1

Marine sediments illuminate Chlamydiae diversity and evolution

Jennah E. Dharamshi[1], Daniel Tamarit[1]†, Laura Eme[1]†, Courtney Stairs[1], Joran Martijn[1], Felix Homa[1], Steffen L. Jørgensen[2], Anja Spang[1,3], Thijs J. G. Ettema[1,4]*

[1] Department of Cell and Molecular Biology, Science for Life Laboratory, Uppsala University, SE-75123 Uppsala, Sweden
[2] Department of Earth Science, Centre for Deep Sea Research, University of Bergen, N-5020 Bergen, Norway
[3] Department of Marine Microbiology and Biogeochemistry, NIOZ Royal Netherlands Institute for Sea Research, and Utrecht University, NL-1790 AB Den Burg, The Netherlands
[4] Laboratory of Microbiology, Department of Agrotechnology and Food Sciences, Wageningen University, NL-6708 WE Wageningen, The Netherlands.

† These authors contributed equally
* Correspondence to: Thijs J. G. Ettema, Email: thijs.ettema@wur.nl

The bacterial phylum Chlamydiae, which is so far comprised of obligate symbionts of eukaryotic hosts, are well-known as human and animal pathogens[1-3]. However, the Chlamydiae also include so-called environmental lineages[4-6] that primarily infect microbial eukaryotes[7]. Studying environmental chlamydiae, whose genomes display extended metabolic capabilities compared to their pathogenic relatives[8-10] has provided first insights into the evolution of the pathogenic and obligate intracellular lifestyle that is characteristic for this phylum. Here, we report an unprecedented relative abundance and diversity of novel lineages of the Chlamydiae phylum, representing previously undetected, yet potentially important, community members in deep marine sediments. We discovered that chlamydial lineages dominate the microbial communities in the Arctic Mid-Ocean Ridge[11], which revealed the dominance of chlamydial lineages at anoxic depths, reaching relative abundances of up to 43% of the bacterial community, and a maximum diversity of 163 different species-level taxonomic unit. Using genome-resolved metagenomics, we reconstructed 24 draft chlamydial genomes, thereby dramatically expanding known interspecies genomic diversity in this phylum. Phylogenomic and comparative analyses revealed several deep-branching Chlamydiae clades, including a sister clade of the pathogenic Chlamydiaceae. Altogether, our study provides new insights into the diversity, evolution and environmental distribution of the Chlamydiae.

During a previous metagenomics study aimed at exploring microbial diversity of deep marine sediments from the Arctic Mid-Ocean Ridge[12], we detected several Chlamydiae-related contiguous sequences (contigs). This finding prompted us to systematically screen marine sediment cores from a region surrounding Loki's Castle hydrothermal vent field (Fig. 1a). We extracted DNA from 69 sediment samples (Supplementary Table 1) from various core depths, followed by screening using Chlamydiae-specific 16S rRNA gene primers. Chlamydiae were identified in 51 (74%) of the samples ranging in depth from 0.1 to 9.4 meters below seafloor (mbsf). We investigated the chlamydial relative abundance and diversity in 30 of these samples using bacterial-specific 16S rRNA gene amplicon sequencing, resulting in the identification of 252 operational taxonomic units (OTUs; clustered at 97% identity; Supplementary Data 1) that could be reliably assigned to Chlamydiae. This analysis revealed notable differences in chlamydiae relative abundance and diversity between samples (Supplementary Data 2), with individual samples showing relative abundances of up to 43% of the total bacterial community and up to 163 OTUs (Fig. 1b). Furthermore, we found that 155 of the 252 chlamydiae OTUs from our samples (hereafter referred to as "marine sediment chlamydiae") could be identified in at least two samples. We further investigated the diversity of these marine sediment chlamydiae by performing a phylogenetic analysis in which the presently discovered chlamydial OTUs were placed amidst a recently compiled dataset of diverse chlamydiae 16S rRNA gene sequences[6]. This analysis revealed that the overall diversity of marine sediment chlamydiae spanned, and expanded, the known chlamydial diversity (Fig. 1c), revealing several deeply branching clades.

To obtain genomic information from these marine sediment chlamydiae, we employed a genome-resolved metagenomics approach (Supplementary Fig. 1) in which we generated 249.6 gigabase pairs (Gbp) of paired-end reads from four sediment samples with a high relative abundance and diversity of chlamydial OTUs (Supplementary Fig. 2). Sequence assembly

generated 5.85 Gbp of contigs larger than 1 kilobase pair (kbp). To assess the diversity of Chlamydiae-related sequences in these metagenome assemblies, we performed phylogenetic analyses of contigs containing 16S rRNA gene sequences or at least five genes of a conserved 15-ribosomal protein gene cluster. These analyses revealed numerous Chlamydiae-related contigs (Supplementary Fig. 2), most of which represented novel lineages that were distantly related to known chlamydiae. Contigs were binned into 24 highly complete (median 95% completeness, Supplementary Table 2) metagenome-assembled genomes (MAGs) on the basis of their tetranucleotide frequencies and patterns of sequence coverage across samples. They differed markedly in predicted genome size (1.33-1.99 Mbp), GC-content (26.4-48.9%) and gene content (Fig. 2, Supplementary Figs. 3 and 4).

To robustly infer the evolutionary relationships of the marine sediment chlamydiae to known lineages (Supplementary Table 3), we performed phylogenomic analyses of concatenated conserved marker protein sequence datasets. These analyses, which were designed to minimize potential long branch attraction and compositional bias artefacts, revealed that the newly reconstructed genomes form five new clades of high taxonomic rank (referred to as Chlamydiae Clades (CC) I-IV and Anoxychlamydiales; Fig. 2, Supplementary Fig. 3, Supplementary Discussion). While these phylogenomic analyses form clades that are in agreement with results based on previously available chlamydial genome data[13], the branching order of these clades inferred differs considerably (Supplementary Discussion). Most marine sediment chlamydiae were placed into one of two well-supported, deeply branching clades, CC-II and Anoxychlamydiales, which also include four MAGs associated with estuary sediment, aquifer groundwater and a drinking water treatment plant[14-16] (Fig. 2). Anoxychlamydiales are unique among Chlamydiae as they comprise members with gene repertoires indicative for an anaerobic lifestyle and will be treated in more detail in a complementary study (*manuscript in prep.*).

Together, CC-I, CC-II, CC-III and Anoxychlamydiales form a superclade that is mainly comprised of uncultivated members represented by MAGs (Fig. 2, Supplementary Fig. 3). While the lifestyles of most members of this superclade remain elusive, the distinctive gene repertoires of the various clades point at functional differences (Supplementary Figs. 3 and 4). A second chlamydial superclade is comprised of the environmental chlamydiae, CC-IV and Chlamydiaceae (Fig. 2, Supplementary Fig. 3) and includes five of the marine sediment chlamydiae MAGs. While two of these MAGs affiliate with the environmental chlamydiae, which comprise symbionts of single-celled eukaryotes such as amoebozoa[7], three MAGs resolve into a previously unknown clade referred to as CC-IV. Previous studies have suggested that Chlamydiaceae, a family that includes important animal pathogens such as the human pathogen *Chlamydia trachomatis*[17], represents a deep-branching clade of the Chlamydiae phylum[13]. However, our phylogenetic analyses strongly support that Chlamydiaceae and CC-IV represent sisterclades, which together share a common ancestor with environmental chlamydiae (Fig. 2). Hence our results indicate that Chlamydiaceae represent a clade that was formed relatively late in chlamydiae evolution (Fig. 2), and that features specifically associated with pathogenicity in the Chlamydiaceae evolved much more recently than previously assumed.

Uncovering the sister relationship of CC-IV and the pathogenic Chlamydiaceae allowed us to re-evaluate the evolutionary events leading to their emergence (Fig. 3a). The genome sizes of CC-IV MAGs, while generally smaller than those of environmental chlamydia (Fig 2.), are larger than those of Chlamydiaceae (Fig. 2), suggesting that the latter have been subjected to genome reduction, a feature often observed in animal pathogens[18]. However, when considering only the gene set conserved across environmental chlamydiae, these differences are less drastic (Fig. 3b), and distribution patterns of Clusters of Orthologous Group (COG) categories found in CC-IV lineages more closely resemble that of environmental chlamydiae than Chlamydiaceae

(Supplementary Fig. 5). Since their divergence from CC-IV, Chlamydiaceae have experienced reductive genome evolution (Fig. 3a-b, Supplementary Fig. 5), including the loss of several core components of central carbon metabolism, and *de novo* biosynthesis of nucleotides and amino acids (Supplementary Fig. 4, Supplementary Discussion). At the same time, they have acquired a small number of genes primarily linked to host-interaction and virulence, in addition to a notable set of conserved genes with unknown functions (Supplementary Fig. 6, Supplementary Discussion). It is likely that some of these unique features of the Chlamydiaceae are linked to the emergence of host-specificity. Further comparative analyses of CC-IV and Chlamydiaceae genomes revealed that only seven gene families are uniquely shared between these clades (Supplementary Fig. 7, Supplementary Discussion). This gene set is highly conserved across the Chlamydiaceae family, hinting at their putative importance in their adaptation to a pathogenic lifestyle. In addition, phylogenetic analyses of gene families with multiple homologs reveal gene duplication events have occurred both before and after the divergence of the Chlamydiaceae from CC-IV (Supplementary Fig. 7, Supplementary Discussion). However, the exact functions of these genes in Chlamydiaceae are currently unknown and our analyses suggest that they should represent a priority for future functional investigations. CC-IV genomes also encode homologs of several flagellar proteins (Supplementary Discussion), which clustered with flagellar components recently identified within distantly-related chlamydiae sampled from marine waters[19] (Supplementary Fig. 8), indicating motility as a sharp difference between CC-IV and Chlamydiaceae.

All previously known members of the Chlamydiae, including environmental lineages, are obligate intracellular symbionts of eukaryotic hosts[20] and display an obligate host-association for replication[10]. Similar to previously characterized lineages[20], the herein identified marine sediment chlamydiae encode various homologs of proteins associated with this typical chlamydial lifestyle

(Supplementary Fig. 4, Supplementary Discussion). For example, the marine sediment chlamydiae contain NF-T3SS (Supplementary Figs. 8 and 9) and NF-T3SS-specific effectors (Supplementary Data 3, Supplementary Discussion). NF-T3SS components are typically present in obligate bacterial symbionts that have a eukaryotic host, but they are also observed in free-living lineages (Supplementary Discussion). Similarly, most of these genomes contain genes encoding other secretion systems (Supplementary Fig. 10 and 11, Supplementary Discussion), some of which have been described to participate in adhesion and invasion in Chlamydiaceae. However, these systems have been shown to have alternative functions in other bacteria (Supplementary Discussion). Nucleotide transporters (NTTs), and in particular ATP/ADP transporters, are a characteristic feature of Chlamydiae and other bacteria typically associated with eukaryotic hosts[21]. All marine sediment chlamydiae genomes encode multiple NTT homologs (Supplementary Figs. 4 and 12, Supplementary Discussion), including possible ATP/ADP NTTs, which cluster together with other chlamydial sequences in phylogenetic analyses (Supplementary Fig. 12). Pathways for the *de novo* biosynthesis of nucleotides and amino acids are often incomplete in obligate intracellular symbionts[8,18]. Indeed, despite identifying chlamydial lineages with the genetic capacity to synthesize both purine and pyrimidine nucleotides *de novo*, we did not observe lineages capable of *de novo* synthesis of all nucleotides and amino acids (Supplementary Fig. 4, Supplementary Discussion).

Since the genomes of these marine sediment chlamydiae encode the typical features of host-dependency found in previously characterized chlamydiae[9], we expected to find indications of potential eukaryotic hosts in the marine sediment samples. While attempts to amplify 18S rRNA genes from DNA isolated from these sediment samples were unsuccessful, screening the four marine sediment metagenome datasets resulted in the identification of a few 18S rRNA gene fragments (Supplementary Table 4, Supplementary Discussion). However, no such sequences

could be identified in the metagenome of the marine sediment sample with the highest Chlamydiae relative abundance (Fig. 1b). Given the absence of obvious host candidates, we explored the possibility that these genome sequences derive from persistent marine sediment chlamydiae cells, such as from elementary bodies, which can survive outside of host cells and even remain metabolically active, despite being unable to replicate[22,23]. Estimation of replication rates based on sequence read-coverage[24] indicated that these MAGs are derived from actively dividing marine sediment chlamydiae (Fig. 4b). Thus, our results suggest that at least some of these marine sediment chlamydiae are not associated with eukaryotic hosts.

Finally, we investigated chlamydial diversity in environments other than marine sediments. An analysis of publicly available 16S rRNA gene amplicon datasets from a variety of environments revealed that fresh water, ground water, salt marshes and wastewater often harbour diverse and abundant chlamydiae (Fig. 4b). These findings likely represent considerable underestimations of chlamydial diversity and relative abundance since mismatches in commonly used 16S rRNA primer sets (e.g., Earth Microbiome Project[25], Supplementary Table 5) generally do not capture known Chlamydiae diversity (Supplementary Discussion).

Our work reports the existence and genomic characterization of an extended diversity of Chlamydiae-related lineages in deep marine sediments and provides insights into the evolution and diversification of the Chlamydiae phylum. Using sophisticated phylogenomic methods, we used a robust phylogenomic framework for investigating Chlamydiae evolution. Furthermore, we identified several new Chlamydiae clades of high taxonomic rank, including a sister clade of the pathogenic Chlamydiaceae, which provided insights into the early evolution of this family. Altogether, our findings indicate that Chlamydial diversity and abundance has been underappreciated in environmental surveys, and our observations represent a shift in our view of the environmental distribution of Chlamydiae. These results indicate differences in basic ecology

and lifestyle across this phylum, and contribute to a comprehensive understanding of its evolution, including the emergence of host dependency and pathogenicity.

References

1 Bachmann, N. L., Polkinghorne, A. & Timms, P. Chlamydia genomics: providing novel insights into chlamydial biology. *Trends Microbiol* **22**, 464-472 (2014).

2 Nunes, A. & Gomes, J. P. Evolution, phylogeny, and molecular epidemiology of Chlamydia. *Infect Genet Evol* **23**, 49-64 (2014).

3 Elwell, C., Mirrashidi, K. & Engel, J. Chlamydia cell biology and pathogenesis. *Nat Rev Microbiol* **14**, 385-400 (2016).

4 Amann, R. *et al.* Obligate Intracellular Bacterial Parasites of Acanthamoebae Related to Chlamydia spp. *Applied and Environmental Microbiology* **63**, 115–121 (1997).

5 Horn, M. & Wagner, M. Evidence for additional genus-level diversity of Chlamydiales in the environment. *FEMS Microbiology Letters* **204**, 71-74 (2001).

6 Lagkouvardos, I. *et al.* Integrating metagenomic and amplicon databases to resolve the phylogenetic and ecological diversity of the Chlamydiae. *ISME J* **8**, 115-125 (2014).

7 Horn, M. Chlamydiae as symbionts in eukaryotes. *Annu Rev Microbiol* **62**, 113-131, doi:10.1146/annurev.micro.62.081307.162818 (2008).

8 Horn, M. *et al.* Illuminating the evolutionary history of chlamydiae. *Science* **304**, 728-730, doi:10.1126/science.1096330 (2004).

9 Omsland, A., Sixt, B. S., Horn, M. & Hackstadt, T. Chlamydial metabolism revisited: interspecies metabolic variability and developmental stage-specific physiologic activities. *FEMS Microbiol Rev* **38**, 779-801, doi:10.1111/1574-6976.12059 (2014).

10 Taylor-Brown, A., Vaughan, L., Greub, G., Timms, P. & Polkinghorne, A. Twenty years of research into Chlamydia-like organisms: a revolution in our understanding of the biology and pathogenicity of members of the phylum Chlamydiae. *Pathog Dis* **73**, 1-15 (2015).

11 Pedersen, R. B. *et al.* Discovery of a black smoker vent field and vent fauna at the Arctic Mid-Ocean Ridge. *Nat Commun* **1**, doi:10.1038/ncomms1124 (2010).

12 Spang, A. *et al.* Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521**, 173-179, doi:10.1038/nature14447 (2015).

13 Pillonel, T., Bertelli, C. & Greub, G. Environmental Metagenomic Assemblies Reveal Seven New Highly Divergent Chlamydial Lineages and Hallmarks of a Conserved Intracellular Lifestyle. *Front Microbiol* **9**, 79, doi:10.3389/fmicb.2018.00079 (2018).

14 Baker, B. J., Lazar, C. S., Teske, A. P. & Dick, G. J. Genomic resolution of linkages in carbon, nitrogen, and sulfur cycling among widespread estuary sediment bacteria. *Microbiome* **3**, 14, doi:10.1186/s40168-015-0077-6 (2015).

15 Anantharaman, K. *et al.* Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat Commun* **7**, 13219, doi:10.1038/ncomms13219 (2016).

16 Pinto, A. J. *et al.* Metagenomic Evidence for the Presence of Comammox Nitrospira-Like Bacteria in a Drinking Water System. *mSphere* **1**, doi:10.1128/mSphere.00054-15 (2016).
17 Stephens, R. S. *et al.* Genome Sequence of an Obligate Intracellular Pathogen of Humans: Chlamydia trachomatis. *Science* **282**, 754-759 (1998).
18 Toft, C. & Andersson, S. G. E. Evolutionary microbial genomics: insights into bacterial host adaptation. *Nature Reviews Genetics* **11**, 465-475, doi:10.1038/nrg2798 (2010).
19 Collingro, A. *et al.* Unexpected genomic features in widespread intracellular bacteria: evidence for motility of marine chlamydiae. *ISME J* **11**, 2334-2344, doi:10.1038/ismej.2017.95 (2017).
20 Collingro, A. *et al.* Unity in variety--the pan-genome of the Chlamydiae. *Mol Biol Evol* **28**, 3253-3270 (2011).
21 Schmitz-Esser, S. *et al.* ATP/ADP Translocases: a Common Feature of Obligate Intracellular Amoebal Symbionts Related to Chlamydiae and Rickettsiae. *Journal of Bacteriology* **186**, 683-691, doi:10.1128/jb.186.3.683-691.2004 (2004).
22 Haider, S. *et al.* Raman microspectroscopy reveals long-term extracellular activity of Chlamydiae. *Mol Microbiol* **77**, 687-700, doi:10.1111/j.1365-2958.2010.07241.x (2010).
23 Sixt, B. S. *et al.* Metabolic features of Protochlamydia amoebophila elementary bodies--a link between activity and infectivity in Chlamydiae. *PLoS Pathog* **9**, e1003553, doi:10.1371/journal.ppat.1003553 (2013).
24 Brown, C. T., Olm, M. R., Thomas, B. C. & Banfield, J. F. Measurement of bacterial replication rates in microbial communities. *Nat Biotechnol* **34**, 1256-1263, doi:10.1038/nbt.3704 (2016).
25 Thompson, L. R. *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457-463, doi:10.1038/nature24621 (2017).

Acknowledgements We thank M. Horn, L. Guy, S. Abby, L. Juzokaite, A. E. Lind, K. Zaremba-Niedzwiedzka and E. C. Fernandez for technical assistance and/or for useful advice and discussions. We also acknowledge the help from chief scientist R. B. Pedersen, the scientific party and the entire crew on board the Norwegian research vessel G.O. Sars during the summer 2008, 2010 and 2014 expeditions. All sequencing was performed by the National Genomics Infrastructure sequencing platforms at the Science for Life Laboratory at Uppsala University, a national infrastructure supported by the Swedish Research Council (VR-RFI) and the Knut and Alice Wallenberg Foundation. We thank the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) at Uppsala University and the Swedish National Infrastructure for Computing (SNIC) at the PDC Center for High-Performance Computing for

245 providing computational resources. This work was supported by grants of the European Research
246 Council (ERC Starting grant 310039-PUZZLE_CELL), the Swedish Foundation for Strategic
247 Research (SSF-FFL5) and the Swedish Research Council (VR grant 2015-04959) to T.J.G.E.
248 C.W.S. is supported by a European Molecular Biology Organization long-term fellowship
249 (ALTF-997-2015) and the Natural Sciences and Engineering Research Council of Canada
250 postdoctoral research fellowship (PDF-487174-2016). Funding was received from the European
251 Union's Horizon 2020 research and innovation program under the respective Marie Sklodowska-
252 Curie grant agreements 625521 (to A.S.) and 704263 (to L.E.). A.S. is supported by the Swedish
253 Research Council (VR starting grant 2016-03559) and the NWO-I foundation of the Netherlands
254 Organisation for Scientific Research (WISE fellowship).

255

256 **Author Contributions** T.J.G.E. and A.S. conceived the study, and J.E.D. and T.J.G.E.
257 performed the experimental design. S.L.J. provided environmental samples. S.L.J. and J.E.D.
258 performed DNA extractions and performed PCR-based screening. J.E.D. generated 16S rRNA
259 gene amplicons. F.H. and J.E.D. performed metagenomic sequence assemblies. J.E.D., J.M. and
260 F.H. performed genome-resolved metagenomics analyses. J.E.D., F.H. and T.J.G.E. analyzed 16S
261 rRNA gene amplicon sequence data. J.E.D., C.S., D.T., L.E., J.M., and A.S. analyzed genomic
262 data and performed phylogenetic analyses. J.E.D., C.S., L.E., D.T., A.S. and T.J.G.E. interpreted
263 the obtained data and results. T.J.G.E. and J.E.D wrote, and all authors edited and approved, the
264 manuscript.

265

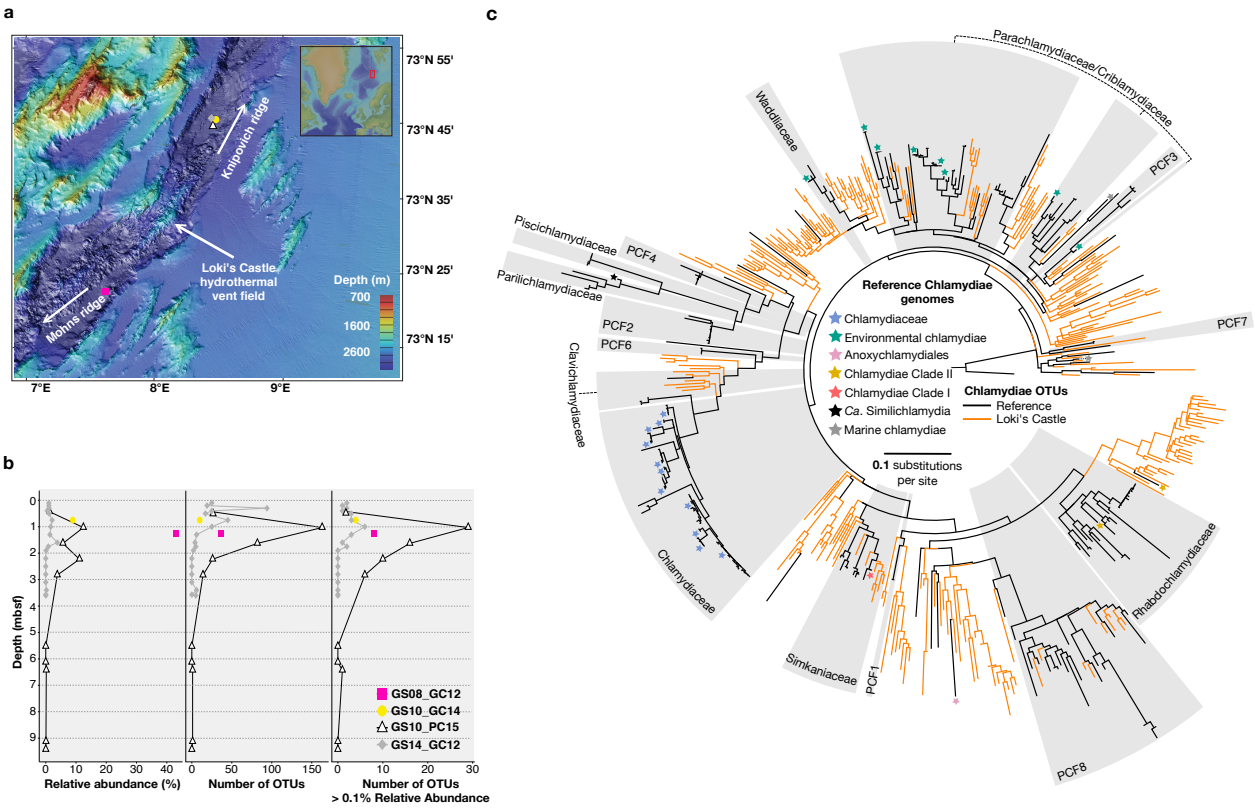

Figure 1. *Chlamydiae are diverse and abundant in Loki's Castle marine sediments*. **a**, Bathymetric map of sediment core sampling locations taken northeast (GS10_GC14, GS10_PC15 and GS14_GC12) and southwest (GS08_GC12) of Loki's Castle hydrothermal vent field. **b**, Chlamydial relative abundance, OTU number and abundant OTUs, as a factor of depth (meters below seafloor, mbsf) based on bacterial 16S rRNA gene amplicons. **c**, Maximum-likelihood (ML) phylogenetic tree (520 taxa, 476 sites) of marine sediment chlamydiae 16S rRNA gene sequences from amplicon OTUs (orange) and a reference (black) dataset, rooted using PVC taxa as an outgroup, inferred using IQ-TREE with the GTR+R7 model of evolution. Previously sequenced chlamydial genomes and metagenomic bins are labelled with stars.

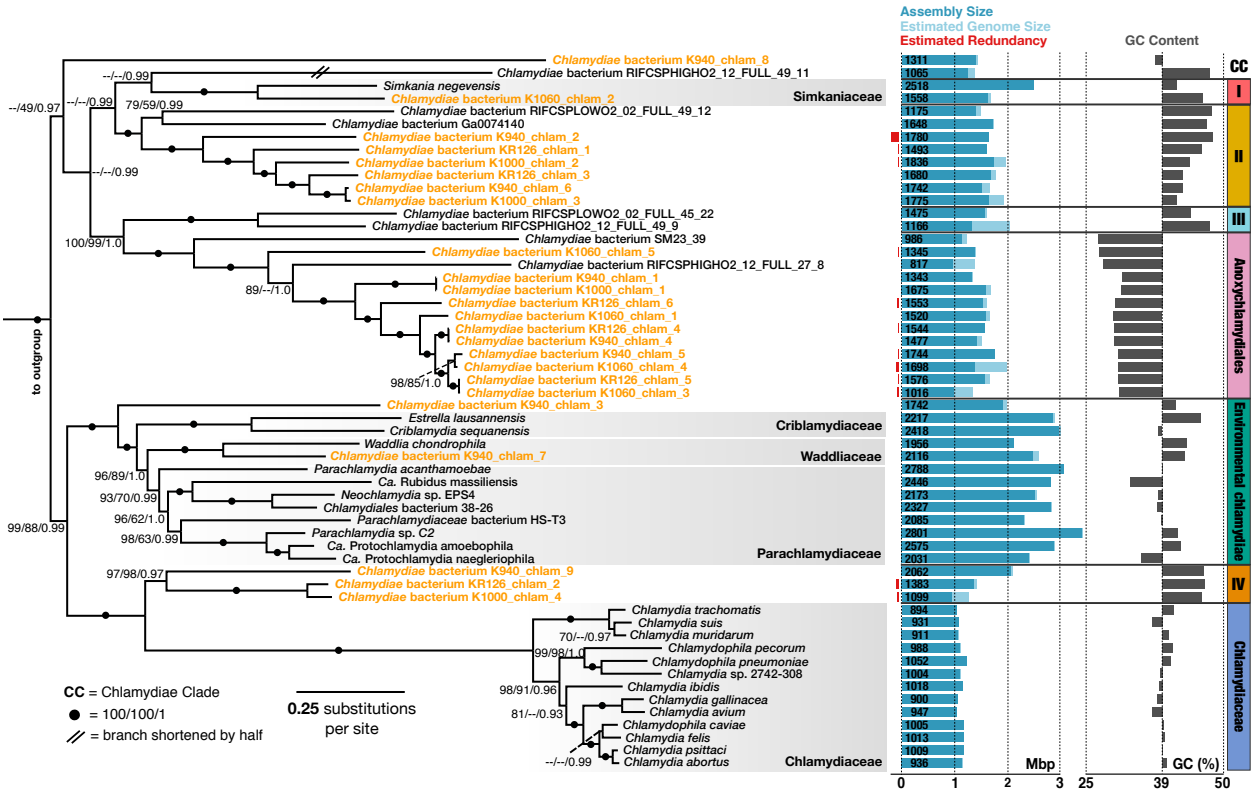

Figure 2. *Marine sediment lineages span Chlamydiae species phylogeny*. Bayesian phylogenetic tree of 38 concatenated single-copy conserved marker proteins in which marine sediment chlamydiae and previously characterized chlamydial representatives are shown in orange and black fonts, respectively. Previously established chlamydial families are shaded in grey. Branch support values were mapped onto the tree in the following order: non-parametric bootstrap support values (BV) for the full alignment (8006 sites) and reduced alignment (6005 sites after removal of the top 25% compositionally heterogeneous sites), each under the LG+C60+G+F derived PMSF approximation estimated by IQ-TREE, and posterior probability (PP) support values under the CAT+GTR+Γ4 model of evolution inferred with Phylobayes. For each species, the genome bin size, estimated genome size and redundancy are reported in Mbp along with the number of predicted open reading frames, and GC content.

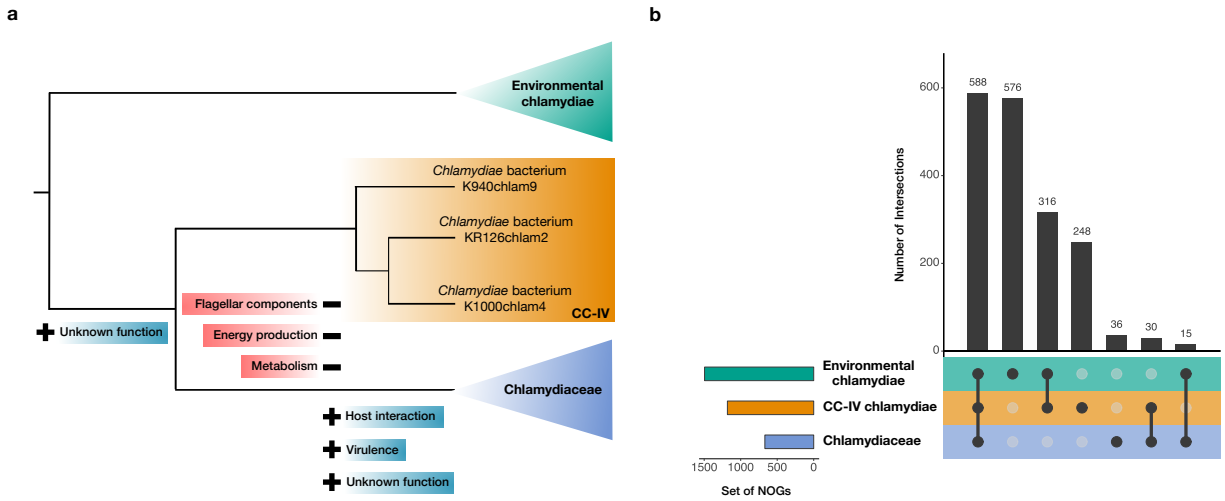

Figure 3. *Gene content evolution in Chlamydiaceae*. **a**, Schematic overview of lost and acquired cellular features, based on presence and absence patterns of NOGs (Supplementary Data 4), in pathogenic Chlamydiaceae, inferred from sister clade CC-IV. **b**, Plot showing intersections of NOGs conserved (in a third of lineages affiliated with each clade, Supplementary Table 2 and 3) across environmental chlamydiae, CC-IV and Chlamydiaceae.

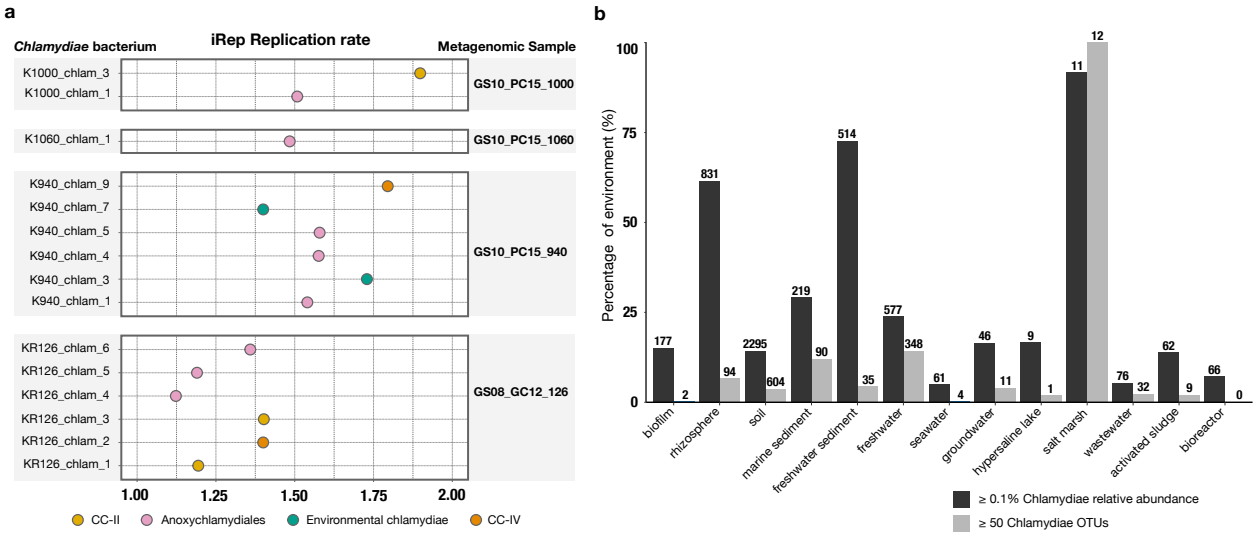

Figure 4. *Estimated iRep replication rate of marine sediment chlamydiae, and bar plot showing presence of chlamydiae in diverse environments*. **a**, Replication rates of marine sediment chlamydiae inferred using iRep. Indicates the proportion of the microbial population represented by the metagenome-assembled genome that is actively dividing. **b**, Percentage of samples from selected environments which contain a chlamydial relative abundance of ≥ 0.1%, or ≥ 50 OTUs based on publicly available 16S rRNA gene amplicon datasets. Values above each bar represent the absolute number of samples.

Methods

Sample acquisition

Sediment cores were retrieved from the Arctic Mid-Ocean Ridge near Loki's Castle[11] hydrothermal vent field over multiple sampling expeditions: GS08_GC12 in 2008 (3.3 m)[26], GS10_GC14 (2 m) and GS10_PC15 (11.2 m) in 2010[27], and GS14_GC12 (3.6 m) in 2014. Sediment samples at various depths (Supplementary Table 1) were collected on-board and immediately frozen for prospective microbiological analysis. Geochemistry and sediment characteristics of these samples have been published previously[26,27], except in the case of GS14_GC12.

Sample screening and bacterial 16S rRNA gene amplicon sequencing

DNA was extracted using the PowerLyzer® PowerSoil® DNA kit in conjunction with the PowerLyser® 24 homogenizer as per manufacturer's instructions (MOBIO) with minor modifications using 0.5-0.7 g of sediment and the addition of polyadenosine to increase DNA yield[28]. Taxonomic coverage of all primer pairs, for both sample screening and amplicon sequencing, was tested *in silico* using SILVA TestPrime[29] with the SSU r132 RefNR database (Supplementary Table 5). All primers used have been published previously[29,30] except for Chla-310-a-20 which was designed with PRIMROSE[31] (Supplementary Table 5). Reaction conditions for each primer pair are found in Supplementary Table 5. Primer pair Chla-310-a-20 and S-*-Univ-1100-a-A-15[29], which amplifies an approximately 800 bp region of the 16S rRNA gene, was used to screen 69 Loki's Castle marine sediment samples (Supplementary Table 1) for Chlamydiae. Primer pair 574*f[30] and 1132[30] were used to screen for 18S rRNA genes in these same sediment samples. The latter primer pair did not yield positive amplification products, in

spite of its broad taxonomic coverage of known eukaryotic 18S rRNA genes (Supplementary Table 5).

Thirty sediment samples with positive PCR screening results for Chlamydiae were selected for further investigation. Bacterial-specific 16S rRNA gene amplicons were sequenced using a two-step PCR approach. Primer sequences and reaction conditions are reported in Supplementary Table 5. In the first step, a ca. 500 bp region of the 16S rRNA gene was amplified in triplicate, to account for random PCR drift[32], using bacterial primers (S-D-0564-a-S-15 and S-D-Bact-1061-a-A-17[29]). For each reaction, replicates were pooled and purified using magnetic AMPure XP beads (Agencourt). In the second PCR step, libraries were constructed using adaptor sequences and reaction conditions from the TruSeq DNA LT Sample Prep Kit (Illumina), before sequencing with Illumina MiSeq (2x300 bp). Using cutadapt[33] v. 1.10 sequence reads shorter than 100 bases were filtered out, 3' ends trimmed to a minimum Phred quality score of 10, and primer sequences removed. Forward and reverse reads were merged using VSEARCH[34] v. 1.11.1 (--fastq-minovlen 16), de-replicated (--derep_fulllength), and clustered into centroid OTUs (threshold = 97%)**.** Chimeras were detected and removed using UCHIME[35] with the SILVA123.1_SSUref_tax:99 database[36]. Taxonomy was assigned using the LCAClassifier[37] from CREST-2.0.5 with silvamod106 as the reference database.

Metagenome sequencing and assembly

The Fast DNA Max Spin kit and Fast DNA Spin kit (for GS10_PC15_940 only, 4 replicates pooled) were used according to manufacturer's protocols (MP Biomedical), with the addition of polyadenosine, to extract DNA from sediment samples chosen for further analysis (GS08_GC12_126, GS10_PC15_940, GS10_PC15_1000, and GS10_PC15_1060). Libraries were prepared using the Nextera DNA Library Prep kit (Illumina) with 25 ng of input DNA, and sequenced with Illumina HiSeq in rapid-mode (2x 250 bp). For GS10_PC15_1060, reads from

three separate HiSeq runs and undetermined reads (reads with barcode mismatches) from one run were combined. Quality control to remove low-quality reads was performed using Trimmomatic[38] 0.35 with the options: SLIDINGWINDOW:4:12, MINLEN:50, ILLUMINACLIP:TruSeq Illumina Universal Adaptors. FastQC[39] v.011.4 was used to visually evaluate sequence quality before and after processing. Using the fq2fa program from the IDBA-UD[46] package, paired reads were interlaced and Ns removed, using the options 'merge' and 'filter' (except for GS10_PC15_1060 reads, where Ns were retained, as it improved the assembly). Trimmed reads from each sample were subjected to iterative *de novo* assembly using IDBA-UD 1.0.9 (minimum k-mer size = 20 and maximum k-mer size = 124, except for GS08_GC12_126 for which maximum k-mer size = 100). Assembly quality and statistics (Supplementary Fig. 2) were assessed using QUAST[40] v3. Open reading frames (ORFs) across assembled contigs in each metagenome were called with Prodigal[41] v.2.6.3.

Assesment of metagenome microbial composition

To investigate the microbial composition of the samples, 'ribocontigs', i.e. a contig encoding at least 5 of 15 ribosomal proteins found in an operon conserved across prokaryotes[42], were identified in the metagenomic assemblies using the RP15 pipeline[43]. As part of this, a maximum-likelihood (ML) phylogeny using the LG+C60+G model of evolution (Supplementary Fig. 2) was inferred from a dataset consisting of concatenated ribosomal proteins extracted from metagenomic ribocontigs and 90 phylogenetically diverse reference taxa including bacteria and archaea[44] in addition to PVC representative species (Supplementary Tables 3 and 6).

Obtaining Chlamydiae metagenome assembled genomes

A differential coverage binning approach was used to obtain metagenome-assembled genomes (MAGs). For each metagenome assembly, contig coverage was estimated using pseudoalignment with Kallisto[45] 0.42.5, with sequence reads from each of the four samples. Differential coverage

444 profiles were generated for each assembly using a freely available script
445 (github.com/EnvGen/toolbox/tree/master/scripts/kallisto_concoct/input_table.py), provided by
446 Johannes Alneberg. To give more statistical weight to longer contigs and to reduce the impact of
447 chimeric sequences[46], contigs larger than 20 kb were split into 10 kb fragments. CONCOCT[46]
448 v.0.40 was used to cluster contigs within each focal assembly into MAGs, using their differential
449 coverage profile[47], tetranucleotide frequency, and different contig length cut-offs (1 kb, 2 kb and
450 3 kb). Due to the large diversity in sample GS10_PC15_1060, the maximum number of bins was
451 adjusted (1500 bins for 1 kb, and 1000 bins for 2 and 3 kb length cut-offs). Putative chlamydial
452 metagenomic bins were identified by phylogenomic analyses of concatenated ribosomal proteins
453 encoded on ribocontigs (Supplementary Fig. 2). Completeness and redundancy was assessed
454 using the micomplete[48] pipeline (without weighting), provided by Lionel Guy
455 (https://bitbucket.org/evolegiolab/micomplete) using a custom marker set (Supplementary Table
456 7) corresponding to genes present in complete Chlamydiae genomes (Supplementary Table 3).
457 For each ribocontig, the corresponding metagenomic bin with the highest completeness and
458 lowest redundancy across the CONCOCT iterations was selected for further analysis. Chlamydial
459 metagenome bins were subjected to manual cleaning using mmgenome[49], resulting in medium
460 and high quality MAGs (Supplementary Table 2) based on MIMAG standards[56]. Differential
461 coverage across samples, GC content, linkage, the presence of chlamydial-specific marker
462 proteins[20], and single-copy bacterial marker proteins (Supplementary Table 7) were visualized
463 using mmgenome[50]. Contigs with profiles diverging from the majority of contigs were removed.
464 Linkage information, i.e. information about which contigs are connected by read pairs, was
465 calculated with read mapping using Bowtie2[51], followed by application of the script
466 bam_to_linkage.py from CONCOCT[46]. After generating the final MAGs, contigs that had been
467 split into 10 kb contig fragments were joined again if more than half of the fragments of a

specific contig were assigned to a specific MAG. Otherwise, all fragments of the contig in question were discarded.

Selection of published PVC genomes for comparative and phylogenetic analysis

Selection of PVC superphylum representatives was facilitated by a phylogenetic analysis of ribocontigs from PVC member genomes available in NCBI (as of February 6th, 2017). Using the RP15 pipeline[43], a maximum-likelihood (ML) phylogeny of these ribocontigs was inferred, using RAxML[52] 8.2.4 under the PROTCATLG model of evolution. Branch support was estimated through 100 rapid bootstrap replicates ('-f a') (Supplementary Fig. 13). Phylogenetically diverse representatives from non-Chlamydiae PVC phyla were selected (Supplementary Table 6) to be used as an outgroup for phylogenomic analyses. With the exception of Chlamydiaceae (for which we used genomes classified as reference or representative in NCBI), all Chlamydiae genomes (Supplementary Table 3) were used in protein phylogenies and comparative genomics analyses. For determining interspecies relationships within the Chlamydiae phylum (Supplementary Fig. 13), only one representative was kept whenever several chlamydial genomes had a near-identical phylogenetic placement (Supplementary Table 3). Chlamydial species representative MAGs and single-cell assembled genomes (SAGs), which were made available on NCBI subsequently (between February 6th, 2017 and April 18th, 2018, Supplementary Table 3), were included in comparative genomics and phylogenetic analyses (Supplementary Fig. 4).

Protein clustering and gene annotation

Gene features of marine sediment chlamydiae MAGs (Supplementary Table 2) were annotated with Prokka[53] v1.12, using a version that allows for partial gene prediction (GitHub pull request #219). All protein sequences from both chlamydiae in NCBI (Supplementary Table 3) and from those in this study (Supplementary Table 2) were searched against databases as follows: top hits with and excluding Chlamydiae against the *nr* database and taxonomic classification (Lowest

Common Ancestor (LCA) algorithm, '-f 102') using blastp (--more-sensitive) from DIAMOND[54] aligner v0.9.19.120; PFAM (PF)[55] and Interpro (IPR)[56] domains, and MetaCyc[57] and KEGG[58,59] pathway annotations, were assigned using Interproscan[60] version 5.22-61.0; KEGG 'KO' numbers were assigned using GhostKOALA[61]. Protein sequences were also mapped to the eggNOG orthologous groups[62] version 4.5 using eggNOG-mapper[63], at both the universal '-d NOG' and bacterial level '-d BACT'. The presence of proteins of interest across Chlamydiae genomes were assessed using these annotations and database searches (Supplementary Data 3). The presence of amino acid and nucleotide *de novo* biosynthesis and central carbon metabolism pathways (Supplementary Data 3) were manually investigated using KEGG[58] KO number assignments.

Detection of flagellar genes, secretion systems, NF-T3SS secreted proteins, eukaryotic-like domains and subcellular targeting signals in the host

Genes related to secretion systems and the flagellum were detected for most chlamydiae using MacSyFinder[71] with the protein models built by Abby et al.[64], using the mode 'gembase' for complete genomes and 'unordered' for incomplete genomes. We predicted NF-T3SS secreted proteins, eukaryotic-like domains (ELD), and putative subcellular targeting signals to eukaryotic cellular compartments for all predicted proteins from chlamydiae (Supplementary Table 2 and 3) and from nine well-characterized PVC representatives with free-living lifestyles. For this, we used EffectiveDB[65] in 'genome mode', i.e. enabling the prediction of secretion systems and the discovery of novel ELD (Supplementary Data 3).

Phylogenetic analyses

Unless otherwise stated, gene or protein sequences were aligned using MAFFT-L-INS-i[74] v7.271 and trimmed with trimAl[66] v1.4 (--gappyout). Identical sequences were removed, and alignments were inspected manually. ML phylogenetic analyses were performed using IQ-TREE[67] 1.5.3 with

automated model selection[68] among the following models: the empirical LG model[69] the empirical profile mixture models (C10 to C60) combined with the LG exchangeability matrix (e.g., LG+C10)[70], with or without empirically determined amino acid frequencies (+F), and free or gamma-distributed rates (+R or +G)[71]. Bootstrap support values were inferred from 1000 ultrafast bootstrap (ufBV) replicates and from 1000 replicates of the SH-like approximate likelihood ratio test (SH-aLRT). All unprocessed phylogenetic trees can be found in Supplementary Data 4.

Phylogenetic inference of interspecies relationships within the Chlamydiae phylum

Identification and phylogenetic analysis of 16S rRNA gene amplicon and metagenome sequences

Barrnap[53] 0.8 was used to identify 16S/18S rRNA genes in the metagenomic assemblies, three iterations with 'kingdom' set to 'euk', 'arc' and 'bac' were run, with 'reject' set to 20% of the rRNA gene. Sequences were taxonomically classified using LCAClassifier[37] from CREST-3.0.5 with silvamod128 as the reference database. The microbial composition of each metagenome can be found in Supplementary Table 4. A reference dataset of near-full length chlamydiae 16S rRNA gene sequences from various metagenomic and amplicon sequence databases[6] was used for determining the phylogenetic placement of these sequences. ML phylogenetic inference (Fig. 1c, Supplementary Fig. 2) was performed with the GTR+R7 model of evolution[72] (based on model selection).

Selection of single-copy marker proteins

We selected marker proteins using NOGs that were present in a single-copy in 95% of near-complete PVC genomes (Supplementary Fig. 14). For each of the 149 markers protein initially identified, alignments were generated and manually curated to remove divergent sequences prior to final alignment and trimming with trimAl[66] v1.4 (--automated1). ML phylogenies were inferred for each alignment, and proteins displaying patterns of vertical inheritance were selected.

Due to strain microdiversity, it is not uncommon to have contigs from two different, but closely related, lineages together in a MAG, resulting in redundancy. For each marker, in cases of multiple copies from the same MAG, alignments and corresponding single-gene phylogenies were manually inspected to determine if it represented a paralog, a redundant gene copy, or partial sequences from the same gene. If redundant sequences overlapped with non-identical regions, all sequences from the same genome were removed; if they were placed at the end of a contig and shared an identical overlapping region (longer than 30 nucleotides), the sequences were merged; and if they were partial non-overlapping protein fragments, the longer fragment was selected and the shorter one removed. This inspection resulted in 126 remaining protein markers which were subjected to discordance filtering (or "χ^2 trimming") to remove markers with the most conflicting phylogenetic signal[73]. The 20% of markers whose taxon bipartition profiles were least concordant with the others, resulting in a high discordance score (Supplementary Fig. 15), were removed. This final set of 98 marker proteins was concatenated and used for species-tree reconstruction (Supplementary Table 8).

Phylogenetic inference of interspecies relationships using concatenated marker proteins

Chlamydial species representatives (released prior to February 6th, 2017, Supplementary Table 3) and marine sediment chlamydiae MAGs (Supplementary Table 2) were used to investigate interspecies phylogenomic relations within the Chlamydiae phylum, using other PVC representatives as an outgroup (Supplementary Table 6). Marker proteins were separately aligned and trimmed before being concatenated into a supermatrix. ML inference was performed on a supermatrix of all 98 identified single-copy marker proteins (28,286 amino acid positions), and a sub-selection of the 55 (14,212 amino acid positions) and 38 (7,894 amino acid positions) markers with the highest representation among lineages (Supplementary Table 8), using the LG+C60+ G+F model of evolution. The tree topologies inferred were similar across all three

datasets (Supplementary Fig. 16), indicating that the 38 marker protein sub-selection was sufficient for further inferences. Several more in-depth phylogenetic analyses were applied to the supermatrix of 38 marker proteins (Fig. 2), using a smaller outgroup to allow for more computationally intensive analyses (Supplementary Table 6). PMSF is a site-heterogeneous mixture model that can closely approximate complex mixture models such as LG+C60+G+F while reducing computational time several-fold[74], making full bootstrapping practical. A ML phylogeny was inferred using the PMSF model implemented in IQ-TREE[67] 1.5.5, with a guide-tree inferred using the LG+C60+G+F model of evolution, with 100 nonparametric bootstrap replicates (BV). The same analysis was performed on the alignment, after it was subjected to χ^2-trimming. Here, the proportion of most heterogeneous sites were was removed in a step-wise fashion from 5% to 95% of sites, as previously described[48,75]. The resulting χ^2 test statistics for each lineage under the various heterogenous site removal treatments were visualized for the chlamydiae and PVC outgroup (Supplementary Fig. 17) and 25% removal was chosen as the treatment which best lowered compositional heterogeneity while retaining the largest number of informative sites. Bayesian analysis of this alignment was performed under the CAT+GTR+Γ4 model with PhyloBayes MPI[76] 1.7a. Four independent Markov chain Monte Carlo chains were run for ~55,000 generations, after which, three chains converged (maxdiff = 0.12; burn-in = 15,000). Subsequently, these 38 markers were updated with additional Chlamydiae species representatives (released between February 6th, 2017 and April 18th, 2018, Supplementary Table 3). Single-protein phylogenies and alignments including these lineages were inferred and manually inspected, as described above, followed by concatenation of each of the 38 trimmed alignments. A robust ML phylogeny was inferred based on the resulting supermatrix using the selected LG+C60+PMSF model, as described above (Supplementary Fig. 3).

Phylogenetic analyses of proteins of interest

Flagellar genes and NF-T3SS

Sequences identified for each NF-T3SS/flagellar gene (see above) were gathered, separately aligned and trimmed with BMGE[77] v. 1.12 (-m BLOSUM30). ML phylogenies were inferred with IQ-TREE[67] v1.6.5 with model selection among the following models: the empirical LG model[69] the empirical profile mixture models (C10 to C60) combined with the LG exchangeability matrix (e.g., LG+C10)[70], with or without empirically determined amino acid frequencies (+F), and specific free rates (+R0, +R2, +R4 or +R6)[71]. An additional ML reconstruction was run with the same alignments using the PMSF[74] approximation of the selected model and the previously obtained tree as a guide tree, with 100 BV. These reconstructions were used to reclassify sequences between the NF-T3SS/flagellar homologues and outgroups. Phylogenies of individual components were largely congruent (Supplementary Data 4), allowing us to reconstruct a concatenated phylogeny using the separately trimmed alignments of the *sctJ*, *sctN*, *sctR*, *sctS*, *sctT*, *sctU*, and *sctV* homologues with IQ-TREE[67] v1.6.5 as above (Supplementary Data 4).

Nucleotide transporters

The diversity and phylogenetic placement of nucleotide transporter (NTT) proteins from marine sediment chlamydiae was investigated. The region corresponding to the NTT PF[78] domain (PF03219) with a single-domain structure was predicted using hmmscan from the HMMER v3.1 toolkit (http://hmmer.org) and extracted from the full protein sequences. An initial phylogeny (not shown) was reconstructed based on this conserved region using FastTree2[79], and indicated two monophyletic clades separated by a long branch as identified previously[80]. Further phylogenetic analyses were performed separately on the clade containing "canonical NTT" and the "other NTT" clade (Supplementary Fig. 12). NTT proteins containing a HEAT domain were also investigated (Supplementary Fig. 12). Proteins homologous to the *Chlamydia trachomatis*

query were retrieved from marine sediment chlamydiae using BLASTP and were combined with NTT-HEAT sequences identified in a prior study[80]. All three NTT datasets were each aligned before being trimmed with TrimAl[81] v1.4, using the 'gappyout' option for the "canonical NTT" alignment and the 'automated1' option for the "sister NTT" and NTT-HEAT alignments. The LG+F+R8 model of evolution was selected for ML phylogenetic inference, except for the NTT-HEAT alignments, where the LG+F+R6 model was selected instead. Chlamydiae NTTs that have been functionally characterized were annotated in the resulting phylogenies (Supplementary Fig. 12, Supplementary Discussion).

Proteins conserved in CC-IV and Chlamydiaceae

NOGs[62] and PF[78] domains identified uniquely in clade CC-IV and Chlamydiaceae among chlamydial lineages (Supplementary Table 3) were compiled (Supplementary Fig. 6 and 7) and phylogenetic investigation of these sequences with the PF domains PF04518 and PF05302 performed (Supplementary Fig. 7). Hmmalign, from the HMMER v3.1 toolkit (http://hmmer.org), was used to extract the region corresponding to the domains from sequences, which were aligned and trimmed as outlined previously. ML phylogenies were inferred using the PMSF model implemented in IQ-TREE[67] 1.5.5, with a guide-tree inferred using the LG+C20+G+F model of evolution, with 100 BV. An hmmsearch against the *nr* database confirmed that these PF domains are only present in CC-IV and Chlamydiaceae chlamydiae.

Determination of replication rates

We used iRep[24] to determine the replication rate of the microbial population represented by each MAG. The tool uses differences in sequencing coverage that arise bi-directionally across the genomes of replicating bacteria, due to the single origin of replication, to infer a population-level rate of replication[24]. Since iRep calculations require MAGs that are estimated to be 75% complete with 5X coverage, we analyzed only 15 of the 24 MAGs that met these criteria.

636 Sequenced reads from each metagenome were mapped to corresponding assembled contigs using
637 Bowtie2[51] with the 'reorder' option, before applying iRep[24] v.1.10 with default settings (Fig. 4a).

638 **Chlamydial environmental diversity**

639 Using the Integrated Microbial NGS platform (IMNGS)[82] (accessed March 5th, 2018), which
640 systematically screens prokaryotic 16S rRNA gene amplicon datasets deposited as sequence read
641 archives in NCBI, the percentage of samples from select environments with a relative abundance
642 of over 0.1% Chlamydiae, and a percentage with at least 50 OTUs was assessed (Fig. 4b).

643 **Data visualization**

644 Plots in figures were made with R v.3.2.2 (R Development Core Team, 2008) using the packages
645 ggplot2[83] and gplots[84]. NOG absence and presence profiles across chlamydial genomes were
646 evaluated using top NOG hits identified by eggNOG mapper[63] as described above
647 (Supplementary Data 3). A binary distance matrix of NOG presence patterns was hierarchically
648 clustered using hclust with the 'average' agglomeration method in R, and a heatmap generated
649 using heatmap.2 from the gplots[84] package (Supplementary Fig. 3). Intersection plots were
650 implemented using the R package UpSetR[85] (Fig. 3b). Synteny figures for NF-T3SS/flagellar
651 systems (Supplementary Figs. 8, 9, 10 and 11) were generated using the R package GenoPlotR[86].
652 Phylogenetic trees were visualized and edited using Figtree[87] v1.4.2 and iTOL[88]. Protein domains
653 were visualized and mapped to phylogenetic trees using iTOL[88]. Bathymetry was uploaded and
654 visualized using GeoMapApp V. 3.6.6 (http://www.geomapapp.org). Figures were made and
655 edited using Inkscape and Adobe Illustrator.

656 **Data availability**

657 Raw sequence reads for both 16S rRNA gene amplicons and metagenomes have been deposited
658 to the NCBI Sequence Read Archive repository under BioProject PRJNA504765. Whole
659 Genome Shotgun projects for metagenome assemblies GS08_GC12_126, GS10_PC15_940,

GS10_PC15_1000 and GS10_PC15_1060 have been deposited at DDBJ/ENA/GenBank under the accessions SDBU00000000, SDBV00000000, SDBS00000000 and SDBT00000000, respectively. The versions described in this paper are versions SDBU01000000, SDBV01000000, SDBS01000000 and SDBT01000000. Accessions for MAGs generated in this study can be found in Supplementary Table 2, and are linked to BioProject PRJNA504765. Files containing sequence datasets, alignments, all 16S rRNA gene amplicon OTUs and additional data generated in this study are archived at the Dryad Digital Repository: https://datadryad.org/resource/XXX.

Method References

26 Jorgensen, S. L. *et al.* Correlating microbial community profiles with geochemical data in highly stratified sediments from the Arctic Mid-Ocean Ridge. *PNAS* **109**, E2846-E2855, doi:10.1073/pnas.1207574109 (2012).

27 Jorgensen, S. L., Thorseth, I. H., Pedersen, R. B., Baumberger, T. & Schleper, C. Quantitative and phylogenetic study of the Deep Sea Archaeal Group in sediments of the Arctic mid-ocean spreading ridge. *Front Microbiol* **4**, 299, doi:10.3389/fmicb.2013.00299 (2013).

28 Hugenholtz, P., Pitulle, C., Hershberger, K. l. & Pace, N. r. Novel Division Level Bacterial Diversity in a Yellowstone Hot Spring. *Journal of Bacteriology* **180**, 366-376 (1998).

29 Klindworth, A. *et al.* Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res* **41**, e1, doi:10.1093/nar/gks808 (2013).

30 Hugerth, L. W. *et al.* Systematic design of 18S rRNA gene primers for determining eukaryotic diversity in microbial consortia. *PLoS One* **9**, e95567, doi:10.1371/journal.pone.0095567 (2014).

31 Ashelford, K. E., Weightman, A. J. & Fry, J. C. PRIMROSE: a computer program for generating and estimating the phylogenetic range of 16S rRNA oligonucleotide probes and primers in conjunction with the RDP-II database. *Nucleic Acids Res* **30**, 3481-3489 (2002).

32 Acinas, S. G., Sarma-Rupavtarm, R., Klepac-Ceraj, V. & Polz, M. F. PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. *Appl Environ Microbiol* **71**, 8966-8969, doi:10.1128/AEM.71.12.8966-8969.2005 (2005).

33 Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal* **17**, 10-12, doi:10.14806/ej.17.1.200 (2011).

34 Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahe, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584, doi:10.7717/peerj.2584 (2016).

35 Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C. & Knight, R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**, 2194-2200, doi:10.1093/bioinformatics/btr381 (2011).

36 Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41**, D590-596, doi:10.1093/nar/gks1219 (2013).

37 Lanzen, A. *et al.* CREST--classification resources for environmental sequence tags. *PLoS One* **7**, e49334, doi:10.1371/journal.pone.0049334 (2012).

38 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120, doi:10.1093/bioinformatics/btu170 (2014).

39 Andrews, S. FastQC: A quality control tool for high throughput sequence data. (2010).

40 Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072-1075, doi:10.1093/bioinformatics/btt086 (2013).

41 Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119, doi:10.1186/1471-2105-11-119 (2010).

42 Nomura, M. & Morgan, E. A. Genetics of bacterial ribosomes. *Annual Reviews Genetics* **11**, 297-347 (1977).

43 Martijn, J., Vosseberg, J., Guy, L., Offre, P. & Ettema, T. J. G. Deep mitochondrial origin outside the sampled alphaproteobacteria. *Nature* **557**, 101-105, doi:10.1038/s41586-018-0059-5 (2018).

44 Raymann, K., Brochier-Armanet, C. & Gribaldo, S. The two-domain tree of life is linked to a new root for the Archaea. *Proc Natl Acad Sci U S A* **112**, 6670-6675, doi:10.1073/pnas.1420858112 (2015).

45 Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**, 525-527, doi:10.1038/nbt.3519 (2016).

46 Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat Methods* **11**, 1144-1146, doi:10.1038/nmeth.3103 (2014).

47 Hugerth, L. W. *et al.* Metagenome-assembled genomes uncover a global brackish microbiome. *Genome Biol* **16**, 279, doi:10.1186/s13059-015-0834-7 (2015).

48 Martijn, J. *et al.* Single-cell genomics of a rare environmental alphaproteobacterium provides unique insights into Rickettsiaceae evolution. *ISME J* **9**, 2373-2385, doi:10.1038/ismej.2015.46 (2015).

49 Karst, S. M., Kirkegaard, R. H. & Albertsen, M. mmgenome: a toolbox for reproducible genome extraction from metagenomes. *bioRxiv*, doi:10.1101/059121 (2016).

50 Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol* **35**, 725-731, doi:10.1038/nbt.3893 (2017).

51 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).

52 Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313, doi:10.1093/bioinformatics/btu033 (2014).

53 Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-2069, doi:10.1093/bioinformatics/btu153 (2014).

54 Buchfink, B., Xie, C. & Huson, D. H. fast and sensitive protein alignment using DIAMOND. *Nature Methods* **12**, 59-60, doi:10.1038/nmeth.3176 (2015).

55 Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Res* **42**, D222-230, doi:10.1093/nar/gkt1223 (2014).

56 Finn, R. D. *et al.* InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res* **45**, D190-D199, doi:10.1093/nar/gkw1107 (2017).

57 Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res* **42**, D459-471, doi:10.1093/nar/gkt1103 (2014).

58 Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* **44**, D457-462, doi:10.1093/nar/gkv1070 (2016).

59 Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28**, 27-30 (2000).

60 Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236-1240, doi:10.1093/bioinformatics/btu031 (2014).

61 Kanehisa, M., Sato, Y. & Morishima, K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *J Mol Biol* **428**, 726-731, doi:10.1016/j.jmb.2015.11.006 (2016).

62 Huerta-Cepas, J. *et al.* eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res* **44**, D286-293, doi:10.1093/nar/gkv1248 (2016).

63 Huerta-Cepas, J. *et al.* Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol Biol Evol* **34**, 2115-2122, doi:10.1093/molbev/msx148 (2017).

64 Abby, S. S. *et al.* Identification of protein secretion systems in bacterial genomes. *Sci Rep* **6**, 23080, doi:10.1038/srep23080 (2016).

65 Eichinger, V. *et al.* EffectiveDB--updates and novel features for a better annotation of bacterial secreted proteins and Type III, IV, VI secretion systems. *Nucleic Acids Res* **44**, D669-674, doi:10.1093/nar/gkv1269 (2016).

66 Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-1973 (2009).

67 Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32**, 268-274, doi:10.1093/molbev/msu300 (2015).

68 Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* **14**, 587-589, doi:10.1038/nmeth.4285 (2017).

69 Le, S. Q. & Gascuel, O. An improved general amino acid replacement matrix. *Mol Biol Evol* **25**, 1307-1320, doi:10.1093/molbev/msn067 (2008).

70 Quang le, S., Gascuel, O. & Lartillot, N. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* **24**, 2317-2323, doi:10.1093/bioinformatics/btn445 (2008).

71 Le, S. Q., Dang, C. C. & Gascuel, O. Modeling Protein Evolution with Several Amino Acid Replacement Matrices Depending on Site Rates. *Molecular Biology and Evolution* **29**, 2921-2936, doi:10.1093/molbev/mss112 (2012).

72 Tavaré, S. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on mathematics in the life sciences* **17**, 57-86 (1986).

73 Williams, K. P. *et al.* Phylogeny of gammaproteobacteria. *J Bacteriol* **192**, 2305-2314, doi:10.1128/JB.01480-09 (2010).

74 Wang, H. C., Minh, B. Q., Susko, E. & Roger, A. J. Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation. *Syst Biol* **67**, 216-235, doi:10.1093/sysbio/syx068 (2018).

75 Viklund, J., Ettema, T. J. & Andersson, S. G. Independent genome reduction and phylogenetic reclassification of the oceanic SAR11 clade. *Mol Biol Evol* **29**, 599-615, doi:10.1093/molbev/msr203 (2012).

76 Lartillot, N., Rodrigue, N., Stubbs, D. & Richer, J. PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst Biol* **62**, 611-615, doi:10.1093/sysbio/syt022 (2013).

77 Criscuolo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evolutionary Biology* **10**, doi:https://doi.org/10.1186/1471-2148-10-210 (2010).

78 Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res* **44**, D279-285, doi:10.1093/nar/gkv1344 (2016).

79 Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS one* **5**, doi:doi:10.1371/journal.pone.0009490 (2010).

80 Major, P., Embley, T. M. & Williams, T. A. Phylogenetic Diversity of NTT Nucleotide Transport Proteins in Free-Living and Parasitic Bacteria and Eukaryotes. *Genome Biol Evol* **9**, 480-487, doi:10.1093/gbe/evx015 (2017).

81 Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-1973 (2009).

82 Lagkouvardos, I. *et al.* IMNGS: A comprehensive open resource of processed 16S rRNA microbial profiles for ecology and diversity studies. *Sci Rep* **6**, 33721, doi:10.1038/srep33721 (2016).

83 Wickham, H. ggplot2: Elegant Graphics for Data Analysis. *Springer-Verlag New York* (2009).

84 Gregory R Warnes, B. B., Lodewijk Bonebakker, Robert Gentleman, Wolfgang Huber, Andy Liaw, Thomas Lumley, Martin Maechler, Arni Magnusson, Steffen Moeller, Marc Schwartz, Bill Venables. gplots: Various R programming tools for plotting data. *R package version* (2009).

85 Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R. & Pfister, H. UpSet: Visualization of Intersecting Sets. *IEEE Trans Vis Comput Graph* **20**, 1983-1992, doi:10.1109/TVCG.2014.2346248 (2014).

86 Guy, L., Roat Kultima, J. & Andersson, S. G. E. genoPlotR: comparative gene and genome visualization in R. *Bioinformatics* **26**, 2334-2335, doi:10.1093/bioinformatics/btq413 (2010).

87 Rambaut, A. FigTree v1.3.1. *Institute of Evolutionary Biology, University of Edinburgh, Edinburgh.*, doi:http://tree.bio.ed.ac.uk/software/figtree/ (2010).

88 Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res* **44**, W242-245, doi:10.1093/nar/gkw290 (2016).