravind L: **Comparative genomics of protists: new insights into the evolution of eukaryotic signal transduction and gene regulation.** *Annu Rev Microbiol* 2007, **61**:453-475.
87. Aderem A, Underhill DM: **Mechanisms of phagocytosis in macrophages.** *Annu Rev Immunol* 1999, **17**:593-623.
88. Boettner DR, Huston CD, Linford AS, Buss SN, Houpt E, Sherman NE, Petri WA Jr: **Entamoeba histolytica phagocytosis of human erythrocytes involves PATMK, a member of the transmembrane kinase family.** *PLoS Pathogens* 2008, **4**:e8.
89. Sun T, Kim L: **Tyrosine phosphorylation-mediated signaling pathways in dictyostelium.** *J Signal Transduction* 2011, **2011**:894351.
90. Turkarslan S, Reiss DJ, Gibbins G, Su WL, Pan M, Bare JC, Plaisier CL, Baliga NS: **Niche adaptation by expansion and reprogramming of general transcription factors.** *Mol Systems Biol* 2011, **7**:554.
91. Lohan AJ, Gray MW: **Analysis of 5′- or 3′-terminal tRNA editing: mitochondrial 5′ tRNA editing in Acanthamoeba castellanii as the exemplar.** *Methods Enzymol* 2007, **424**:223-242.
92. Spencer DF, Schnare MN, Gray MW: **Isolation of wheat mitochondrial DNA and RNA.** In *Modern Methods of Plant Analysis New Series* Edited by: Linskens HF, Jackson JF: Springer-Verlag, Berlin 1992, **14**:347-360.
93. Weissenmayer BA, Prendergast JG, Lohan AJ, Loftus BJ: **Sequencing illustrates the transcriptional response of Legionella pneumophila during infection and identifies seventy novel small non-coding RNAs.** *PloS One* 2011, **6**:e17570.
94. Parkhomchuk D, Borodina T, Amstislavskiy V, Banaru M, Hallen L, Krobitsch S, Lehrach H, Soldatov A: **Transcriptome analysis by strand-specific sequencing of complementary DNA.** *Nucleic Acids Res* 2009, **37**: e123.
95. Craig DW, Pearson JV, Szelinger S, Sekar A, Redman M, Corneveaux JJ, Pawlowski TL, Laub T, Nunn G, Stephan DA, Homer N, Huentelman MJ: **Identification of genetic variants using bar-coded multiplexed sequencing.** *Nat Methods* 2008, **5**:887-893.
96. Zerbino DR, Birney E: **Velvet: algorithms for de novo short read assembly using de Bruijn graphs.** *Genome Res* 2008, **18**:821-829.
97. Denoeud F, Aury JM, Da Silva C, Noel B, Rogier O, Delledonne M, Morgante M, Valle G, Wincker P, Scarpelli C, Jaillon O, Artiguenave F: **Annotating genomes with massive-scale RNA sequencing.** *Genome Biol* 2008, **9**:R175.
98. Korf I: **Gene finding in novel genomes.** *BMC Bioinformatics* 2004, **5**:59.
99. Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Sanchez Alvarado A, Yandell M: **MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes.** *Genome Res* 2008, **18**:188-196.
100. **MAKER..** [http://www.yandell-lab.org/software/maker.html].
101. Lewis SE, Searle SM, Harris N, Gibson M, Lyer V, Richter J, Wiel C, Bayraktaroglir L, Birney E, Crosby MA, Kaminker JS, Matthews BB, Prochnik SE, Smithy CD, Tupy JL, Rubin GM, Misra S, Mungall CJ, Clamp ME: **Apollo: a sequence annotation editor.** *Genome Biol* 2002, **3**:RESEARCH0082.
102. **Apollo Genome Annotation Curation Tool..** [http://apollo.berkeleybop.org/current/index.html].
103. Lorenzi HA, Puiu D, Miller JR, Brinkac LM, Amedeo P, Hall N, Caler EV: **New assembly, reannotation and analysis of the Entamoeba histolytica genome reveal new genomic features and protein content information.** *PLoS Negl Trop Dis* 2010, **4**:e716.
104. **Uniref..** [http://www.ebi.ac.uk/uniref/].
105. **Pfam..** [http://pfam.sanger.ac.uk/].
106. **TIGRFAMs..** [http://www.jcvi.org/cgi-bin/tigrfams/index.cgi].
107. **Prosite..** [http://prosite.expasy.org/].
108. **InterPro..** [http://www.ebi.ac.uk/interpro/].
109. Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, Binns D, Harte N, Lopez R, Apweiler R: **The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology.** *Nucleic Acids Res* 2004, **32**:D262-266.
110. **Broad Institute..** [http://www.broadinstitute.org/].
111. Rattei T, Tischler P, Gotz S, Jehl M-A, Hoser J, Arnold R, Conesa A, Mewes H-W: **SIMAP–a comprehensive database of pre-calculated protein sequence similarities, domains, annotations and clusters.** *Nucleic Acids Res* 2010, **38**:D223-226.
112. Frickey T, Lupas AN: **PhyloGenie: automated phylome generation and analysis.** *Nucleic Acids Res* 2004, **32**:5231-5238.
113. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: **MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods.** *Mol Biol Evol* 2011, **28**:2731-2739.

doi:10.1186/gb-2013-14-2-r11
Cite this article as: Clarke *et al.*: Genome of *Acanthamoeba castellanii* highlights extensive lateral gene transfer and early evolution of tyrosine kinase signaling. *Genome Biology* 2013 **14**:R11.

Chapter VII

Synthesis

Synthesis

The identification of a single protein that is uniquely shared among all sequenced members of the phyla *Planctomycetes*, *Verrucomicrobia, Chlamydia, Lentisphaera and Omnitrophica* (i.e the PVC superphylum), described in Chapter III, acts as a genetic link unifying those phyla. On the other hand, the absence of this signature protein (SP) homolog in the draft genomes of *Poribacteria* is arguing against the inclusion of this phylum within the PVC superphylum as it was initially postulated [1]. This suggestion was recently verified by systematic phylogenetic analysis of the phylum establishing the exclusion of *Poribacteria* from the superphylum [2]. Based on the accumulated evidence, the evolutionary origin of *Chlamydiae* is now conclusively placed within the PVC superphylum clade of bacteria although the exact evolutionary history and relation to the other bacterial phyla still remains to be determined.

The conservation level of the PVC SP is also suggesting a conserved function. Based on all available experimental and informatics data we hypothesize that the PVC signature protein represents an analog of the L30 ribosomal protein. Similar cases of non-homologous proteins that have converged to perform similar functions have been shown before but never for such a fundamental function as a ribosomal protein [3]. Targeted experiments like cellular fractioning, mass spectrometry or immunogold staining, are necessary to help verify this claim.

At a more theoretical level, concerning the hypothesized genetic link among *Planctomycetes* and eukaryotes, a PVC specific ribosomal SP would favor the eukaryogenesis fusion hypothesis postulated by Forterre [4]. According to this hypothesis in a rapid fusion event an ancient PVC bacterium merged with an archaeum resulting in a chimeric live form, the first eukaryote. In this fusion the PVC bacterium contributed the complex membrane system and metabolism and the archaeum the translation machinery. Following the fusion, the SP together with all the other ribosomal proteins of the PVC partner were replaced by their analogs or homologs from the archaeum leading to the absence of the signature (ribosomal) protein from the eukaryotes and the presence of the archaeal L30 instead.

This highly conserved nature of the SP across all members of the PVC superphylum was exploited for the screening of metagenomic data for the presence of PVC members. The successful extraction and phylogenetic reconstruction of signature protein homologs from available metagenomes supports the suitability of the SP for the taxonomic characterization of metagenomic fragments as a genomic marker. This search although resulted in an overall dramatic increase of available SP homologs, brought only few novel sequences that could be characterized as of chlamydial origin.

The low diversity of phylogenetically affirmed chlamydial SPs in environmental metagenomes was reconfirmed in a later study (Chapter IV) were all known chlamydial proteins were used for phylogeny based identification of chlamydial proteins in metagenomic datasets. In contrast to the low diversity of chlamydial proteins, the exploration of chlamydial diversity based on 16S rRNA amplicon studies resulted in the discovery of a massive, previously unrecognised, diversity of chlamydia. Using the most conservative estimations the sequences were supporting the existence of an at least 10 -20 times higher family level diversity within the phylum *Chlamydiae*.

Given that the depth of amplicon sequencing far exceeds that of metagenomic studies it can be anticipated that rare microbes would not be as adequately represented in microbial metagenomes as in 16S rRNA based amplicon studies. This logic is partly explaining the discrepancies among the detected chlamydial diversities within amplicon and metagenomic data as a result of low abundance of chlamydia in the investigated samples. Furthermore, especially since all known chlamydia are strict intracellular symbionts, sample preparation technics employing size fractionation effectively eliminate non-microscopic eukaryotic organisms from analyzed samples and together their associated chlamydia. Therefore the observed diversity from environmental samples is still a minor piece of the real diversity and it is likely to represent mostly chlamydia associated with microbial eukaryotes (e.g. amoebae or zooplankton) and only a fraction of the chlamydia in transmission among larger hosts (e.g. insects, fishes or mammals).

From the families formed using all the available genomic, metagenomic, and amplicon 16S rRNA sequences, the *Rhabdochlamydiaceae* was by far the one containing the

most unique sequences. The second largest family, in terms of diversity, was the *Parachlamydiaceae,* but it contained less than half of the sequences of the *Rhabdochlamydiaceae*. Considering the association of all known members of the *Rhabdochlamydiaceae* with arthropods and that of *Parachlamydiaceae* with protists there is a distinguishable pairing of the largest “animal” groups and their associated chlamydia families’ in terms of diversity. Given the evidences for host specificity, this pairing is supportive of a view anticipating a chlamydial counterpart for almost every eukaryotic species (plants excluded).

Concerning the origin of sequences, for both metagenomic proteins and 16S rRNA amplicons, the majority of supported novel chlamydial families contained sequences purely derived from marine environmental samples. This finding points towards the existence of an equally large number of marine eukaryotes that should act as chlamydia hosts. The amount of chlamydial novelty to be discovered in the oceans in combination with the global significance of saline ecosystems should attract researcher’s attention to unravel the full role of chlamydia in marine life.

Furthermore, the integrative methodology used in this study for the estimation of chlamydial diversity is by itself an innovative achievement. The described workflow was applied in other taxonomic groups of bacteria in order to extract relevant information about their abundance, distribution, and diversity, utilizing the accumulating wealth of data produced by next generation sequencing platforms. For example the environmental distribution and diversity of a bacterial symbiont infecting the nucleus of amoebae [5] or the global distribution of the thermophilic bacteria that were identified by an amplicon study [6] had been determined using the tools and approach developed in this study for the investigation of chlamydia diversity. To enhance the usability of data integration a fully automated web front for the query and analysis of all publically available 16S rRNA amplicon studies is under development.

In order to fully appreciate the evolutionary and ecological role of the predicted novel chlamydia diversity, genetic and experimental analysis of actual isolates is necessary. This wide gap between the numbers of chlamydial species supported by sequence evidence and the actual chlamydia isolates available illustrates the methodological

limitations in chlamydia isolation. Amoebae represent both important environmental reservoirs of chlamydia and also flexible and broadly susceptible co-cultivation hosts [7]. Therefore, optimization of the isolation and axenization of amoebae would augment the chlamydia isolation efforts. The proposed usage of hypersensitive *Escherichia coli* as a fully controllable food source for bacteriovoric eukaryotes (Chapter V) could serve to diminish the gap between the number of predicted chlamydia species and the available isolates. As it was shown in this study, *tolC* knockout *E. coli* performs significantly better than heat inactivated *E. coli* in the axenization of amoebae and it is now feasible to perform high-throughput screenings of environments samples using direct or co-cultivation approaches. This ability, when coupled with the knowledge of chlamydia diversity patterns by sequence evidences, should help in the targeted isolation of novel chlamydia from identified sources.

Besides the methodological innovation, the usage of Δ*tolC E. coli* already revealed significant insights into the chlamydia - amoeba association. The isolation of members of *Parachlamydiaceae* from every *Acanthamoeba* containing sample (in this study and other not published isolation attempts) strengthens the view of *Parachlamydiaceae* as a primarily protist symbiotic family. This inherent association, although persistent in the amoeba populations, lack the pattern of prevalence (100%) of a beneficial symbiosis. Although a slightly negative effect of chlamydia to infected *Acanthamoeba* has been described [8] the occurrence levels in all *Acanthamoeba* populations suggest a so far unclear occasionally beneficial role for the *Acanthamoeba* associated with chlamydia. Such a role could be, for example, a case of parasite mediated competition were an amoeba population with members harboring slightly parasitic chlamydia would benefit when competing against other more susceptible populations [9, 10]. Further investigations are needed in order to explore this intriguing role and the general prevalence patterns.

The high rate of isolates recovered from this study containing two taxonomically different symbionts was also a surprising outcome of the axenization method development. A widespread trilateral symbiosis puts one more dimension in the

discussion about the role of *Acanthamoeba* as a melting pot for lateral gene transfers among bacteria [11].

It has been proposed that the establishment of primary plastids and the emergence of plants involved the synchronous association of an early eukaryote with a chlamydium and a cyanobacterium [12, 13]. According to this hypothesis the chlamydium provided important molecular machinery (e.g. type III secretion system) and mediated the establishment of communication between the engulfed cyanobacterium and the early eukaryote. Given the observed abundance of trilateral symbiosis in present-day amoebae, the closest to the primordial eukaryote, this hypothesis sounds more plausible at least on the chance and therefore the possibility of such a co-association.

Although obtaining novel chlamydial isolates can enhance our knowledge (e.g. by genomic sequencing), it is the analysis of the interactions and responses within the host that would eventually shed light on their actual physiology. The genome of *Acanthamoeba castellanii* (chapter VI), a versatile natural host of many chlamydia, should provide the means for the successful analysis of the functional basis of symbiosis under different experimental conditions as already done for other symbiotic bacteria [14].

The analysis of the genome of *Acanthamoeba* itself gave important insights in the intriguing process of lateral gene transfer (LGT). Among others, the finding of an unprecedented number of LGTs in the genomes of *Acanthamoeba* and *Naegleria* was a striking result considering the low number observed in other protists and especially in multicellular organisms. Nevertheless, the analysis of putative donors did not show an enrichment of symbiotic bacteria origins for the LGTs. The majority of the LGTs show homology to proteins from bacteria in the specific environmental niche of each protist. For example the LGTs identified in the parasitic *Entamoeba* predominantly originate from gut associated bacteria. However, a substantial number of genes seem still to originate from environmentally separated niches. Therefore, we have to assume that an unknown LGT mechanism exists in *Acanthamoeba and*, in combination with the phagocytotic lifestyle, drives this enhanced genetic exchange. It is tempting to speculate that viral vectors mediate such transfers acting as bridges over the genetic

pools across non overlapping niches. Such view becomes more appealing especially under the light of the recent discoveries of the giant viruses and their chimeric genomic content [15, 16].

Overall, this work helped consolidate the indications for a common *Planctomycetes Verrucomicrobia* and *Chlamydiae* origin and shed more light into the size of the yet undiscovered members of the *Chlamydiae* phylum. The discovery of evidence for hundreds of putative novel chlamydia species and the realization that marine environments constitute a major chlamydial reservoir shaped new perceptions for the chlamydia research and broadened the field's boundaries. By applying novel isolation techniques like the Δ*tolC E. coli* we should now be able to probe this novel diversity and extract information about its ecological importance. Finally, with the published genome of the model chlamydia host *Acanthamoebae castellanii* further functional insights of the physiology of the chlamydia-host interactions can be revealed.

References

1. Wagner, M. and M. Horn, *The Planctomycetes, Verrucomicrobia, Chlamydiae and sister phyla comprise a superphylum with biotechnological and medical relevance.* Curr Opin Biotechnol, 2006. **17**(3): p. 241-9.
2. Kamke, J., et al., *The candidate phylum Poribacteria by single-cell genomics: new insights into phylogeny, cell-compartmentation, eukaryote-like repeat proteins, and other genomic features.* PLoS One, 2014. **9**(1): p. e87353.
3. Bork, P., C. Sander, and A. Valencia, *Convergent evolution of similar enzymatic function on different protein folds: the hexokinase, ribokinase, and galactokinase families of sugar kinases.* Protein Sci, 1993. **2**(1): p. 31-40.
4. Forterre, P., *A new fusion hypothesis for the origin of Eukarya: better than previous ones, but probably also wrong.* Res Microbiol, 2011. **162**(1): p. 77-91.
5. Schulz, F., et al., *Life in an unusual intracellular niche: a bacterial symbiont infecting the nucleus of amoebae.* ISME J, 2014.
6. Muller, A.L., et al., *Endospores of thermophilic bacteria as tracers of microbial dispersal by ocean currents.* ISME J, 2014. **8**(6): p. 1153-65.

7. Kebbi-Beghdadi, C. and G. Greub, *Importance of amoebae as a tool to isolate amoeba-resisting microorganisms and for their ecology and evolution: the Chlamydia paradigm.* Environ Microbiol Rep, 2014.
8. Collingro, A., et al., *Chlamydial endocytobionts of free-living amoebae differentially affect the growth rate of their hosts.* Eur J Protistol, 2004. **40**(1): p. 57-60.
9. Kuo, C.H., V. Corby-Harris, and D.E. Promislow, *The unavoidable costs and unexpected benefits of parasitism: population and metapopulation models of parasite-mediated competition.* J Theor Biol, 2008. **250**(2): p. 244-56.
10. Price, P.W., M. Westoby, and B. Rice, *Parasite-Mediated Competition: Some Predictions and Tests.* Am Nat, 1988. **131**(4): p. 544-555.
11. Moliner, C., P.E. Fournier, and D. Raoult, *Genome analysis of microorganisms living in amoebae reveals a melting pot of evolution.* FEMS Microbiol Rev, 2010. **34**(3): p. 281-294.
12. Subtil, A., A. Collingro, and M. Horn, *Tracing the primordial Chlamydiae: extinct parasites of plants?* Trends Plant Sci, 2014. **19**(1): p. 36-43.
13. Huang, J. and J. Gogarten, *Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids?* Genome Biol, 2007. **8**(6): p. R99.
14. Hoffmann, C., C.F. Harrison, and H. Hilbi, *The natural alternative: protozoa as cellular models for Legionella infection.* Cell Microbiol, 2014. **16**(1): p. 15-26.
15. Moreira, D. and C. Brochier-Armanet, *Giant viruses, giant chimeras: the multiple evolutionary histories of Mimivirus genes.* BMC Evol Biol, 2008. **8**: p. 12.
16. Yoosuf, N., et al., *Related Giant Viruses in Distant Locations and Different Habitats: Acanthamoeba polyphaga moumouvirus Represents a Third Lineage of the Mimiviridae That Is Close to the Megavirus Lineage.* Genome Biol Evol, 2012. **4**(12): p. 1324-1330.

Chapter VIII

Abstract

Abstract (English)

Chlamydiae represent a bacterial phylum with obligate intracellular members of important medical relevance. In order to investigate the evolutionary origin of the phylum, we used comparative genomics among all sequenced members of *Chlamydiae* and the proposed sister phyla *Planctomycetes*, *Verrucomicrobia*, *Lentisphaera* and *Omnitrophica* (PVC superphylum) to identify their genomic links. A single protein, unique and universally shared among all members of this superphylum, was identified and characterized. Based on the accumulated evidence, this protein is proposed to perform a function analogous to that of the L30 ribosomal protein, which is missing among all members of this superphylum. This protein proved to be a good taxonomic and phylogenetic marker as it was successfully used to extract PVC phyla diversity insights from all available metagenomic datasets. The diversity of the phylum *Chlamydiae* was revisited with the collection and analysis of all publicly available genomic and metagenomic data for chlamydia-like proteins or 16S rRNA genes sequences. In addition to the eight currently described families in the phylum, this analysis supports the existence of more than 200 unknown families, the majority of which contain sequences originating exclusively from marine ecosystems. In order to probe this yet undiscovered diversity of environmental chlamydia, a novel method for axenization of amoebae, which serve as natural hosts for chlamydiae, was developed. Using *ΔtolC Escherichia coli*, rendered hypersensitive to antibiotics, control and elimination of *E.coli* from amoeba cultures was achieved by the addition of otherwise sub-lethal amounts of antibiotics. Application of this approach resulted in improved axenization efficiency of amoebae and suggested that infection with more than one bacterial symbiont is more frequent than recognized previously. This observation led to the idea that amoeba hosts may be sites of increased genetic exchange. The possible effect of symbiosis-mediated lateral gene transfer (LGT) onto host amoeba, were thus investigated through the genomic analysis of the model host *Acanthamoeba castellanii*, as well as four other protists. Using a conservative, phylogeny-based approach it was shown that the *A. castellanii* and *Naegleria gruberi* genomes encode the highest number of LGTs (450 and 431 genes, respectively) attributed to prokaryotic origins. For

all analyzed amoeba, the closest phylogenetic neighbors of the identified HGTs, belong to microorganisms present at the respective ecological niche of each amoeba species, with a strong selection for genes encoding metabolic functions. Overall this work improved our evolutionary understanding of the phylum and provided additional means for the future analysis of its members.

Abstract (German)

Chlamydien sind eine Gruppe an intrazellulären Bakterien, die bedeutende Krankheitserreger von Mensch und Tier aber auch Symbionten von Einzellern umfasst. Die vergleichende Analyse der Genome der Chlamydien und ihren freilebenden Verwandten der Schwesterphyla *Planctomycetes*, *Verrucomicrobia*, *Lentisphaera* und *Omnitrophica* (PVC-Superphylum) erlaubte Einblicke in die frühe Evolutionsgeschichte dieser Mikroorganismen. Wir konnten ein Protein unbekannter Funktion identifizieren, das in allen Mitgliedern des PVC-Superphylums vorkommt – und nur dort. Die heterologe Expression dieses Proteins und funktionelle Assays legen den Schluss nahe, dass dieses Protein DNA und RNA bindet und möglicherweise Bestandteil des Ribosoms ist. Zudem eignet sich das Signaturprotein als phylogenetischer Marker zu Untersuchung der Diversität des PVC-Superphylums. Einen umfassenden Überblick über die Diversität und Verbreitung der Chlamydien konnten wir mithilfe eines innovativen Computergestützten Ansatzes gewinnen. Hierfür wurden zunächst verschiedene Sequenzdatenbanken integriert, die neben Genom- auch Metagenom- und Amplikonsequenzen enthalten. Die detaillierte Untersuchung der Chlamydien-ähnlichen 16S rRNS-Sequenzen in diesem Datensatz zeigte, dass es neben den bekannten acht Familien des Phylums *Chlamydiae*, mehr als 200 weitere Familien gibt, die vorwiegend in marinen Habitaten zu finden sind. Frei-lebende Amöben sind die natürlichen Wirte vieler Chlamydien in der Umwelt. In einem nächsten Schritt konnten wir ein Protokoll für die Isolierung und Axenisierung von Amöben entwickeln, das durch die Verwendung eines hypersensitiven *ΔtolC Escherichia coli* Stamms deutlich schneller

und effizienter als traditionelle Methoden ist. Diese Protokoll wird helfen, neue Chlamydien zu isolieren; erste Untersuchungen sind vielversprechend und zeigten, dass Amöben oft mit mehr als einer Bakterienart infiziert sind. Im Rahmen der Sequenzierung und Analyse des Genoms von *Acanthamoeba castellanii* haben wir den Einfluss der Interaktion dieser Amöben mit Bakterien – als Symbionten oder Nahrungsgrundlage – auf deren Genomevolution untersucht. Mithilfe eines phylogenomischen Ansatzes und konservativer phylogenetischer Analysen konnten wir zeigen, dass das Genom von *A. castellanii* in hohem Maß von horizontalem Gentransfer geprägt ist. Mehr als 400 Gene scheinen prokaryotischen Ursprungs zu sein. Der Vergleich mit den Genomen anderer Amöben zeigte, dass insbesondere Gene die eine Rolle im Stoffwechsel spielen betroffen sind. Zudem hat die ökologische Nische einen entscheidenden Einfluss auf die Herkunft der horizontal erworbenen Gene. Zusammengefasst führte diese Arbeit zu einem vertieften Verständnis der Evolution dieses Phylums und stellt zusätzlich neue Werkzeuge für die zukünftige Analyse seiner Mitglieder bereit.

Appendix I

Evaluation of interdomain lateral gene transfer in the genomes of *Acanthamoeba* and other protists

Introduction

Lateral gene transfer (LGT) has been well described in prokaryotes as an important factor for acquisition of novel genes and rapid evolution [1]. For eukaryotes the separation of somatic to gametic cells together with their highly different genetic organization form a barrier to the flow of genes among domains although some well documented cases of such transfers have been reported [2]. An exception to this rule is the unicellular protists were by default this barrier seizes to exist as all genetic modifications would pass to the daughter cell. Over the last years the progress in sequencing technologies enabled a detailed look over representative genomes of several unicellular protists that revealed numerous cases of laterally acquired genes [3-8]. The genome sequencing project of a phylogenetic different unicellular protist, the naked amoeba *Acanthamoeba castellanii*, enabled the investigation of the extent of interdomain LGT in this organism and the comparison with the so far known unicellular eukaryotes. In this work, we analyze the proteomes of *Acanthamoeba castellanii*, *Naegleria gruberi* [8], *Dictyostelium discoideum* [7], *Entamoeba histolytica* [9], *Entamoeba dispar*, *Micromonas* sp. [10], *Trypanosoma brucei* [6] and *Cryptosporidium parvum* [4] for the presence of genes with possible microbial origin. Additionally, in order to estimate the rate and extent of LGT events, the intergenic space of the *Acanthamoeba* genome was searched for remnants of past non-utilized foreign DNA.

Identification and analysis of LGT candidates

The complete predicted proteome of *Acanthamoeba castellanii* was imported to SIMAP (**SI**milarity **MA**trix of **P**roteins), a database of precalculated homologies of all against all proteins based on the FASTA algorithm [11]. For each protein, of all 8 investigated proteomes, the closest homologs that do not belong to the same family were extracted. This filter was implemented in order to avoid bias by families with multiple sequenced representatives that would otherwise mask LGT events that happen prior of the in-family speciation. Since the main scope of this analysis was the identification of the extent of LGT events in the investigated genomes, in the cases were multiple proteins from the same organism have as best homolog the same prokaryotic protein, only the representative with the best homology was kept. The rationale behind this selection was

that multiple homolog proteins in one genome are expected to rise from duplication rather than multiple LGT events. Therefore, in order not to overestimate the count of LGT events in the history of each organism we exclude all paralogs from the analysis. Finally, those proteins with a significant (E value < E^{-10}) best homology to a non-eukaryote, were selected for phylogenetic tree calculation and analysis.

The genome-wide screening of the investigated proteomes, revealed a plethora of candidates with best homologies to proteins of prokaryotic origin. Among the 8 analyzed unicellular organisms the genomes of *Acanthamoeba* and *Naegleria* contain more than half of all the identified prokaryotic-like genes with around 15% of their proteome falling into this category. From those genes the vast majority show highest similarity to bacterial homologs, leaving an accumulative percentage of less than 1% for the archaeal and viral like genes.

The analysis of the taxonomic distribution of the closest homologs of all the candidate proteins did not reveal any enrichment of symbiotic/intracellular organisms nor of any particular taxonomic family but rather a wide range of different phyla. In addition, the distributions of sequence similarity values to the identified bacterial homologs follow a normal distribution with the bell peak around 65% protein similarity for all organisms (Figure 1). Taken together, the absence of an overrepresented non-eukaryotic organism and the distribution and level of sequence similarity observed vote against a contamination scenario as the origin of the found homologies.

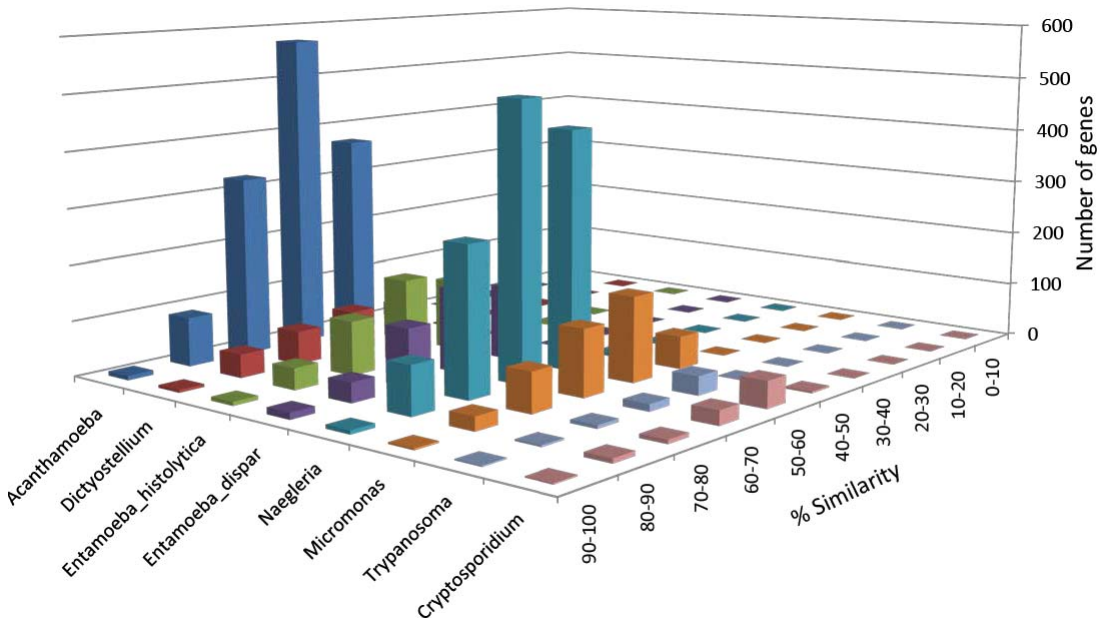

Figure 1: Similarity plot of all bacteria like proteins to their best bacterial homolog.

Calculation and filtering of phylogenetic protein trees

The software package PhyloGenie [12] was used for the calculation and analysis of protein phylogenetic trees. In short, PhyloGenie uses a protein sequence input (seed) to extract homologs by PSI BLAST [13] searches that are then used to construct a multiple sequence alignment. This alignment is then used to create a tree based on the method of choice. In this analysis the minimum of 5 homologs was used as a cut off for the creation of an alignment and the trees were calculated using a Maximum Likelihood [14] approach with 100 bootstraps [15].

All the calculated trees were filtered using PHAT (included in the PhyloGenie package) for nodes, which contain the seed leaf together with bacterial, archaeal or viral leafs and no more than two other unicellular eukaryotic organisms, with bootstrap support above 75%. The tolerance for eukaryotic leafs was implemented because it was observed that in many cases the seed proteins were grouping together with other unicellular organism, that are very distant phylogenetically, inside otherwise prokaryotic clades. The selected trees were further inspected manually, and the final selection was made based on the trees topologies.

From the total set of 4713 candidate proteins, 3476 phylogenetic trees were calculated and after filtering and manual inspection, 1504 trees remained that support a LGT scenario (Table 1).

Species	Candidates (SIMAP)	Calculated (PhyloGenie)	Quality filtering	Manual curation
Cryptosporidium	105	46	15	10
Dictyostelium	277	203	106	91
E. dispar	488	407	229	195
E. histolytica	469	383	215	199
Micromonas	451	251	135	117
Naegleria	1379	1078	506	431
Trypanosoma	65	49	14	11
Acanthamoeba	1479	1059	477	450
All	**4713**	**3476**	**1697**	**1504**

Table 1: Numbers of putative and phylogenetically supported LGT proteins.

The selected proteins with trees supportive of a lateral gene transfer scenario were compared with the published lists of putatively LGT-derived genes when available. In the publication of the *Naegleria* genome [8] 184 genes were reported having high homology to bacteria but no homology to eukaryotes. Using phylogenetic trees analyses, they reported that 44 of these cases have trees with good bootstrap support (>75%) and are candidates for prokaryotic LGT origin. The fate of this set of 44 published genes in our LGT identification pipeline was followed in order to estimate the efficiency of the process.

In the selection of candidates by homology using SIMAP, 42 out of the 44 proteins were found to still have good homology to bacteria while for two the best hit is now eukaryotic. From the 42 candidates submitted for tree calculation, 33 have trees calculated. The 9 missing trees fail to pass the defined PhyloGenie selection criteria, namely the low coverage to the query protein (<50%) or the low number (< 5) of the available homologues. From the 33 calculated trees 31 have topologies supportive of

LGT while two were considered negative. Overall 70% of the reported candidates from that publication were recovered while 30% was excluded based on our criteria.

There are 96 published LGT candidate trees for *Entamoeba histolytica* [16] and 78 of them pass the homology based selection step. The 18 missing genes are due to annotation changes in the *Entamoeba* genome that render some proteins obsolete or difficult to map back to Genebank accession numbers. Only two were excluded in the analyses based on the tree topology (76 positive) with the reason being our conservative approach concerning the classification of the trees. Therefore almost all of the previously proposed LGT candidates from *Entamoeba histolytica* were rediscovered in this analysis.

The genome of *Cryptosporidium pavum* [4] has been reported to contain 24 genes that appeared to be xenologs with prokaryotic origin. Following the procedure already described only 7 of this set were rediscovered. The rest had either best homology to other protists and not included in the analysis or they had tree topologies that was not favoring a LGT scenario. Nevertheless, 3 previously not detected genes show consistent LGT supportive tree topologies. Therefore, our findings suggest that the genome of *Cryptosporidium pavum* had assimilated only 10 genes of prokaryotic origin.

For the case of *Dictyostelium discoideum* [7] the situation was similar. Out of the 18 reported LGT candidates, from the genome paper of *Dictyostelium*, 11 had best homology to Eukaryotes and only 7 had best homology to bacteria. From those only 4 pass the pipeline to the final list of LGT candidates. In this point it has to be mentioned that almost all of the 11 genes not included due to Eukaryotic homology are most similar to other sequenced dictyostelids like *Polysphondylium pallidum*. Nevertheless, besides the 4 previously known LGT candidates, 87 novel candidates were suggested by this analysis. The presence of a sequenced genome of an organism related to *Dictyostelium*, due to the design of the study, retrieve only those LGT events that happen after the divergence of these two organisms. Older LGT events would not be

detected and it should be kept in mind when the occurrences of LGTs among the organisms are compared.

Identification and analysis of putative donors

From all the manually verified trees, the closest protein leaf with non-eukaryotic origin was extracted for each seed protein. The proximity was determined out of the sum of edge distances of each protein from the seed tree leaf. Based on the assumption that, the phylogenetically closest protein carries taxonomic information similar to the most likely donor of the LGT event, the collected sets of putative donors was collected and analyzed. Two approaches were followed for the analysis of the LGT donors.

In the first approach the taxonomic information was extracted for the phylum and class level and the results were used for the estimation of the contribution of each phylum or class in the total number of LGTs (including viruses). These data were then used in order to calculate the Bray-Curtis [17] similarities among the analyzed organisms. The similarities tables were then visualized as heat maps using the Java application JColorGrid [18]. For the second approach environmental information were collected for each donor species using primarily the information available at NCBI page for sequenced prokaryotic genomes (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi) and literature research for those missing.

Data were collected for the oxygen requirements (aerobic, facultative and anaerobic), the preferred habitat (host associated or environmental), the environmental preferences (aquatic, terrestrial, multiple or special) and the pathogenicity or not of the host associated donor species. Especially *for Entamoeba dispar* and *Entamoeba histolytica*, the lists of LGTs were compared and the proteins that are uniquely present in one of them were identified. The analysis of the environmental characteristics of the proposed donors for these proteins was then repeated for the unique proteins donors.

Due to the nature of the analysis the viral donors were excluded. In addition, given the low number of LGT identified for *Cryptosporidium* and *Trypanosoma*, the results of the

environmental analysis for them were not shown, as they are not statistically significant and are prone to biases and misinterpretation.

The taxonomic analysis of the proposed donors of the LGTs showed that the phylum of *Proteobacteria* contribute the most for all the organisms besides the *Entamoeba* sp. where the intestinal dominant phyla of *Firmicutes* and *Bacteroidetes* are the most prevalent donors. The Bray-Curtis analysis of the phyla and classes contributions as LGT donors showed that *Acanthamoeba* and *Naegleria* have similar taxonomic distributions at these levels. Besides the expected clustering of the two *Entamoeba* species weak but noticeable clustering was observed for the two *Apicomplexa* species and for *Dictyostelium* and *Micromonas* (Figure 2A).

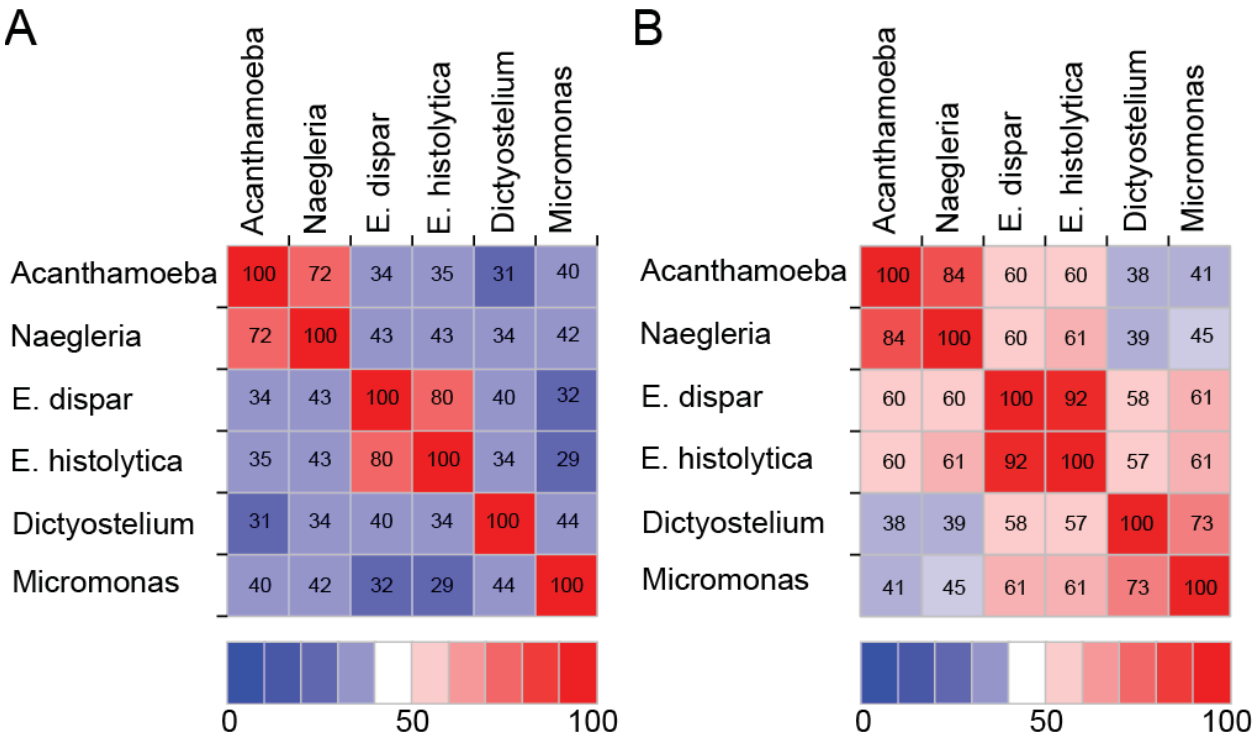

Figure 2: Similarity of analyzed protists based on their LGT genes. A) Bray-Curtis similarity matrix based on the taxonomic assignment of the putative donors (class level) of the LGT genes. B) Bray-Curtis similarity of the functional profiles of the proteins from predicted LGTs. The general functional classification of the COGs that each protein was assigned to was used for the similarity calculations.

The analysis of the habitat and oxygen requirement of the proposed donors of the LGTs showed that the two *Entamoeba* species have 3 times more anaerobic than aerobic donors when for the other organisms the opposite pattern is observed (Figure 3A). In addition half of the *Entamoeba* sp. donors are associated with a host, meaning that their habitat is a human, an animal or a cellular organism in general (Figure 3B). Half of these donor species are pathogenic for their hosts. Interestingly, the distribution of habitats of the environmental donors is similar for all organisms with around 20% originating from specialized habitats like hydrothermal vents (marine and terrestrial) and acid mines drainages.

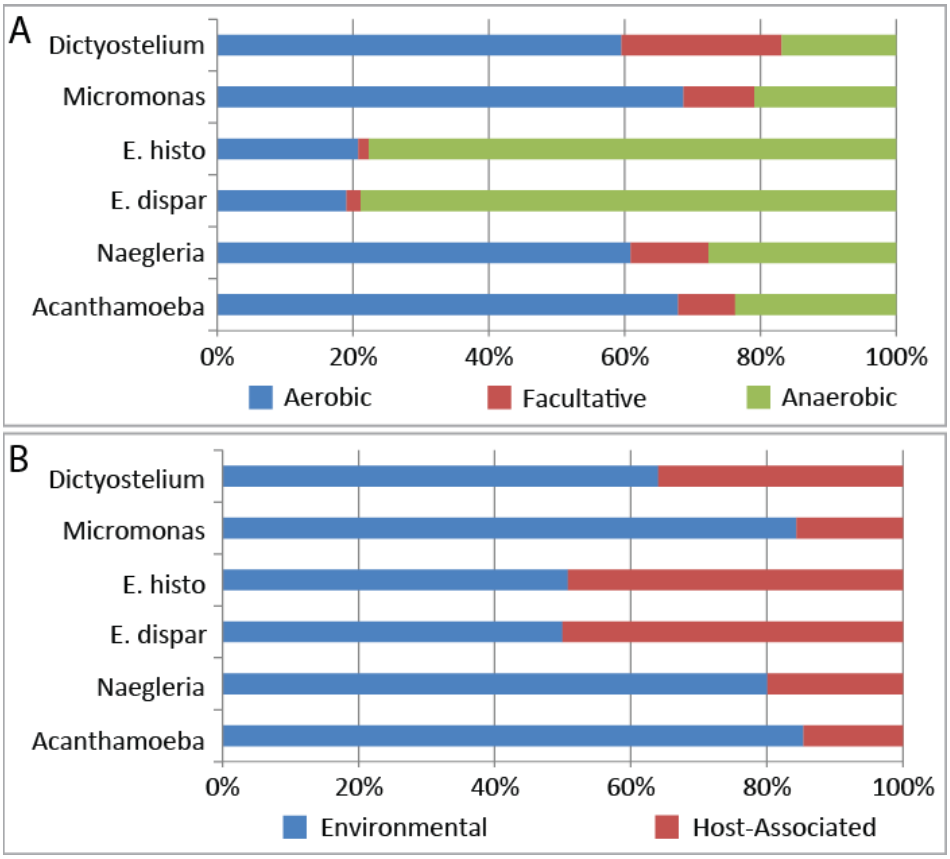

Figure 3: Assessment of protists LGT candidate donors. A) Oxygen requirements of LGT candidate donors of various protists. B) Host association classification of candidate LGT donors of various protists.

Functional Analysis of the candidate LGT acquired proteins

The list of candidate LGT proteins were assigned to Clusters of Orthologus Groups of proteins (COGs) based on the COG assignment of their most similar protein in SIMAP [11]. In the same way the closest proteins (see Identification and analysis of the most likely donors) was also assign to COGs. Since in most cases the COG assignments for the candidate LGTs and the proposed donors were overlapping, the rest of the analysis was based on the COG assignments of the candidate LGTs only.

For each assigned COG the functional description[19] was extracted from the eggNog database [20] . Given that each COG can have multiple functional assignments and that the purposes of this analysis is the identification of the coverage and the extent of laterally imported functions, when the presence of functions was counted the COGs with multiple functional assignments were counted multiple times. Similarly to the putative donor characteristics analysis, the results for *Cryptosporidium* and *Trypanosoma* were not presented as they are not statistically significant.

When the functional profiles of the LGTs are used to calculate a Bray-Curtis similarity among the analyzed protists, a similar pattern emerges as with the analysis of the donor phylum taxonomy. A relation of Acanthamoeba and *Naegleria* emerge as well as the expected similarity among the two *Entamoebas*. The pairs of *Dictyostelium* and *Micromonas* and the two *Apicomplexa* also reveal a higher similarity than to the rest of the organisms in the analysis (Figure 2B).

Concerning the actual function assignments, the most prominent functional categories were the “General function prediction only” and “Function unknown” belonging to the supergroup of “Poorly Characterized Proteins” followed by “Energy production and conversion”, “Carbohydrate transport and metabolism” and “Amino acid transport and metabolism” that belong to the supergroup of “Metabolism”. This trend was better visualized by supergroup level classification comparison that makes clear that the groups of “Poorly Characterized Proteins” and “Metabolism” have each three times

more proteins assigned to them than the groups “Information Storage and Processing” and “Cellular Processes and Signaling” (Figure 4). Nevertheless, it’s noticeable that the underrepresented groups have some functional categories that are assigned to similar number of proteins for different organisms.

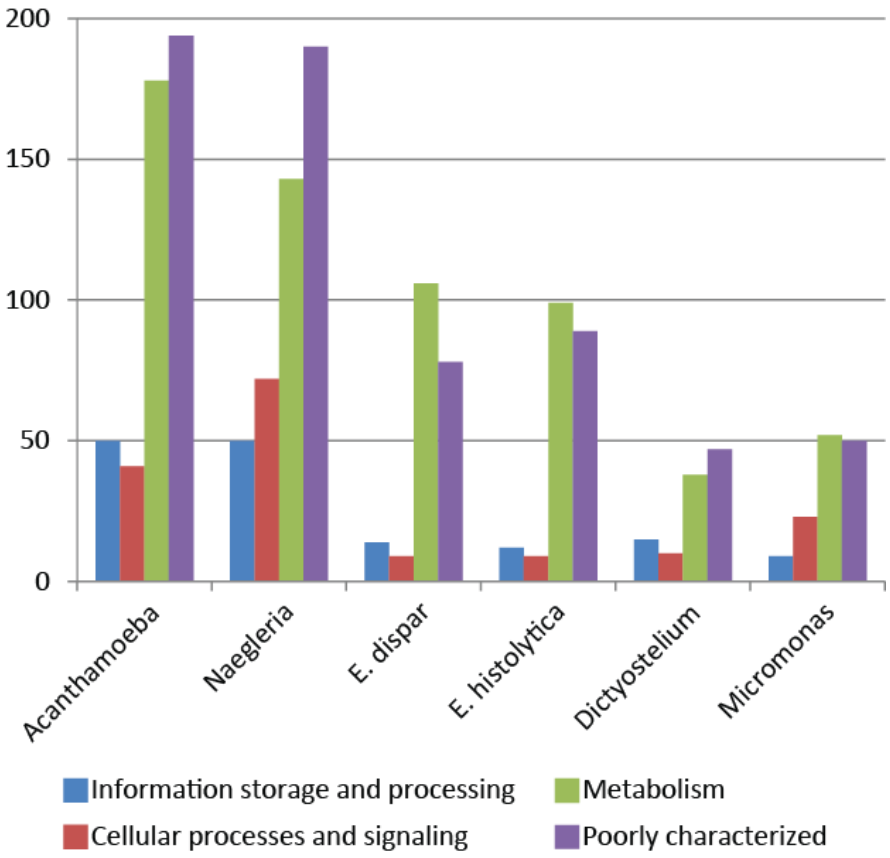

Figure 4: Functional categories of the proteins coded by the predicted LGT genes among different protists.

Search in intergenic space of *Acanthamoeba*

The intergenic spaces (IGSs) of the *Acanthamoeba* genome were used in two kinds of homology search approaches in order to identify regions with significant homology to genes that can be attributed to remains of LGTs. Firstly the IGSs were blasted (blastX) against NCBI (nr) and secondly the IGSs were used to construct a local database and then the UniRef50 protein set was used as tBlastn queries against it. For both analyses for shake of simplicity only the best pair of IGS and protein hit was considered, although

meaningful multiple hits can be explained as each IGS can contain multiple sites homologous to different proteins. To increase the chance to detect these regions the threshold for both homology searches was set to 10E-4. The combined non redundant set of IGSs from both approaches was filtered for size, bit score and Expected value. The cut off values used were >50 amino acids, >50 bit score and <10-5 Expected value respectively.

From the 10641 IGSs of *Acanthamoeba* genome, 8578 have a hit to a protein in NCBIs (nr) database using Blastx and from those for 847 it is significant (E-value < 10-4). The tBlastn of Uniref50 against the IGS database result in 1498 IGS with significant homology to Uniref50 proteins. From both approaches combined 1572 unique cases of IGS with significant hit to proteins was found. Their alignment length plot shows a bell-like distribution with a peak around 50 bp and a skewed right flank. In order to construct a set of more robust cases, this list was further filtered for alignment length (>50), bitscore (>50) and Evalue ($<10^{-5}$) ending up on a set 516 cases of relative good homologies. Overall, depending on the approach used, 5-14% of the *Acanthamoeba* intergenic spaces contain remnants of possibly LGT transferred genes with 10-20% of these being most similar to prokaryotic homologs (Table 2).

	Blastx	tBlastn	Combined	Filtered
Eukaryotic	730	1153	1211	447
Bacterial	93	309	314	52
Archaeal	5	10	10	5
Viral	16	26	29	9
Total	844	1498	1564	513

Table 2: The number of intergenic spaces in the genome of *Acanthamoeba* containing detectable foreign gene fragments.

Intronization and expression of the candidate LGTs in *Acanthamoeba*

The number of introns for each LGT protein was inferred based on the predicted number of exons (Introns = Exons - 1). For the estimation of the expression level, since

multiple experiments were available, the average value of expression among these experiments was taken for each LGT protein (Figure 5). The average number of introns for the 405* LGTs of *Acanthamoeba* with predicted exons is 6 with only 32 cases with no predicted introns. The expression of the proteins follows a power low distribution with a high number of proteins having very low expression and a very low number of proteins having a very high expression level. Overall the median expression of the LGT genes was the same with that of the non LGT genes.

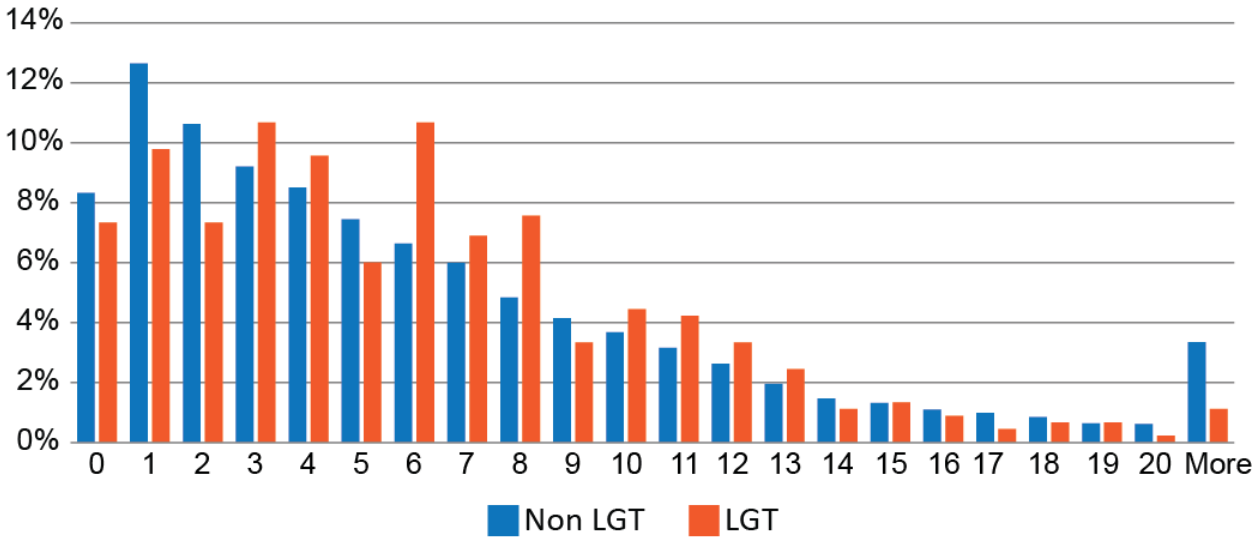

Figure 5: Relative distribution of the intronization levels across the predicted LGT genes and the non LGT genes in *Acanthamoeba* genome.

Discussion

Unicellular protists genomes contain many proteins with non-eukaryotic homologs. For a third of these proteins, the phylogenetic analysis verify their affiliation to bacterial, archaeal or viral proteins and the topologies of the constructed trees strongly support a LGT scenario for their acquisition. Especially for *Acanthamoeba*, the intronization and expression levels of these genes indicate that they are fully integrated in their host biology.

The amount of detected LGTs in the genomes of the bacteriovoric *Acanthamoeba* and *Naegleria* was two times bigger than the parasitic *Entamoeba*, four times bigger than

Micromonas and *Dictyostelium* and 40 times higher than those detected in *Apicomplexa*. Therefore, as expected, bacteriovoric amoebas tend to acquire higher number of proteins laterally due to their diet on prokaryotes. Although the predatory life style of *Acanthamoeba* and *Naegleria* explains the higher number of LGTs in these organisms, life style itself is not sufficient to explain the presence of prokaryotic origin LGT to the green algae *Micromonas* and the eukaryotic parasites.

The methodology followed for the detection of possible LGT proteins in the genomes of the 8 analyzed unicellular protists succeed in recovering previously published cases and even increased the total number of candidates. Besides the inevitable differences arising from the expansion of the available genomic information in public databases, the methodological difference that explains a major part of the increased number of detected LGTs is the consideration of homologous replacement. In previous studies, the screenings of eukaryotic proteins for indications of prokaryotic origin, they exclude those with significant homology to eukaryotes. In this study, the existence of best homology to bacteria archaea or viruses was sufficient to include the protein in the pipeline of phylogenetic analysis. The cases of homologous replacement can then be identified by the inspection of the phylogenetic trees were the LGT protein cluster for example in a bacterial clade and away from the eukaryotic clade.

The study of the taxonomy of the putative donors of the verified LGTs show that *Acanthamoeba* and *Naegleria* have acquired proteins from similar donors (at phylum and Class levels) as also the two *Entamoeba*, and the two *Apicomplexa*. The analysis of the oxygen requirements of the donors showed that the anaerobic *Entamoeba* contains significantly more proteins from donors with anaerobic life style and that 50% of the donors are associated with a host as their proliferation habitat when for the other amoebas its around anaerobic donors are only about 20%. This finding supports the hypothesis that the unicellular protists are mostly assimilating genes from donors that they encounter in their environments. Nevertheless, the high number of cases (30%) of free living environmental donors that have specialized habitats, like hydrothermal vents,

and should not be accessible by the protists, indicate the presence of an unknown vector that shuttle these genes to the protists.

The functional classification of the LGT acquired proteins revealed that what the organisms assimilate are at large related to metabolism. It seems that irrespectively of where they live and what they eat, at the end these organisms acquire similar functions. The analysis of the intergenic space of Acanthamoeba revealed that a high number of remnants of LGT genes can be detected among the predicted genes. This amount of degrading foreign DNA supports the observed number of assimilated LGT and explains the fate of those genes without functional importance for the host.

Overall, it is clear that unicellular protists contain a high number of proteins that are more likely of bacterial, archaeal or viral origin. Although the diet and the habitat of the organism can explain some of these acquisitions, only a gene shuttling vector, different than the proposed donors can account for the rest of the LGTs. It has been proposed that giant viruses can fulfill this role [21, 22]. Although we find this an attractive idea, due to the limited sampling of giant virus genomes and the expected high deletion rate of non-utilized genomic fragments in the virus genomes we could not sufficiently investigate and verify this claim. Going beyond establishing the importance of LGTs in the shaping and evolution of the unicellular protist genomes, the new challenge is the elucidation of the underlying mechanism that makes this surprising rate of LGTs possible.

References

1. Ochman, H., J.G. Lawrence, and E.A. Groisman, *Lateral gene transfer and the nature of bacterial innovation.* Nature, 2000. **405**(6784): p. 299-304.
2. Andersson, J.O., *Lateral gene transfer in eukaryotes.* Cellular and Molecular Life Sciences, 2005. **62**(11): p. 1182-1197.
3. Gardner, M.J., et al., *Genome sequence of the human malaria parasite Plasmodium falciparum.* Nature, 2002. **419**(6906): p. 498-511.

4. Abrahamsen, M.S., et al., *Complete genome sequence of the apicomplexan, Cryptosporidium parvum.* Science, 2004. **304**(5669): p. 441-445.
5. Huang, J.L., et al., *Phylogenomic evidence supports past endosymbiosis, intracellular and horizontal gene transfer in Cryptosporidium parvum.* Genome Biology, 2004. **5**(11).
6. Berriman, M., et al., *The genome of the African trypanosome Trypanosoma brucei.* Science, 2005. **309**(5733): p. 416-22.
7. Eichinger, L., et al., *The genome of the social amoeba Dictyostelium discoideum.* Nature, 2005. **435**(7038): p. 43-57.
8. Fritz-Laylin, L.K., et al., *The genome of Naegleria gruberi illuminates early eukaryotic versatility.* Cell, 2010. **140**(5): p. 631-42.
9. Loftus, B., et al., *The genome of the protist parasite Entamoeba histolytica.* Nature, 2005. **433**(7028): p. 865-8.
10. Worden, A.Z., et al., *Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes Micromonas.* Science, 2009. **324**(5924): p. 268-72.
11. Rattei, T., et al., *SIMAP--a comprehensive database of pre-calculated protein sequence similarities, domains, annotations and clusters.* Nucleic Acids Res, 2010. **38**(Database issue): p. D223-6.
12. Frickey, T. and A.N. Lupas, *PhyloGenie: automated phylome generation and analysis.* Nucleic Acids Res, 2004. **32**(17): p. 5231-8.
13. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: A new generation of protein database search programs.* Nucleic Acids Research, 1997. **25**(17): p. 3389-3402.
14. Stamatakis, A., *RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models.* Bioinformatics, 2006. **22**(21): p. 2688-2690.
15. Pattengale, N.D., et al., *How many bootstrap replicates are necessary?* Journal of Computational Biology, 2010. **17**(3): p. 337-354.
16. Loftus, B., et al., *The genome of the protist parasite Entamoeba histolytica.* Nature, 2005. **433**: p. 865 - 868.

17. Bray, J.R. and J.T. Curtis, *An Ordination of the Upland Forest Communities of Southern Wisconsin.* Ecological Monographs, 1957. **27**(4): p. 326-349.
18. Joachimiak, M.P., J.L. Weisman, and B.C.H. May, *JColorGrid: software for the visualization of biological measurement.* Bmc Bioinformatics, 2006. **7**.
19. Tatusov, R.L., E.V. Koonin, and D.J. Lipman, *A genomic perspective on protein families.* Science, 1997. **278**(5338): p. 631-637.
20. Bork, P., et al., *eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations.* Nucleic Acids Research, 2010. **38**: p. D190-D195.
21. Filee, J. and M. Chandler, *Gene Exchange and the Origin of Giant Viruses.* Intervirology, 2010. **53**(5): p. 354-361.
22. Raoult, D. and P. Colson, *Gene Repertoire of Amoeba-Associated Giant Viruses.* Intervirology, 2010. **53**(5): p. 330-343.

Appendix II

Scientific experience

Acknowledgements

Curriculum vitae

Scientific Experience

Publications list

1. Clarke M., Lohan A. J., Liu B., **Lagkouvardos I.**, Roy S., Zafar N. et al. Genome of *Acanthamoeba castellanii* highlights extensive lateral gene transfer and early evolution of pattern recognition and tyrosine kinase signaling. **Genome Biol** 14(2), R11 (2013)

2. Dolinšek J., Dorninger C., **Lagkouvardos I.**, Wagner M., Daims H. Depletion of Unwanted Nucleic Acid Templates by Selective Cleavage: LNAzymes Open a New Window for Detecting Rare Microbial Community Members. **Appl Environ Microbiol** 79(5), 1534-1544 (2013)

3. Dolinšek J., **Lagkouvardos I.**, Wanek W., Wagner M., Daims H. Interactions of Nitrifying Bacteria and Heterotrophs: Identification of a Micavibrio-like, Putative Predator of Nitrospira. **Appl Environ Microbiol** 79 (6), 2027-2037 (2013)

4. **Lagkouvardos I.**, Jehl M.-A. Rattei T., Horn M. Signature Protein of the PVC Superphylum. **Appl Environ Microbiol** 80 (2), 440-5 (2013)

5. **Lagkouvardos I.**, Weinmaier T., Lauro FM., Cavicchioli R., Rattei T., Horn M. Integrating metagenomic and amplicon databases to resolve the phylogenetic and ecological diversity of the Chlamydiae. **ISME J** 8(1), 115–125 (2014)

6. Müller A., Rezende J., Hubert C., Kjeldsen K., **Lagkouvardos I.**, Berry D., Jørgensen B., Loy A. Endospores of thermophilic bacteria as tracers of microbial dispersal by ocean currents. **ISME J** 8(6), 1153-65 (2014)

7. Bright M., Espada-Hinojosa S., **Lagkouvardos I.**, Volland JM. The giant ciliate Zoothamnium niveum and its thiotrophic epibiont Candidatus Thiobios zoothamnicoli: a model system to study interspecies cooperation. **Front Microbiol** 5(145), (2014)

8. **Lagkouvardos I.**, Shen J., Horn M. Improved axenization method reveals complexity of symbiotic associations between bacteria and acanthamoebae. **Environ Microbiol Rep**. 6 (4), 383-388 (2014)

9. Schulz F**., Lagkouvardos I.**, Aistleitner K., Kostanjšek R., Horn M. Life in an unusual intracellular niche – a bacterial symbiont infecting the nucleus of amoebae. **ISME J**. 8(8) **,** 1634–1644 (2014)

10. Domman D.B., Collingro A., **Lagkouvardos I.**, Gehre L., Weinmeier T., Rattei T., Subtil A., Horn M. Massive expansion of ubiquitination-related gene families within the Chlamydiae. **Mol Biol Evol** (2014), doi: 10.1093/molbev/msu227

11. Kläring K., Just S., **Lagkouvardos I.**, Hanske L., Blaut M., Haller D., Wenning M., Clavel T. Murimonas intestini gen. nov., sp. nov., an acetate-producing bacterium of the family *Lachnospiraceae* isolated from the mouse gut. Accepted in **IJSEM** (2014)

12. Tsao H.-F., **Lagkouvardos I.**, Horn M. Legionella arctica sp nov isolated from arctic sediments shows adaptation to lower temperature. In preparation

13. Speth D., **Lagkouvardos I.**, Horn M., Jetten MSM. Assessment of global diversity and distribution of annamox bacteria based on available amplicon data. In preparation.

14. Schmidt A., **Lagkouvardos I.**, Clavel T., Haller D. The gut microbial profile of anemic patients suffering from IBD is drastically altered after iron administration. In preparation

15. Schaubeck M., **Lagkouvardos I.**, Calasan J., Kollias G., Clavel T., Haller D. Induction of ileitis in TNF mutant mice is directly associated to specific bacteria. In preparation

16. **Lagkouvardos I.**, Kläring K., Platz S., Scholz B., Huptas C., Wenning M., Engel K-H, Hauner H., Scherer S., Haller D., Rohn S., Skurk T. Clavel T. Flaxseed intervention revealed complex interactions among gut microbiota and blood enterolignans content. In preparation.

Scientific meetings

6th meeting of European Society for Chlamydia Research (Jul 2008). COMP – A database for predicted chlamydial outer membrane proteins (Poster).

8th German Chlamydia Workshop (Feb 2010). The Signature of the PVC superphylum (Poster).

1st EMBO PVC superphylum workshop (Mar 2013). Tapping the metagenomic universe: Diversity, abundance, and distribution of Chlamydiae (Talk). A signature protein for members of the PVC superphylum (Talk).

Rowett-INRA (Jun 2014).metaSRA: Exploration of bacterial diversity and distribution in publically available amplicon sequences (Poster).

7th Seeon Conference (Jul 2014) Forgotten cultivable bugs: diversity and functions of novel mouse intestinal bacteria (MiBC-the Mouse intestinal Bacterial Collection)(Poster)

Acknowledgements

Above and foremost all I would like to thank my supervisor Prof. Dr. Matthias Horn for his trust and patience during the course of this PhD. His guidance and critical thinking shaped every project and was his investment of resources and personal time on myself that made the results possible.

I would like to acknowledge the expert bioinformatics support and training I receive from Prof. Dr. Thomas Rattei and Thomas Weinmaier and to thank Dr. Stephan Schmitz-Esser for been my mentor in the area of amoeba research.

Credits are due to all of my co-authors for their critical contributions in guiding and assisting the completion of all research performed in the context of this dissertation and I also want to express my gratitude to all involved technicians for their excellent support.

Big thanks to all DoME members and fellow members of the IK symbiotic interactions for the fruitful discussions and the friendly atmosphere.

Special thanks to my wife for her patience during the long overtime hours and her encouragement when I was losing my optimism.

Finally, I want to apologize to anyone I forgot to acknowledge or reference properly and to take sole responsibility for possible mistakes or omissions.

Curriculum vitae

Personal Information

Name:	Ilias Lagkouvardos
Date of birth:	25/10/1977
Place of birth:	Chania, Greece

Education

Nov 2007 – Sep 2014	PhD in the Department of Microbial Ecology at the University of Vienna focused on Symbiotic Interactions and *Chlamydia* research
Oct 2004 – Oct 2006	MSc in Life Science Informatics from University of Bonn, at the Bonn-Aachen International Center for Information Technology
Sep 1996 – Sep 2004	MSc in Agricultural Biotechnology from Agricultural University of Athens in the Department of Agricultural Biotechnology
Spring 2001	ERASMUS/SOCRATES student at University of Florence Italy

Research/Work experience

Jan 2014 – present	Postdoctoral fellow in the Chair of Nutrition and Immunology of the Technical University of Munich working in the analysis of gut microbiota.
Nov 2007 – Sep 2014	PhD thesis in the Department of Microbial Ecology at the University of Vienna. Thesis title: “Evolutionary history and phylogenetic diversity of *Chlamydiae*”. Supervisor: Prof. Dr. Matthias Horn.
Apr 2006 – Sep 2006	Master Thesis at Kekulé Institute of Organic Chemistry and Biochemistry in Bonn, Germany. Thesis title: Genomic Analysis of the Primary Metabolism of the uncultivated symbiont “*Pseudomonas paederi*”. Supervisor: Prof. Dr. Joern Piel.
Nov 2005 – Mar 2006	Collaboration with the group of Theoretical Biology at Bonn University for the production of a mathematical and computational model for the motility of epithelial cancer cells. Supervisor: Prof. Dr. Wolfgang Alt.
Jan 2004 – Sep 2004	Diploma dissertation in the Department of Genetics at the Agricultural University of Athens. Thesis title: “Evolution and Systems of Increasing Complexity”. Supervisor: Prof. Dr. John Sourdis
Jul 2000 – Sep 2000	Institute for Soil Science and Plant Nutrition at Martin Luther University of Halle, Germany. Practice exercise working on the building of SOMNET (Soil Organic Matter Network).
Jan 2000 – Jun 2000	Employee of the GUnet (Greek University network) project working on distance learning and teleconferencing services.
Jul 1999 –Sep 1999	Institute for Olive Tree and Subtropical Plants and of Chania, Greece. Practice exercise, working in plant propagation with tissue culture in the laboratory.