FEMS Microbiology Ecology, 94, 2018, fiy185

doi: 10.1093/femsec/fiy185
Advance Access Publication Date: 10 September 2018
Research Article

RESEARCH ARTICLE

The all-intracellular order *Legionellales* is unexpectedly diverse, globally distributed and lowly abundant

Tiscar Graells[1,2,†], Helena Ishak[1], Madeleine Larsson[1] and Lionel Guy[1,*,‡]

[1]Department of Medical Biochemistry and Microbiology, Science for Life Laboratory, Uppsala University, Box 582, 75123 Uppsala, Sweden and [2]Departament de Genètica i Microbiologia, Universitat Autònoma de Barcelona, Edifici C, Carrer de la Vall Moronta, 08193 Bellaterra, Spain

[*]**Corresponding author:** Lionel Guy, Department of Medical Biochemistry and Microbiology, Science for Life Laboratory, Uppsala University, Box 582, 75123 Uppsala, Sweden E-mail: lionel.guy@imbim.uu.se

One sentence summary: The all-intracellular bacterial order of Legionellales is much more diverse, prevalent and globally distributed than previously thought.

Editor: Rolf Kümmerli

[†]Tiscar Graells, http://orcid.org/0000-0002-2376-3559
[‡]Lionel Guy, http://orcid.org/0000-0001-8354-2398

ABSTRACT

Legionellales is an order of the *Gammaproteobacteria*, only composed of host-adapted, intracellular bacteria, including the accidental human pathogens *Legionella pneumophila* and *Coxiella burnetii*. Although the diversity in terms of lifestyle is large across the order, only a few genera have been sequenced, owing to the difficulty to grow intracellular bacteria in pure culture. In particular, we know little about their global distribution and abundance.

Here, we analyze 16/18S rDNA amplicons both from tens of thousands of published studies and from two separate sampling campaigns in and around ponds and in a silver mine. We demonstrate that the diversity of the order is much larger than previously thought, with over 450 uncultured genera. We show that *Legionellales* are found in about half of the samples from freshwater, soil and marine environments and quasi-ubiquitous in man-made environments. Their abundance is low, typically 0.1%, with few samples up to 1%. Most *Legionellales* OTUs are globally distributed, while many do not belong to a previously identified species.

This study sheds a new light on the ubiquity and diversity of one major group of host-adapted bacteria. It also emphasizes the need to use metagenomics to better understand the role of host-adapted bacteria in all environments.

Keywords: legionella; legionellales; metagenomics; amplicons; host-adapted bacteria; geographical distribution

INTRODUCTION

Legionellales is an order composed only of intracellular bacteria within the *Gammaproteobacteria* class. They are gram-negative, non-spore forming, rod-shaped bacteria and are classically divided into two families: the *Legionellaceae* and the *Coxiellaceae* (Garrity *et al.* 2005). In the original description, the former was described as facultative intracellular (e.g. *Legionella pneumophila*), and the latter as obligate intracellular (e.g. *Coxiella burnetii*).

In the environment, *Legionellaceae,* which includes the genus *Legionella,* can be found in natural aquatic environments, sediments and soils as a free form, but is mostly found colonizing amoeba or within biofilms (e.g. Fields 1996). They colonize man-made water systems where the temperature conditions

Received: 20 June 2018; **Accepted:** 8 September 2018

are suitable for their optimal growth. Their hosts include amoebae like *Acanthamoeba*, *Naegleria*, *Balmuthia*, *Dictyostelium* and ciliates such as *Tetrahymena* (Boamah *et al.* 2017). Several species have been described as accidental pathogens of humans (*L. pneumophila, L. longbeachae, L. micdadei*). This family has been proposed to be divided in three genera: *Legionella, Tatlockia* and *Fluoribacter*. However, this classification is not often used by microbiologists as there are no phenotypic differences between them (Garrity *et al.* 1980; Fry *et al.* 1991), and we chose to only use the *Legionella* genus in this contribution. One *Legionella* species has a totally different lifestyle: '*Candidatus* Legionella polyplacis' (hereafter referred to as *L. polyplacis*) (Rihova *et al.* 2017), which has undergone considerable genome reduction, is an obligate intracellular symbiont of the blood-sucking lice *Polyplax* spp.

The *Coxiellaceae* comprise several genera and cover a wider diversity of lifestyles. The arthropod-associated *Rickettsiella* (Leclerque 2008; Bouchon, Cordaux and Grève 2011) have a wide variety of hosts; *Diplorickettsia* (Mediannikov *et al.* 2010) and *Coxiella* (Taylor *et al.* 2012; Gottlieb, Lalzar and Klasson 2015) use ticks as hosts, except *C. burnetii*, which is an obligate intracellular bacterium infecting mammals. Amoeba-associated genera include *Aquicella* (Santos *et al.* 2003), '*Candidatus* Berkiella' (Mehari *et al.* 2016) and '*Candidatus* Cochliophilus' (Tsao *et al.* 2017). *Diplorickettsia massiliensis* (Subramanian *et al.* 2012) and *Coxiella burnetii* have been described as human pathogens (van Schaik *et al.* 2013).

Despite their very broad ecological range, *Legionellales* have significant common characters: they replicate and multiply inside eukaryotic hosts, using a type IVB secretion system (T4BSS). This system, also known as Icm/Dot (intracellular multiplication / defect in organelle trafficking genes), is used to inject effector proteins inside the host (Nagai and Kubori 2011; Christie, Gomez Valero and Buchrieser 2017). This virulence trait is key to avoid lysosomal degradation and to replicate inside intracellular compartments (Richards *et al.* 2013). Imitating different functions of cells in their infection biology cycle has likely contributed to the infection of cattle and human macrophages (Richards *et al.* 2013).

Legionellales seem to be widely distributed but because of their complicated life cycle they have often been unnoticed. Due to their intracellular lifestyle and dependency on their host, growing them in a laboratory setting is challenging. So far, only some species of *Legionella*, *Aquicella* (Santos *et al.* 2003) and *Coxiella burnetii* can be cultivated in axenic media; the development of a protocol for the latter took decades and tremendous efforts (Omsland 2012). Hence, studies have investigated the microbiology composition of soil, sediments and water with independent-culture methods, primarily through amplicon sequencing. For example, significant amounts of *Legionellaceae* have been found in cold waters (Wullings and van der Kooij 2006), even in Antarctica lakes (~0°C) (Carvalho *et al.* 2008). *Legionella* have also been found widely distributed in watersheds but with relatively low abundance (2.1%) (Peabody *et al.* 2017). The same study found a negative correlation between abundance of both bacteria and hosts and human activity, i.e. a higher abundance of *Legionella* and amoebae in pristine environments compared to agricultural soils. Treatment of drinking water with chlorine tends to reduce the abundance of *Legionella*, but higher abundances were restored further away in the supply chain, with phylotypes and abundance differing between cold and warm tap water (Lesnik, Brettar and Hofle 2016). In general, *Proteobacteria* were within the most common bacteria in soils and aquatic environments in different countries (Denet *et al.* 2017; Hosen *et al.* 2017; Naghoni *et al.* 2017; Peabody *et al.* 2017). Many of the predominant amoebae in those soils are *Tetramitus*, *Acanthamoeba* and *Naegleria* (Denet *et al.* 2017; Peabody *et al.* 2017) known to be hosts for *Legionellales*. Surprisingly, *Legionellales* seem to be abundant even in hypersaline environments where archaea, other *Gammaproteobacteria, Firmicutes* and *Bacteroidetes* are otherwise predominant (Naghoni *et al.* 2017).

The microbial diversity of natural environments can be affected by different factors. The global temperatures rising can lead to changes in abundance of certain microorganisms and protists. Human activity has shown to affect water environments, modifying the microbial diversity between forest and urban areas where microbes play key roles in biogeochemical cycles (Hosen *et al.* 2017). The prevalence of vector-borne diseases on the rise (Rosenberg *et al.* 2018; Semenza and Suk 2018) and the amoebae as a potential vector for emerging pathogens (Lamoth and Greub 2010) motivate the need for a global study of the distribution of the exclusively host-adapted *Legionellales*.

Here, the environmental and geographical distribution, as well as the prevalence of the *Legionellales* was studied, both by using publicly available datasets and by analyzing samples taken in different kinds of wetlands and in a silver mine in Sweden. The aim was to better understand the global ecology of this order to predict responses to environmental changes and identify the mechanisms that affect their microbial biodiversity.

MATERIAL AND METHODS

Collection and preparation of environmental samples

A total of 45 water, sediment and soil samples were collected from areas in and around Hedesundafjärden natural reserve (12 samples), Florarna natural reserve (12), Färnebofjärden national park (12) and Stadsskogen natural reserve (9) (Supplementary Table 1) in Uppland, Sweden, during the months of July and August 2016. These samples are referred to as the 'Uppland samples'. In general, samples were collected in duplicates. In a separate sampling campaign, 12 samples were retrieved from different levels and rooms of the Sala silver mine (Sala, Sweden) in April 2017 (Supplementary Table 1). These are referred to as the 'Sala samples'.

To retrieve water, 1 L sterilized glass bottles were immersed halfway as to mainly collect surface water. Sediment and biofilm was acquired by scooping the top layers of the sediment with 50 ml, sterile Falcon tubes. Soil samples were collected using a soil sampler, digging 10–15 cm into the ground. Temperature was measured. The samples were then kept cold during transportation. Water samples were filtered first through 100 μm pore filters to remove large debris such as dust, small insects and large particles. Filtered water was then re-filtered through Whatman filters with a pore size of 2 μm to obtain microorganisms on the filter papers. Sala samples were also filtered a third time with Whatman filters of 0.2 μm pore size to recover even smaller microorganisms. Filtering the Uppland water samples with 0.2 μm filters was not possible due to the higher turbidity of these samples. Samples where the water was very turbid with organic matter were centrifuged at 14 000 x g for 10 minutes to pellet microorganisms.

DNA extraction

For water samples, filters were resuspended in 1 ml of sterile ultrapure water and cut to small pieces, ranging in size of 2–6 mm; parts of the filters and 200 μl of the water were used for extraction. For other samples, 0.5 g of soil or sediment were used. DNA was isolated from the raw material or the filters with the FastDNA® SPIN Kit for Soil and the FastPrep® Instrument (MP Biomedicals, Santa Ana,CA). For water samples ML_10 001 to ML_10 012, no DNA could be retrieved.

Quality control of DNA extraction

Purity control and quantification of raw DNA were performed using a Nanodrop 1000 Spectrophotometer (Thermo Fischer). The 260/280 nm and 260/230 nm ratios were controlled to be within an acceptable range. Since environmental samples may contain PCR inhibitors such as proteins or phenols, the extracted DNA was diluted to reach 1–3 ng/µl to minimize problems in the following PCR.

Two-step polymerase chain reaction

In order to create a 16S rRNA amplicon library, a two-step PCR was used. The first PCR reaction uses two primers that contain an adaptor and a universal primer (Supplementary Table 2) to amplify 16S/18S rDNA genes of the extracted DNA samples, using HotStar Taq polymerase (Qiagen). The PCR ran through 28 cycles and conditions were set to initial denaturation at 95°C for 15 minutes, denaturation at 94°C for 30 seconds, annealing at 57°C for 45 seconds, elongation at 72°C for 1 minute and 20 seconds, final elongation at 72°C for 7 minutes and then resting/cooling at 4°C until retrieval. Prior to the second PCR step, PCR products were checked in an agarose gel, the amplicons were purified using the protocol for GeneJET Gel extraction Kit (Thermo Fischer) and finally quantified using Qubit dsDNA HS (High Sensitivity) Assay Kit with Qubit Fluorometer. Concentration of DNA samples was adjusted to 5 ng/µl. The second PCR step used primers comprising Illumina sequencing primers, barcodes and an adaptor matching the one of the first PCR primer (Supplementary Table 2). The PCR ran for 10 cycles and conditions were set to initial denaturation at 95°C for 15 minutes, denaturation at 95°C for 20 seconds, annealing at 61°C for 30 seconds, elongation at 72°C for 90 seconds, final elongation at 72°C for 7 minutes and rest/cooling until retrieval. Again, PCR products were checked in an agarose gel, purified with GeneJET Gel extraction Kit (Thermo Fischer) and quantified using Qubit dsDNA HS (High Sensitivity) Assay Kit with Qubit Fluorometer.

DNA quality control and sequencing

Samples were pooled together and the final quality control before sequencing was done using High Sensitivity D5000 ScreenTape Assay for Agilent 4200 TapeStation System which quantifies and distributes the DNA molecules by different length ranges.

Purified amplicon libraries were sequenced on the Illumina MiSeq, with 2 × 300 bp setting. All sequencing was performed by NGI, SciLifeLab, Uppsala and Stockholm, Sweden.

Quality control and trimming

The quality of each library was assessed with FastQC v0.11.3 (Andrews 2010). Results were summarized with MultiQC 0.9 (Ewels *et al.* 2016). After demultiplexing and adapter trimming, reads were trimmed with Trimmomatic 0.35 (Bolger, Lohse and Usadel 2014) with the following parameters: MAXINFO:200:0.5. Remaining adaptors were removed with SeqPrep v1.3.2 (St. John 2011). Demultiplexed, untrimmed reads are deposited at ENA under study accession PRJEB26992.

Amplicon analysis

Amplicons were analyzed with IM-Tornado 2.0.3.3 (Ewels *et al.* 2016). Due to stringent criteria and the relatively low quality of sequences for the reverse read, the initial use of both read ends yielded too few results, and we used only the forward read. Taxonomic attribution was done against SILVA SSU Ref release 128 (Yilmaz *et al.* 2014). We used the following settings: MINIMUM_LENGTH = 180; R1_TRIM = 250; R2_TRIM = 180. Clustering (within IM-Tornado) was performed with VSEARCH v2.3.4 (Rognes *et al.* 2016) and preliminary trees run with FastTree 2.1.8 (Price, Dehal and Arkin 2010). The per centage of how much various clades would be identified by the chosen primers was calculated online using TestPrime 1.0 available at the SILVA website (https://www.arb-silva.de/search/testprime/) (Klindworth *et al.* 2013).

Wherever possible, we favored SILVA over greengenes, because (i) the latter does not include eukaryotes, and (ii) the inclusion of other families (e.g. *Francisellaceae*) than *Legionellaceae* and *Coxiellaceae* in the *Legionellales* does not follow the traditional taxonomy of *Gammaproteobacteria* nor is sufficiently supported by multigene phylogenies (Williams *et al.* 2010). The vast majority of the analysis available at EBI metagenomics (v. 2-v. 3.1) are unfortunately based on greengenes 13.8. It is difficult to assess what effect using SILVA instead of greengenes would have on the results presented here without actually reanalyzing all EBI samples, which is beyond the scope of this contribution. However, several facts suggest that the differences between the two taxonomic attributions would be limited: (i) the number of *Legionellales* OTUs in both databases is similar; (ii) the phylogenetic breadth of *Legionellales* is well covered by well-known species (*Legionella*, *Coxiella*, *Aquicella*), and there are not many deep-branching groups where no sequence is known and which would be more difficult to correctly attribute and (iii) the trees inferred from both databases are fairly congruent for the *Legionellales*.

Analysis of publicly available data

Basic data handling, including the interaction with the RESTful API at EBI Metagenomics (Mitchell *et al.* 2018) was performed in python 3.6, with the help of the pandas library (McKinney 2010). The results were analyzed and displayed in R (R Development Core Team 2017), with the help of the ggplot2 package (Wickham 2009).

Basic information under the form of spreadsheet about all available samples were retrieved from EBI Metagenomics (Mitchell *et al.* 2016) in April 2018, representing 90 861 samples in 1687 projects. Basic information about all sequencing runs (n = 110 584) was also retrieved as a spreadsheet. For each sample, if applicable, the following basic metadata was retrieved: project with which the sample is associated; project name and description; biome to which the sample belong; what feature and material the sample consisted of; latitude and longitude of sampling; temperature. This information was (at least partially) available for 87 955 samples. A representative sequencing run was also selected by choosing, among the runs derived from this sample that had at least one *Legionellales* read, the one that contained most OTUs. This way, 20 972 samples (referred to as 'positive samples' thereafter) could be linked to a sequencing run that contained at least one *Legionellales* run.

This procedure was repeated for nine other gammaproteobacterial orders: *Alteromonadales*, *Chromatiales*, *Enterobacteriales*, *Oceanospirillales*, *Pasteurellales*, *Pseudomonadales*, *Thiotrichales*, *Vibrionales* and *Xanthomonadales*.

For each sample positive for any of the 10 gammaproteobacterial orders, an OTU table corresponding to the representative sequencing run was downloaded. If the taxonomic attribution had been performed using several versions of the analysis pipeline, the version 3 or 2 were preferred, because taxonomic attribution is done with the same database (greengenes 13.8), and the OTU ids can be compared. The following metrics were calculated for each representative run: total number of reads for which a taxonomic attribution was available and total number of OTUs in the sample; number of reads that were attributed to *Legionellales* and number of OTUs belonging to *Legionellales*; OTU id and number of reads belonging to the five most abundant *Legionellales* OTUs in this run. It should be noted that in greengenes 13.8, but not in SILVA 128, the families *Francisellaceae* and *Endoecteinascidiaceae* are included in the order *Legionellales*.

To test the effect of temperature on the abundance of *Legionellales*, we calculated the Spearman's correlation coefficient, per biome, using all samples for which the temperature had been recorded, and the non-logarithmically transformed fraction of reads belonging to *Legionellales*. We performed the test only for biomes with temperature data for >10 samples.

The table containing the summarized information for each sample and representative run, as well as most of the code necessary to run the analysis of the public data is available https://bitbucket.org/evolegiolab/legionellalesabundancedata/

SSU rDNA phylogeny

We retrieved all 16S rDNA sequences from SILVA SSU Ref release 128 (Yilmaz *et al.* 2014) that were attributed to the order Legionellales, whose quality was > 90 and that were 900 nt or longer. After a first round of alignment with mafft–linsi (Katoh and Toh 2008) and maximum-likelihood phylogeny inference with FastTree 2.1.8 (Price, Dehal and Arkin 2010) under a GTR substitution matrix, 16 sequences with very long branches were removed from the pool, yielding a set of 2433 sequences. To this pool, we added: (i) representative *Gammaproteobacteria* (82 sequences), representatives for the OTUs obtained from the amplicon libraries from (ii) the Uppland samples (66 sequences) and (iii) the Sala samples (42 sequences). The final pool of sequences was re-aligned with mafft-linsi and a maximum-likelihood tree was inferred with IQ-TREE v. 1.5.3 (GTR+I+Γ4) (Nguyen *et al.* 2015).

To estimate the amount of species and genera in the order *Legionellales*, we clustered the 2433 sequences filtered from Silva 128 with mothur 1.39.1 (Schloss *et al.* 2009), using the dist.seq method with default parameters and clustering then with 0.03 (97% identity) and 0.05 (95% identity) as cut-off, respectively. We also downloaded the taxonomy attributions from greengenes 13.8 (McDonald *et al.* 2012) and filtered the OTU id belonging to the *Legionellales* and to the other selected gammaproteobacterial orders.

Table 1. Number of OTUs in gammaproteobacterial orders, at different cutoffs, according to greengenes 13.8. The rows were ordered by decreasing number of OTUs at 94% similarity. 'NA' represents OTUs for which no taxonomic attribution could be made at order level. The order *Legionellales* is shown in bold.

Order	Cutoff 94%	Cutoff 97%	Cutoff 99%
[Marinicellales]	1760	3918	7615
Legionellales	**535**	**834**	**1042**
Alteromonadales	439	1261	2699
Oceanospirillales	368	846	1831
Chromatiales	340	755	1396
Pseudomonadales	337	1073	3418
Enterobacteriales	276	938	3490
Xanthomonadales	263	794	2029
Thiotrichales	249	611	1331
Vibrionales	129	394	1139
Pasteurellales	81	333	1205
Methylococcales	72	194	362
Aeromonadales	61	204	635
HTCC2188	50	78	115
Thiohalorhabdales	39	87	147
HOC36	37	68	111
Cardiobacteriales	27	77	186
Acidithiobacillales	16	32	101
34P16	8	16	31
Salinisphaerales	7	11	18
RCP1–48	5	8	9
PYR10d3	4	13	41
NA	46256	99322	203452

RESULTS

Diversity and abundance of *Legionellales* in public datasets

We estimated the number of uncultivated genera and species belonging to the order *Legionellales* by clustering publicly available rRNA sequences from the ribosome small subunit (hereafter referred to as 16S). We chose conservative clustering cutoff values (Stackebrandt and Goebel 1994) for delineating genera (95%) and species (97%), respectively. The cutoff to discriminate between species was more recently increased by the same authors to 98.5% (Stackebrandt and Ebers 2006), while others claim this value should be even higher for human-associated pathogens (Rossi-Tamisier *et al.* 2015). All 16S reads published in Silva 128 and classified in the *Legionellales* were clustered at 95 and 97%, resulting in 462 and 756 OTUs, respectively. These estimations gathered from Silva are consistent with the 535 and 834 *Legionellales* OTUs clustered at 94 and 97%, respectively, in the greengenes database v. 13.8. In comparison with other gammaproteobacterial orders (Table 1), *Legionellales* had the second-highest number of OTUs at 94% similarity, and the sixth-highest at 97% similarity.

We further investigated the hidden diversity of *Legionellales*, as well as their abundance, by analyzing the vast quantity of data deposited at EBI metagenomics (Hunter *et al.* 2014; Mitchell *et al.* 2016; Mitchell *et al.* 2018). Among the 87 955 samples for which we could retrieve basic metadata at the time of the analysis, we were able to select a sequencing run containing at least one *Legionellales* read for 20 971 samples (*Legionellales*-positive runs, LPRs; 22.6%). Further, 10.4%, 2.96% and 0.47% contain at least 10, 100 or 1000 reads attributed to *Legionellales*, respectively. Four types of experiments are available at EBI metagenomics:

amplicon (most generally 16S; 17 766 runs), metagenomic (2736 runs), metatranscriptomic (461 runs) and assembly (8 runs), in decreasing numbers. We chose to discard the 8 runs of type ‘assembly’; we also discarded approximately 100 samples for which the metadata was clearly erroneous. Among LPRs, the number of reads for which a taxonomic attribution is available spans 9 orders of magnitude, with two clear peaks around 8000 and 80 000 reads per run, irrespective of the type of experiment (Fig. 1A). The number of operational taxonomic units (OTUs), which is a proxy for the number of species in a sample, ranges from one to several millions, with a clear peak around 1000 (Fig. 1B). The fraction of reads attributed to *Legionellales* ranges from 0 to close to 1, with a peak at 10^{-3} (Figs 1C, 1E); the number of *Legionellales* OTUs reaches 1000, although most LPRs harbors between 1 and 10 *Legionellales* OTUs (Fig. 1F).

Compared to other gammaproteobacterial orders, *Legionellales* are found in an average number of samples, with *Pseudomonales*, *Enterobacteriales*, *Xanthomonadales* and *Alteromonadales* being found in more samples than *Legionellales* (Fig. 1D). The distribution of the fraction *Legionellales* is, on the other hand, very peculiar, with a very sharp peak around 0.1% (Fig. 1E), whereas the other orders had a more uniform distribution, except for *Xanthomonadales*, which seem to represent between 1 and 5% of the reads in most samples. *Legionellales* are the least common gammaproteobacterial order with samples 1% and above (Fig. 1E), but still exhibits an average diversity of OTUs (Fig. 1F).

Legionellales are represented differently in different environments, or biomes, and in variable proportions (Fig. 2). The number of biomes represented in this study amounts to 220, making it impractical to study all of them separately. To reduce this complexity, we took advantage of the hierarchical nature of the GOLD biome naming (Mukherjee *et al.* 2017): biomes that were represented by only a few samples, or that were generally irrelevant for our study, were included in their parent category; parent categories do not include child categories that have been kept separate. For example, all human samples were collapsed in the Host-associated:Human category, except the Host-associated:Human:Respiratory system, which was considered as relevant per se; samples in the latter category are not included in the former one. This way, the number of biomes was reduced to 25 (Supplementary Table 3; Fig. 2).

LPRs are found in the majority of engineered environments, culminating in built environments, where 96.3% of all samples were *Legionellales*-positive. In general, host-associated samples contain proportionally less *Legionellales*, with the exception of plants (41%) and mollusks (78%). The high prevalence of *Legionellales* in mollusks is surprising but may be the result of a bias introduced by one large unpublished study with many samples almost all containing *Legionellales*. Perhaps less surprisingly, 60% and 41% of samples taken from freshwater and soil, respectively, contain *Legionellales*. Over 16% of drinking water samples contain *Legionellales*. The samples displaying the largest fractions (Fig. 2a) and highest numbers of OTUs (Fig. 2b) from *Legionellales* come from aquatic environments and from soil and plants. A few samples from aquatic biomes count almost exclusively *Legionellales*, but these come from a study using a method specifically targeting the *Legionella* genus.

In comparison with other gammaproteobacterial orders, *Legionellales* are present in lower fractions, and with fewer OTUs in most biomes (Supplementary Fig. S1). There are however a number of exceptions: in the built environment and in aquatic (particularly freshwater and drinking water) biomes, *Legionellales* are often among the more present and more diverse gammaproteobacterial orders.

Effects of temperature on the abundance of *Legionellales*

We investigated the effect of temperature on the abundance of *Legionellales* in different environments (Fig. 3). Unfortunately, the number of samples for which the temperature was indicated was limited: only 4074 out of >90 000 samples. Despite that, trends emerge from three of the five environment groups for which enough data was available. In the soil, there seems to be a negative correlation between temperature and *Legionellales* abundance (Spearman’s rho = -0.551; *P*-value = 2.2e-18). Despite what the LOESS curve show, negative correlation are also found in freshwater and in engineered biomes, but are significant only for the latter (Spearman’s rho = -0.131, *P*-value = 0.16 and Spearman’s rho = −0.624, *P*-value = 6e-7, respectively). In mollusk-associated samples, the correlation was positive, with an increase of the abundance of *Legionellales* with the temperature, although with a low rho coefficient (Spearman’s rho = 0.228, *P*-value = 4.9e-10). Removing the *Legionellales*-negative samples did not alter significantly the results above, except for the engineered biome, where the *P*-value increased over 0.05. It should be stressed that the spread of the abundance values is very wide, and that the significance of the correlation coefficients over the whole temperature range has to be taken with caution. Correlation effects might only be found over shorter ranges, as approximated (but not statistically supported) by the LOESS curves.

Environmental distribution of the most abundant *Legionellales*

To gain further resolution on how the different sub-clades of *Legionellales* are distributed, we analyzed, for each sample, the most abundant *Legionellales* OTUs, hereafter referred to as MALOs. We considered the top five MALOs (5MALOs) for each LPR and retrieved their lowest credible taxonomic attribution from greengenes. A total of 804 OTUs are found among 5MALOs for all samples, out of a total of 1042 OTUs (77.2%) available in greengenes (clustered at 99% identity). The distribution of these OTUs is very skewed (Supplementary Fig. S2), with a dozen OTUs being present in the 5MALOs of 500 samples or more; 85 OTUs in the 5MALOs of >200 LPRs, and the majority the 5MALOs of a few samples only.

Among the 25 known *Legionella* species represented in greengenes, 17 are found among the 5MALOs in this study (Supplementary Table 3). The most frequently found known *Legionella* species are *L. pneumophila* (split in two OTUs; found in the 5MALOs of 239 samples), *L. dresdenensis* (in the 5MALOs of 108 samples) and *L. jeonii* (in the 5MALOs of 57 samples) (see Supplementary Table 3 for the other species). Interestingly, *L. pneumophila* ranks 125th among the OTUs most frequently found among 5MALOs.

The distribution of 5MALOs reveals that MALOs are very variable across biomes and show biome-specific patterns (Fig. 4). Although the clustering seems to be mostly influenced by the total abundance of *Legionellales* in the biome, some trends are visible: the biomes associated with plants and soil cluster together, while the marine biome is isolated. Most of the animal-associated biomes, except mammals, were grouped in a larger cluster.

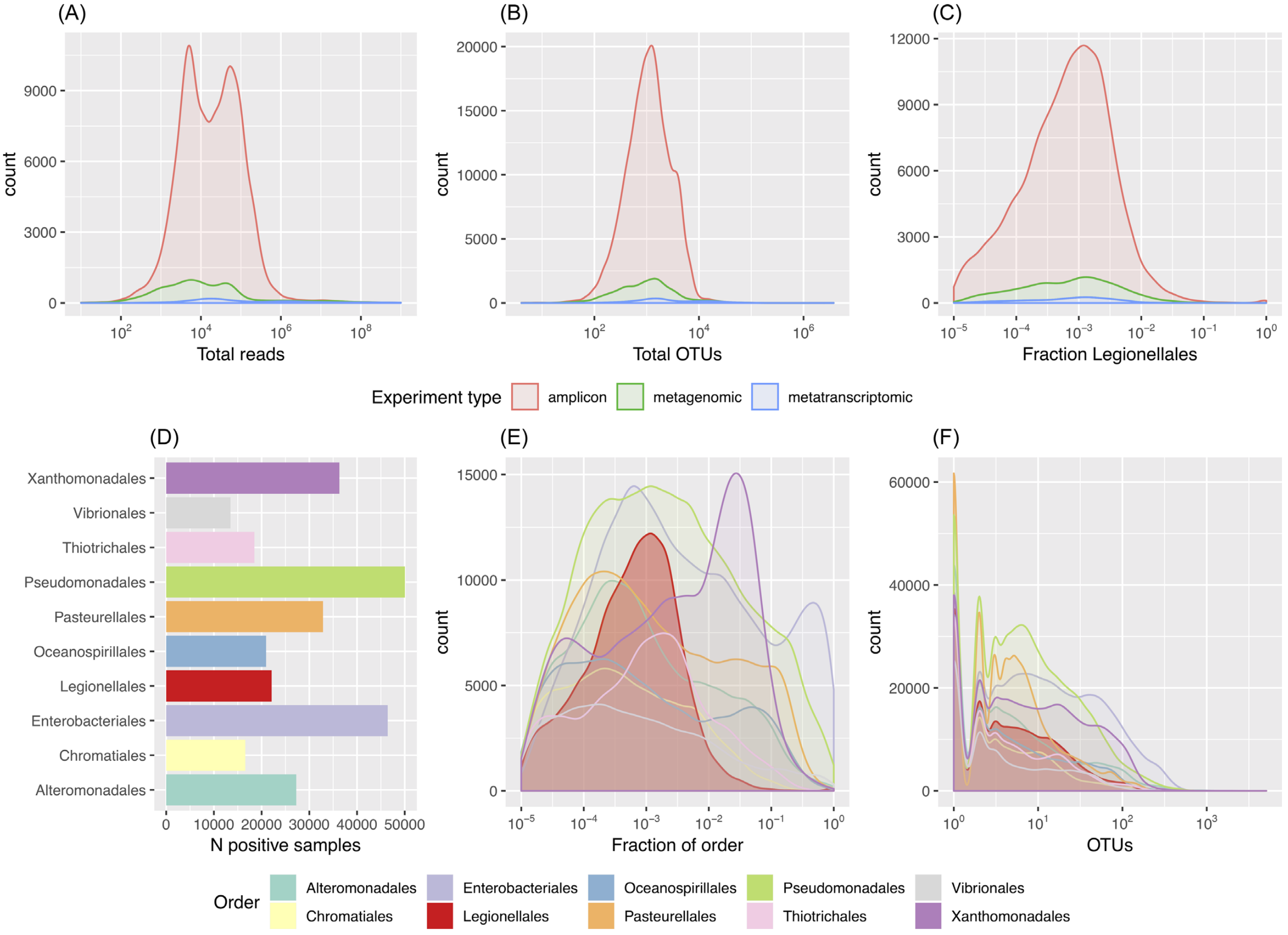

Figure 1. Distribution of samples across experiment types and gammaproteobacterial orders. In all panels except D, x scales are logarithmic and y-axes show the number of samples for that given number of reads. Experiment types (panels A–C) according to the legend right below; order (D–F) according to the lower legend. Distributions, per experiment type, of (A) total number of reads per run, (B) total number of OTUs per run and (C) fraction of reads attributed to *Legionellales*. Number of positive samples per order (D). Distributions, per gammaproteobacterial order, of (E) the fraction of reads attributed to the order, and number of OTUs belonging to the order (F).

Looking at the fraction of the nine identifiable genera among OTUs (Supplementary Fig. S3) across biomes reveals that in almost all biomes, most OTUs could not be attributed to a known genus. It should be noted that the genera *Fangia* and *Caedibacter*, as well as the family *Francisellaceae* (*Francisella* and '*Candidatus* Nebulobacter') are classified as belonging to the order *Thiotrichales*, according to LSPN (Parte 2018). Greengenes classifies however these genera inside the *Legionellales*, based on phylogenetic evidence. Among the OTUs for which an identifiable genus is available, *Legionella* is dominating in most biomes. The exceptions are in soil and plant-associated microbiomes: there, the most frequently encountered genus is *Aquicella*, which are probable facultative intracellular bacteria, found to grow in *Hartmannella* amoebae (Santos *et al.* 2003). In several host-associated biomes, the genus *Rickettsiella* is also abundant. *Rickettsiella* consists in majority of arthropod pathogens, but also includes insect symbionts (Leclerque 2008; Tsuchida *et al.* 2010). It is interesting to note that in marine environments, only a few MALOs could be attributed to known genera, despite the large number of OTUs and the relatively high abundance of *Legionellales* there (Fig. 2).

The low abundance but large diversity of *Legionellales* are also observed in geographically close but environmentally distant biomes (Table 1, Fig. 5; Supplementary Fig. S4). The analysis, with similar methods, of samples taken from water, sediments and soil in or around ponds in Uppland (Sweden), and of samples taken from biofilm and sediment in a disused silver mine near Sala (Sweden), shows that the abundance is more variable within sampling locations than across (Table 2; Supplementary Fig. S5). Even in environments where the temperature is low like the mine, the diversity, in terms of OTUs, is very large, ranging from 1 to 46 in the Uppland samples, while in the Sala samples it ranges from 13 to 52 (Table 2). It was also noticeable on a phylogenetic tree: both sampling campaigns had OTUs covering the largest part of the diversity of the order, although very few were from the *Coxiella* genus.

No reads from free-living amoebae were detected in any of the samples. It should be noted that the universal primers used in this study, while detecting the most common hosts of *Legionella* (*Acanthamoeba*, *Hartmannella*, *Dictyostelium*, etc.), tend not to recognize a large fraction of the free-living amoebae (e.g. *Naegleria*), which are potential hosts for *Legionellales* (Scheikl *et al.* 2014). Interestingly, however, the overall per centage of *Legionellales* reads was higher in the Sala samples (0.22%) than in the Uppland samples (0.09%), whereas the per centage of eukaryotic reads was lower in the Sala samples (0.14%) than in the Uppland samples (0.55%). The total number of eukaryotic OTUs was also significantly lower in the Sala samples (14) than in the Uppland samples (133).

Table 2. Abundance and diversity of *Legionellales* OTUs in the Uppland and Sala samples. *Legionellales* is abbreviated Leg-ales.

Location	Type	Sample	Reads	Percentage reads		OTUs		
			Total	Leg-ales	Eukaryotes	Total	Leg-ales	Eukaryotes
Uppland samples			2 885 349	0.087%	0.545%	8140	66	133
Färnebo-fjärden	sediment	ML_10 013	44 433	0.045%	0.223%	3437	15	7
		ML_10 014	28 485	0.035%	2.580%	3085	7	7
		ML_10 015	955	0.000%	0.733%	473	0	4
		ML_10 016	529	0.000%	4.159%	303	0	5
	soil	ML_10 025	11 507	0.052%	0.009%	1362	4	1
		ML_10 026	27 852	0.093%	0.068%	1603	12	5
		ML_10 027	3670	0.163%	0.218%	720	4	3
		ML_10 028	404	0.248%	0.248%	271	1	1
Florarna	sediment	ML_10 021	119 486	0.020%	0.357%	4352	8	23
		ML_10 022	5938	0.000%	0.236%	1680	0	6
		ML_10 023	98 287	0.012%	1.389%	4239	7	30
		ML_10 024	382 931	0.027%	0.739%	5346	21	51
	soil	ML_10 033	86 820	0.141%	0.016%	2226	19	4
		ML_10 034	62 641	0.198%	0.268%	1696	17	4
		ML_10 035	99 221	0.134%	0.093%	3289	16	9
		ML_10 036	103 665	0.129%	0.070%	3193	13	7
Hedesunda-fjärden	sediment	ML_10 017	71 409	0.069%	1.603%	3095	12	17
		ML_10 018	51 881	0.066%	4.171%	2957	13	22
		ML_10 019	25 806	0.016%	0.058%	2625	2	8
		ML_10 020	185 014	0.017%	0.268%	4274	6	23
	soil	ML_10 029	359 099	0.175%	0.589%	3936	46	27
		ML_10 030	170 239	0.190%	0.490%	2805	30	15
		ML_10 031	151 105	0.054%	0.320%	4021	16	9
		ML_10 032	16 478	0.024%	0.012%	1529	2	2
Stadsskogen	sediment	ML_10038	50 761	0.114%	0.099%	2170	7	10
		ML_10 039	51 958	0.102%	0.756%	2573	4	19
	soil	ML_10 040	55 824	0.199%	0.063%	2466	31	9
		ML_10 041	21 974	0.205%	0.014%	2081	10	2
		ML_10 042	300 647	0.115%	0.232%	2916	33	23
		ML_10 043	80	0.000%	0.000%	69	0	0
	water	ML_10 037	46 616	0.017%	2.287%	738	4	23
		ML_10 044	248 755	0.005%	0.146%	567	1	4
		ML_10 045	879	0.683%	0.000%	117	1	0
Sala samples			*10 070 182*	0.222%	0.136%	2842	123	14
Grisen, Johan/Liljeborg	sediment	TG_1002	730 652	0.124%	0.032%	384	13	3
		TG_1003	1 189 147	0.475%	0.002%	1174	49	2
		TG_1004	946 398	0.565%	0.002%	1243	52	3
Kanslern	sediment	TG_1005	1 053 635	0.062%	0.008%	1321	40	5
		TG_1007	1 047 597	0.004%	0.000%	639	13	0
		TG_1008	1 073 253	0.430%	0.003%	648	25	4
Rödstjärten	water	TG_1009	1 004 503	0.109%	0.001%	1138	27	2
Victoria Salen	sediment	TG_1011	994 448	0.073%	1.323%	737	27	3
Ribbings schakt	water	TG_1012	983 727	0.163%	0.017%	1055	35	4
		TG_1013	1 046 822	0.166%	0.000%	957	46	0

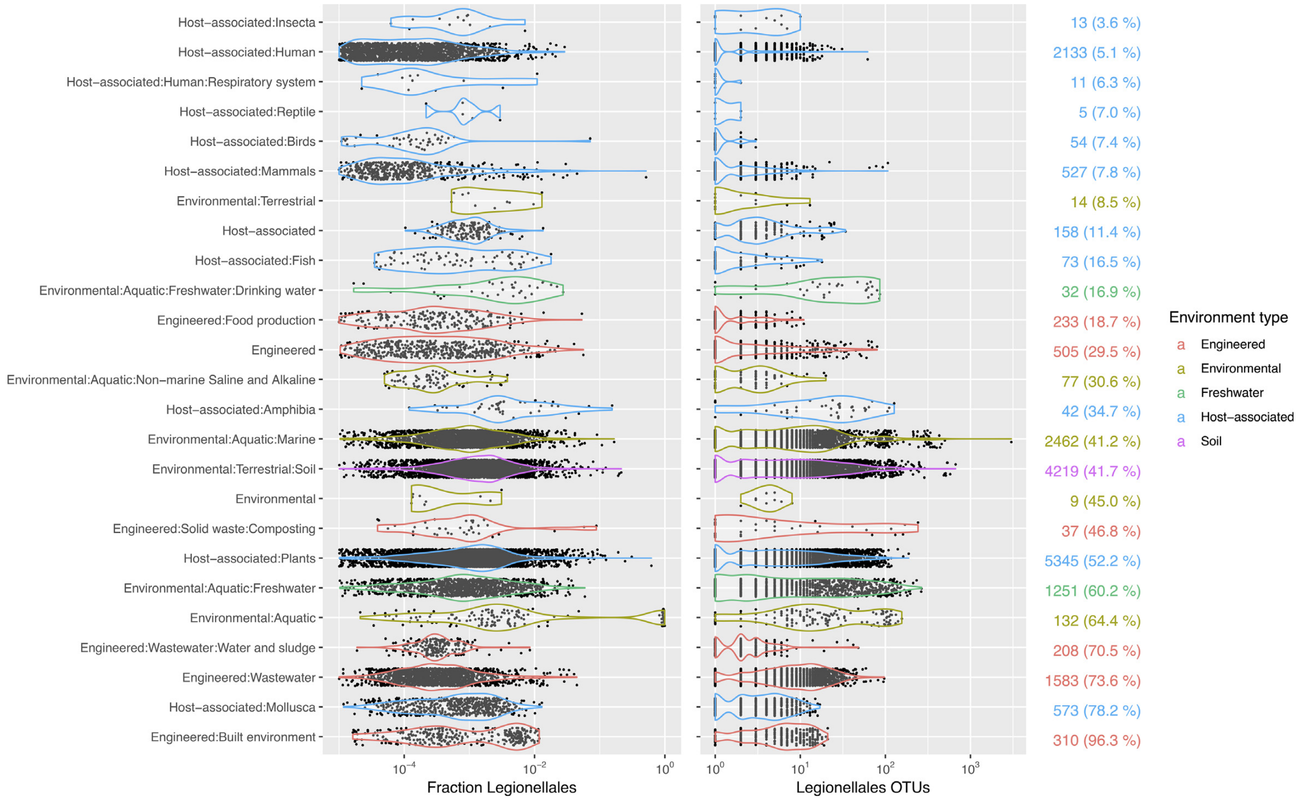

Figure 2. Relative abundance and diversity of *Legionellales* OTUs in different biomes. The left panel (violin plots) represents the fraction of *Legionellales* reads in samples containing at least one *Legionellales* read in a representative run (LPRs, 20 014 out of 87 940 samples or 22.6%). The x scale (logarithmic) extends from 10^{-5} (1 in 10 000 reads) to 1. The right panel displays the number of *Legionellales* OTUs per LPR. The right column indicates the number of positive samples (i.e. samples with at least one *Legionellales* OTU) in that biome and in the categories that have been collapsed into this one (but not the descendant categories that were kept separate), and what per centage of the total samples for that biome it represents. Colors according to the group of biomes. The rows are sorted by increasing fraction of positive samples in that biome (top to bottom).

Geographic distribution of *Legionellales*

Legionellales are globally distributed, with few exceptions (Fig. 6, Supplementary Fig. S6). This is particularly pronounced for land samples (including freshwater), where all continents harbor LPRs, including very cold (Svalbard, Antarctica) and warm climates (Fig. 6; Supplementary Fig. S6). *Legionellales* are also present in all oceans and seas, although they seem to be almost absent from the southern Pacific Ocean, and relatively rare in the northernmost latitudes (Fig. 6; Supplementary Fig. S6). *Legionellales* were also present globally in man-made environments (Fig. 6; Supplementary Fig. S6). *Legionellales* display a similarly broad geographical distribution as other, larger gammaproteobacterial orders (Supplementary Fig. S7). In many biomes, its distribution can be compared to that of *Enterobacteriales* or *Pseudomonadales*, which are the two most commonly found orders (Fig. 1D). *Legionellales* are more globally distributed than *Vibrionales* and *Pasteurellales*, especially in terrestrial samples.

The majority of the most commonly found OTUs is also globally distributed (Supplementary Figs S8 and S9), although a higher level of geographical clustering is observable for some OTUs. For example, the most commonly present OTU (id: 252 003) is mostly present in the northern hemisphere, and in a few cases in the southernmost latitudes of the southern hemisphere. Most of these other OTUs are found on all continents, at all latitudes, and in several types of environments. Among the less frequently found, the level of ubiquity decreases and some more specific OTUs appear (Supplementary Fig. S9).

Incidentally, the most commonly found OTUs among only the most abundant *Legionellales* OTU (1MALO) and among the 5MALOs are largely congruent: the top three are the same, although in a slightly different order: in the 1MALOs, 252 003 and 1 107 824 (ranking first and third among 5MALOs) share the same number of occurrences (763), and 838 066 (second among 5MALOs) ranks third. Among the top 10 OTUs, 8 are found in both lists.

DISCUSSION

In this study, we explored the abundance and distribution (both geographic and environmental) of the gammaproteobacterial order *Legionellales*, and show that this order is (i) more diverse than previously thought, (ii) quasi-ubiquitous, even in environments that are not considered as their primary niches, like marine environments, (iii) rare and typically present in 0.1% of samples. We also show that *Legionellales* are almost as abundant and globally distributed as larger orders of *Gammaproteobacteria* like *Enterobacteriales* and *Pseudomonadales*, which include a much larger number of described genera.

In contrast to most bacterial orders, *Legionellales* are relevant to study at order level: they share traits very likely acquired by their last common ancestor (synapomorphies), not the least their shared intracellular lifestyle (e.g. Qiu and Luo 2017). On the molecular level, the last common ancestor of *Legionellales* most probably acquired the type IV B secretion system (T4BSS, also referred to as Dot/Icm) that allows *Legionella* and *Coxiella* (Segal,

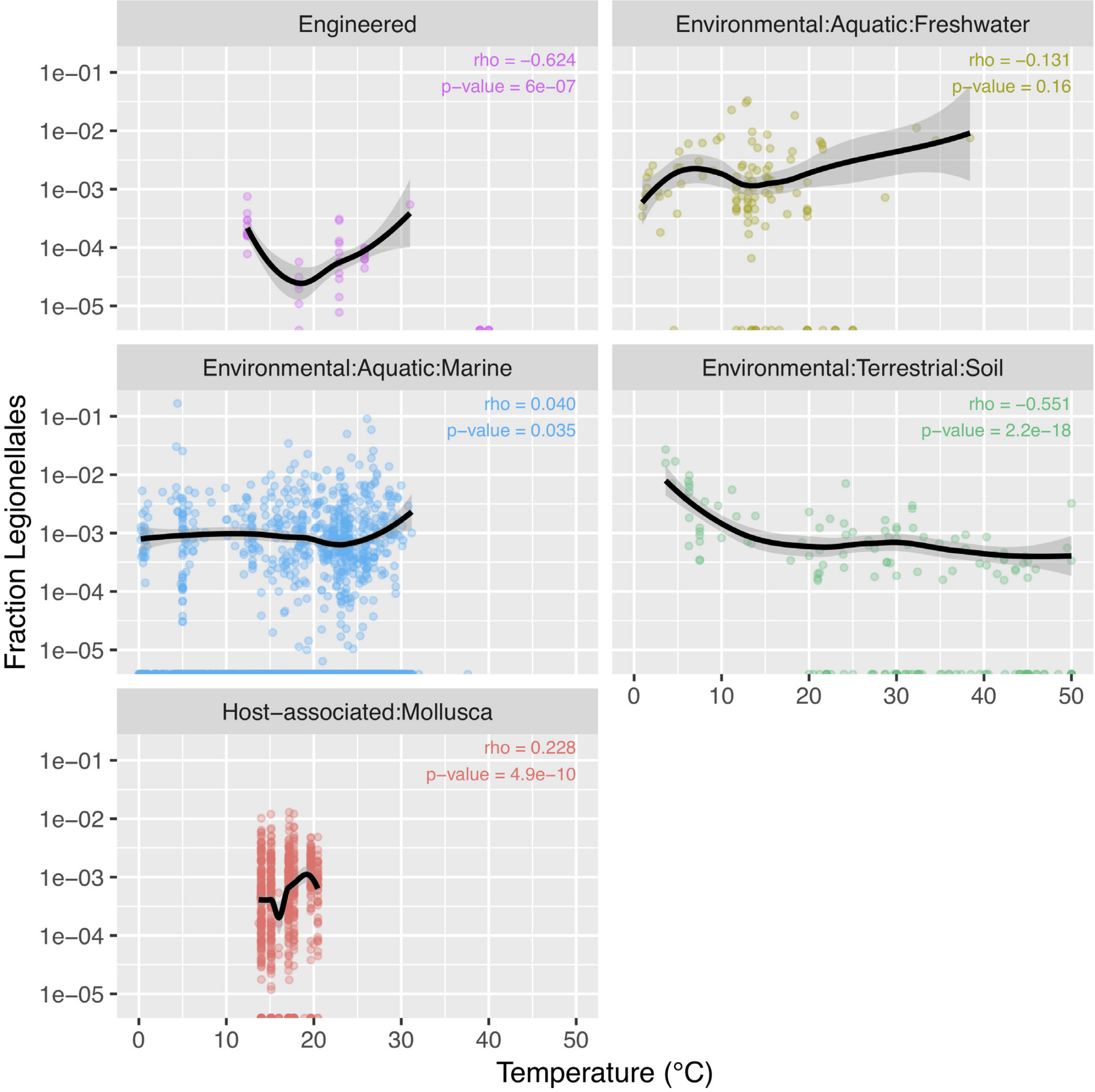

Figure 3. Effect of temperature on the abundance of *Legionellales*. For each biome group for which enough temperature measures were available (n > 10), temperature is represented against the fraction of *Legionellales* reads, in a logarithmic scale (y-axis). Human samples were not considered. Temperature was available for 4074 samples. A local regression curve (LOESS) is displayed on each panel. Samples for which no *Legionellales* reads were found are represented at the bottom of the y-axis but were not used to calculate the regression curve. The rho and P-value of a Spearman's rank correlation test are displayed on each panel: for these, the test was performed on non-logarithmically transformed values, including *Legionellales*-negative samples.

Feldman and Zusman 2005), but also presumably *Rickettsiella* (Leclerque and Kleespies 2008) and *Diplorickettsia* (Mathew *et al.* 2012), to inject proteins into their host and modify its behavior. Given its high level of conservation, the T4BSS has presumably played a key role in the ecological success of the order, enabling *Legionellales* to colonize new hosts. This aspect is relevant to human health: it has been hypothesized that intracellular pathogens of amoebae are likely candidates for emerging bacterial diseases of humans (Lamoth and Greub 2010). Indeed, among *Legionellales*, several clades harbor accidental human pathogens: several species of *Legionella* cause respiratory diseases (Legionnaires' disease and Pontiac fever); *Coxiella burnetii* causes Q-fever; and *Diplorickettsia massiliensis* might also be linked to human infections (Subramanian *et al.* 2012). Researchers have correlated the presence of some of these (potential) pathogens in the natural environment and in man-made water systems where they are most likely to cause diseases, but no large-scale analysis has studied their prevalence and distribution in a global scale, at the order level.

Here, we first show that the genetic diversity of the order is much larger than anticipated from available genomic data. Although there are only six genera for which at least one genome has been sequenced, the order could potentially include over 500 genera. This 'hidden' diversity is not surprising, considering that all *Legionellales* seem to rely on a host for optimal growth. Cultivating them is thus challenging, and they cannot be sequenced through classical genomics, which relies on pure culture. Metagenomics is thus the method of choice to explore the diversity of this, and other, host-adapted clades.

Surveying large quantities of metagenomics data revealed that almost a quarter of all published metagenomics data contain *Legionellales*, with typically a low abundance (about 0.1%) and 1–10 different OTUs, and a very large variation depending on the environment where the samples were taken (Figs 1 and 2). The peak of abundance around 0.1% seems to be specific to *Legionellales*, with other gammaproteobacterial orders having a more uniform distribution. Overall, 22.6% of all samples

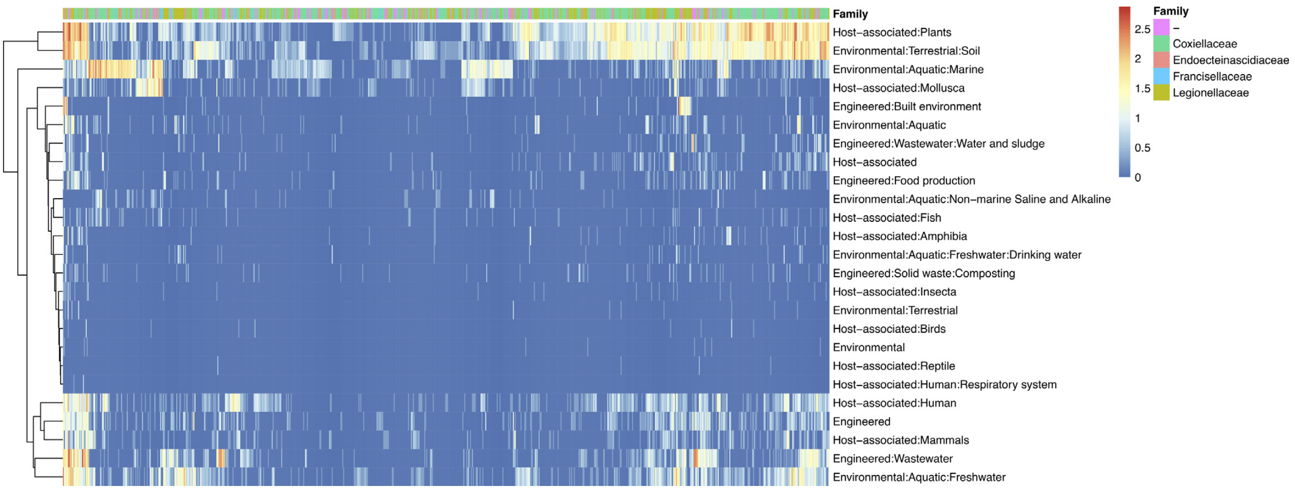

Figure 4. Heatmap based on the prevalence of abundant *Legionellales* OTUs in 25 different biomes. Rows correspond to biomes, and columns to the 804 OTUs appearing at least once in the 5MALOs of any sample. Color scale represents the number of times (in log10) each OTU is found in the 5MALOs most abundant in that biome. The top row gives the family (if available) for each OTU (color legend to the right).

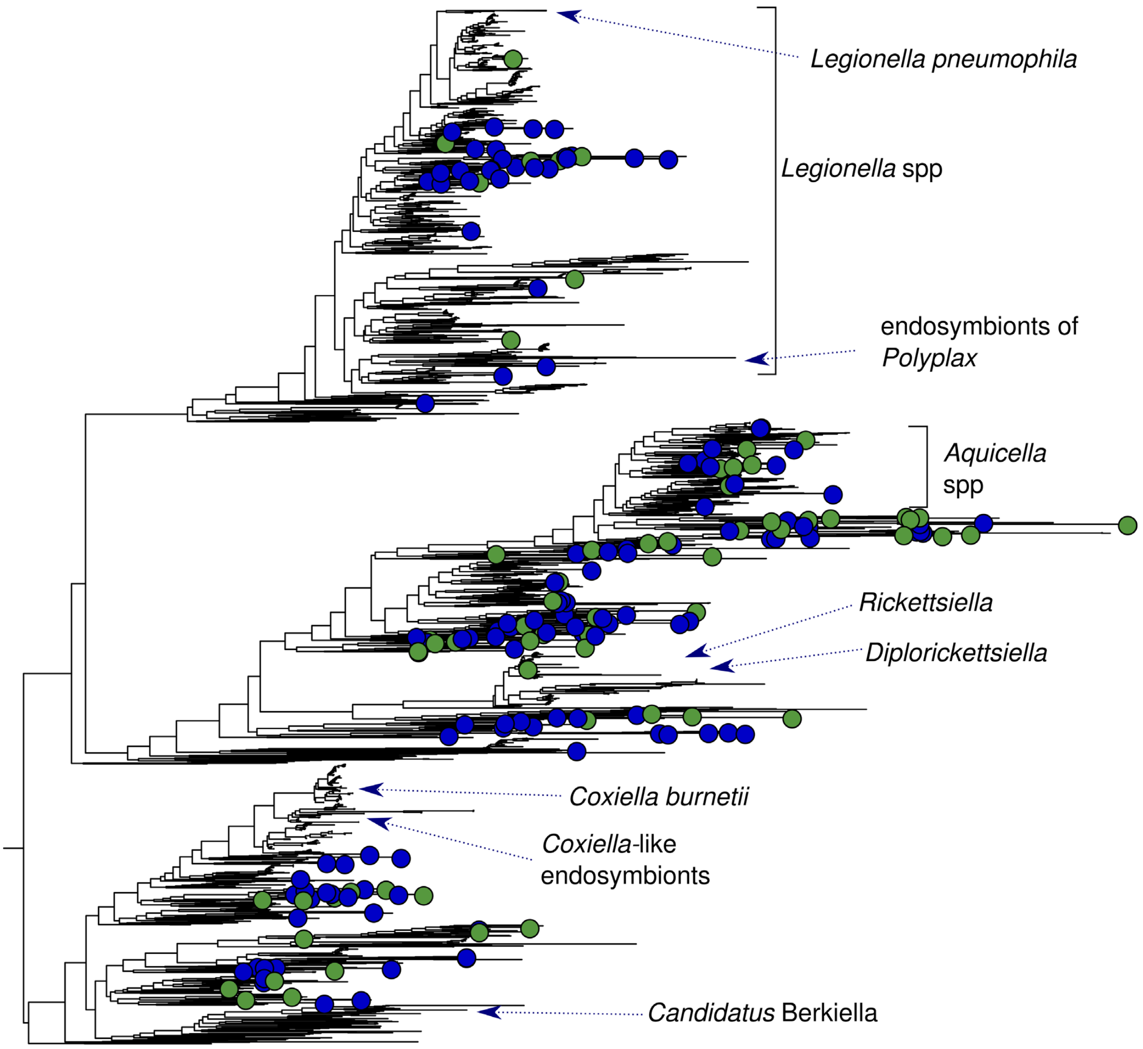

Figure 5. Maximum-likelihood phylogenetic tree of *Legionellales*. The tree is based on all SSU rRNA reads attributed to *Legionellales* in Silva 128 and reads attributed to *Legionellales* in the samples analyzed in this study. The location on the tree of the known genera is indicated by arrows. To improve readability, all branches leading to reads attributed to *L. pneumophila* were collapsed. Blue dots indicate OTUs from the Sala silver mine samples; green dots indicate OTUs from the Uppland samples.

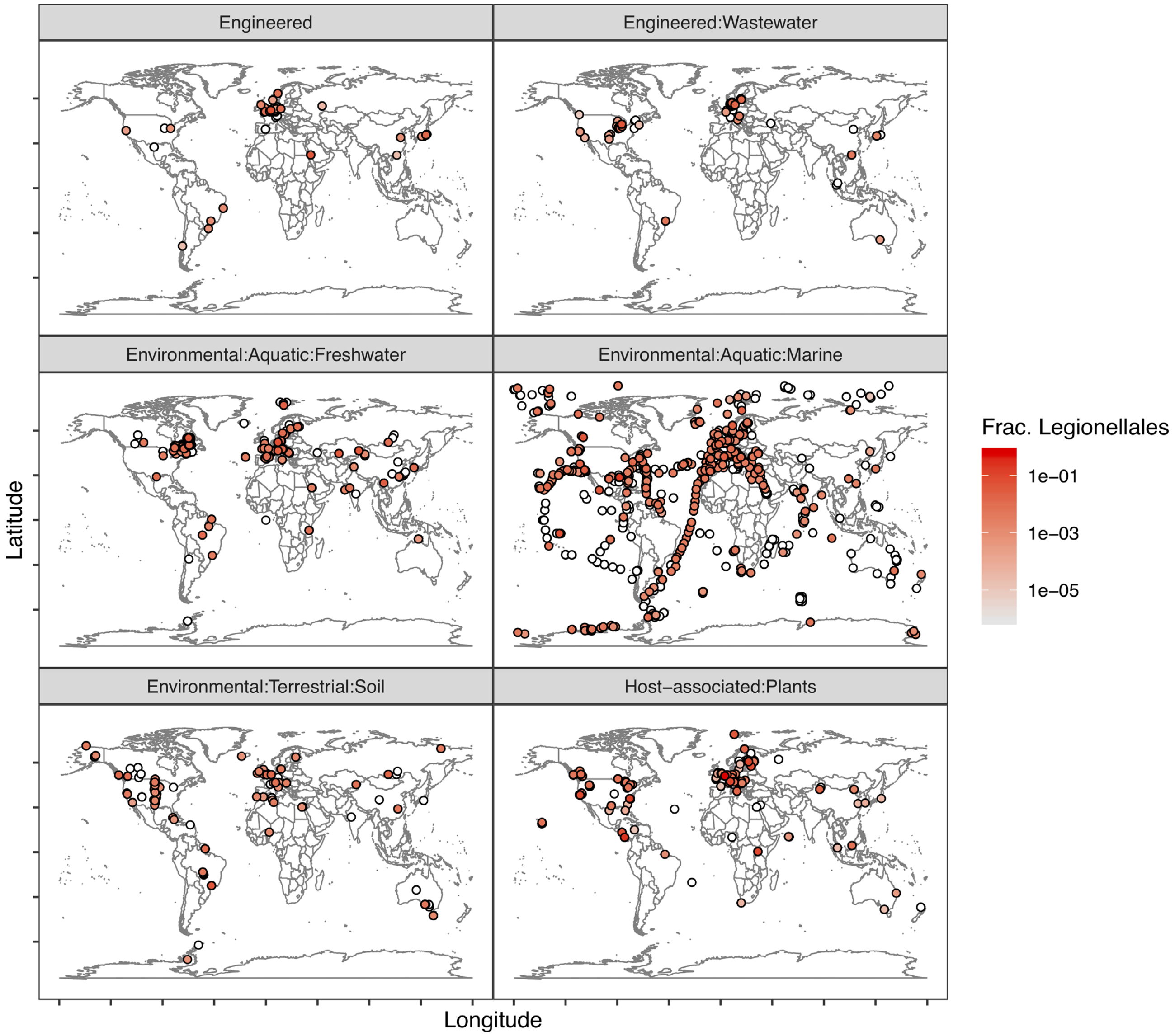

Figure 6. Geographical distribution of *Legionellales* in selected environments. Each panel represents one of the six selected groups of environments or biomes. The sampling location is represented with a dot. *Legionellales*-negative samples are shown in white. LPRs are colored according to the fraction of *Legionellales* reads. A large fraction of the samples in the Engineered category (upper left panel) come from bioreactors and fermenters. It includes all 'Engineered' sub-categories except: Built environment, Food production, Solid waste:Composting, and Wastewater (which is shown separately on the upper right panel). The distribution for all 21 biomes is shown in Supplementary Fig. S6.

contain DNA that can be attributed to *Legionellales* (*Legionellales*-positive runs or LPRs), but this number varies from a few % in hosts (or parts of hosts) that are not commonly colonized by *Legionellales* to over 95% for samples taken from the build environment. In-between, about half of microbiomes associated with soil, plants and freshwater, which are common habitats for *Legionellales*, contain *Legionellales*, with up to several hundred *Legionellales* OTUs, and abundance up to a few %. In line with this, in comparison with other gammaproteobacterial orders, *Legionellales* were particularly diverse and abundant in the built environment, in freshwater and drinking water. Perhaps more surprisingly, marine environments, which are not known to harbor any of the known *Legionellales* species, had similar levels of abundance and richness as soil environments. It would be very interesting to further explore what hosts are colonized in seas and oceans by *Legionellales* bacteria.

Temperature is an important factor for *Legionella* to thrive in man-made water systems (Lesnik, Brettar and Hofle 2016). Its optimal growth temperature is high (37°C) for an environmental bacterium and they survive over 45°C, which makes it prone to proliferate in warm water systems (e.g. Proctor *et al.* 2017). Consistent with that, in freshwater and in mollusk-associated samples, the fraction of *Legionellales* seems to increase for temperatures over 20°C, although the correlations are not statistically significant. However, an inverse tendency is statistically supported in soil samples and in engineered biomes. In the former, the effect is relatively strong for temperatures under 15°C. It should be noted these results are prone to biases: (i) the fraction of samples for which temperature could be retrieved was relatively low (~4.5% of the samples), (ii) the samples available were not controlled for an overrepresentation of a certain type of studies and (iii) the temperature represented in the samples are not uniformly distributed. Nevertheless, although the influence of temperature on the prevalence of legionellosis is disputed (see for example Conza *et al.* 2013; Garcia-Vidal *et al.* 2013; Gleason *et al.* 2016), the global rise in surface temperatures is most likely to change the microbial composition of many biomes, and one of the consequences could be an increase of *Legionellales* in aquatic environments. This might prove problematic in some areas where water is stored at a higher temperature, because the amount of these potential opportunistic pathogens could thus increase in drinking water, thereby potentially increasing the risk of contracting a *Legionellales*-caused disease. More controlled experiments, focusing on the known pathogens in the order, would be necessary to firmly establish the effects of global warming on the abundance of *Legionellales*.

As discussed above, *Legionellales* is a large clade, and we further investigated whether some specific OTUs and subclades had specific distributions across biomes. First, we showed that over 75% of the 1042 OTUs are present at least once in the five most abundant *Legionellales* OTUs (5MALOs), although the distribution of these is very skewed, with only 85 OTUs present in the 5MALOs of over 200 LPRs (Supplementary Fig. S2). It is interesting to note that the OTU present in the most 5MALOs (id 252 003, present in the MALOs of 1563 out of >20 000 LPRs; Supplementary Table 3) is only 88% similar to its closest *Legionellales* relatives. The second-most present OTU (838 066, present in the MALOs of 1456 LPRs) is a yet-unidentified *Legionella* species, 97% similar to other *Legionella* species. A certain degree of biome-specific composition could be observed (Fig. 4), with, for example, biomes from soil and plants being relatively similar, the marine environment clustering on its own, and biomes associated with the chain of drinking water (freshwater, engineered, human- and mammal associated and wastewater biomes) clustering together, which would mean that these groups of biomes share the same OTUs among their 5MALOs. There seems though to be a strong influence of a few single OTUs in biomes that are less represented in our dataset. At genus level, *Legionella* dominate freshwater, terrestrial and engineered biomes, whereas the relatively unknown *Aquicella* species are abundant in soil and plant-associated metagenomes. It is quite interesting to note that among the large amount of *Legionellales*-positive marine samples, only a very small fraction of the most abundant OTUs could be attributed to known species, underlying the importance of more research on intracellular bacteria in marine environments.

From our own sampling campaigns, it appears even more clearly that *Legionellales* OTUs are quasi-ubiquitous, even in environments not known to harbor many host-adapted organisms like mines. In this dark and cold environment, all samples were positive for *Legionellales*, with up to 52 different OTUs in a single sample. In the Uppland samples, which should be the natural environments of *Legionella*, for example, a few samples were negative for *Legionellales* and the diversity was not as large as in the Sala (mine) samples. In contrast, the abundance was in similar ranges: 0–0.7% in the Uppland samples and 0.004–0.6% in the Sala samples (Table 2). It should be noted that the Sala samples yielded much more reads than the Uppland ones. In both samples, the genetic diversity in terms of *Legionellales* was quite high and covered all major clades of the *Legionellales* tree, except the *Coxiella* genus (Fig. 5). Such a wide diversity of host-associated OTUs in the mine environment, which is dark, cold and supposed to have a low biodiversity is surprising, and worth further investigations. No reads from the phylum *Amoebozoa*—which contains all free-living amoebae except *Naegleria*—were retrieved from any of the Uppland or Sala samples, leaving open the question of the potential hosts of the *Legionellales* organisms that live there. Interestingly however, the global diversity and abundance of eukaryotes was noticeably lower in the Uppland sample than in the Sala samples. The lack of *Amoebozoa* reads might be due to the lack of specificity of 'universal' primers for members of that clade (Scheikl *et al.* 2014); alternatively, the highly abundant—but yet unknown—*Legionellales* have hosts other than *Amoebozoa*, or even might be free-living. The latter hypothesis is however unlikely, given that (i) all known *Legionellales* are host-adapted and (ii) there are no known examples of host-adapted bacteria that reverted to a free-living lifestyle (Toft and Andersson 2010). The latest version of the EBI metagenomics pipeline (4.1) now uses the SILVA database, which would allow to also analyze the co-occurrence of *Legionellales* and their hosts at larger scale.

Geographically speaking, *Legionellales* are globally distributed, with very few areas—mostly the South Pacific Ocean—where they were not recovered. The global distribution of *Legionellales* is comparable to that of large gammaproteobacterial order like *Enterobacteriales* and *Pseudomonadales*. Although the fact that *Legionella pneumophila* was ubiquitous in freshwater and built environment was previously known (Sakamoto 2015; van Heijnsbergen *et al.* 2015), the high prevalence of *Legionellales* in marine biomes is surprising. Only few studies have shown the presence of *Legionellales* in marine waters: they have been found in a small per centage of the microbiome of corals (Lawler *et al.* 2016), and in hypersaline environments (Naghoni *et al.* 2017). The fact that *Legionellales* have been identified in cold climates (our study; Fig. 6) is also noteworthy, confirming previous report that *Legionellaceae* were found in freshwater in Antarctica (Carvalho *et al.* 2008) and in the Svalbard island (Ntougias *et al.* 2016).

The global distribution of *Legionellales* is not only observed at order level: the most commonly found OTUs are also, for most of them, globally distributed, both geographically and across biomes (Supplementary Fig. S8). There are exceptions: for example, the most commonly found OTU, (id: 252 003) is mostly found in temperate climates in the Northern hemisphere, and mostly on land.

It is also worth noticing that the vast majority of the most abundant OTUs do not belong to an identified species. For experiments using very short reads or very conserved regions of the rDNA sequence, it might be difficult to correctly identify the species or even the genus, due to the lack of resolution provided by these reads. But each OTU is represented by a full-length rDNA sequence, and it is quite interesting to observe that the most common OTUs have not been isolated and sequenced to this day.

In conclusion, through the analysis of tens of thousands of published metagenomic datasets, we show that the all-host-adapted order *Legionellales* is ubiquitous, both geographically and environment-wise. We also show that the variability in prevalence of these bacteria varies widely, from being rarely observed in most hosts, to being present in half the samples in soil, freshwater and marine environments, and in almost all the samples from man-made environments. In the samples where it was present, its frequency is typically 0.1%, rarely exceeding 1%. The lack of identification of the most common *Legionellales* OTUs emphasizes the need of metagenomics for future studies of host-adapted bacteria. In particular, oceanic waters and cold environments seem to contain many yet-to-be discovered *Legionellales*. We still lack a detailed picture of the order, and future detailed studies on these organisms will bring very valuable knowledge, from both clinical and environmental points of view.

SUPPLEMENTARY DATA

Supplementary data are available at FEMSEC online.

ACKNOWLEDGMENTS

We would like to thank Thijs Ettema, Jennah Dharamshi and Lina Juzokaite for their help with the 16S amplicon protocol. We also would like to thank the support team at EBI metagenomics for their help with retrieving data from their website.

The authors also acknowledge support from Science for Life Laboratory, the Knut and Alice Wallenberg Foundation, the National Genomics Infrastructure funded by the Swedish Research Council and Uppsala Multidisciplinary Center for Advanced Computational Science for assistance with massively parallel sequencing and access to the UPPMAX computational infrastructure.

FUNDING

This project is supported by the Carl Tryggers Foundation [grant number CTS 15:184 to L.G.], Science for Life Laboratory [SciLifeLab National Project 2015 to L.G.]; T.G. is supported by the European Union's Horizon 2020 Research and Innovation programme [grant agreement n° 644669]; work in L.G.'s lab is supported by the Swedish Research Council [grant number 2017–03709 to L.G.].

Conflicts of interest. None declared.

REFERENCES

Andrews S. FastQC. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. 2010.

Boamah DK, Zhou G, Ensminger AW *et al*. From many hosts, one accidental pathogen: The diverse protozoan hosts of legionella. *Front cell infect microbiol* 2017;**7**:477.

Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;**30**:2114–20.

Bouchon D, Cordaux R, Grève P. Rickettsiella, intracellular pathogens of Arthropods. *Manipulative Tenants: Bacteria associated with Arthropods*, In: Zchori-Fein E, Bourtzis K , (eds). pp. 127–45. CRC Press, Boca Raton, FL; 2011.

Carvalho FRS, Nastasi FR, Gamba RC *et al*. Occurrence and diversity of legionellaceae in polar lakes of the antarctic peninsula. *Curr Microbiol* 2008;**57**:294–300.

Christie PJ, Gomez Valero L, Buchrieser C. Biological diversity and evolution of Type IV secretion systems. *Curr Top Microbiol Immunol* 2017;**413**:1–30.

Conza L, Casati S, Limoni C *et al*. Meteorological factors and risk of community-acquired Legionnaires' disease in Switzerland: An epidemiological study. *BMJ Open* 2013;**3**:e002428.

Denet E, Coupat-Goutaland B, Nazaret S *et al*. Diversity of free-living amoebae in soils and their associated human opportunistic bacteria. *Parasitol Res* 2017;**116**:3151–62.

Ewels P, Magnusson M, Lundin S *et al*. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 2016;**32**:3047–8.

Fields BS. The molecular ecology of legionellae. *Trends Microbiol* 1996;**4**:286–90.

Fry NK, Warwick S, Saunders NA *et al*. The use of 16S ribosomal RNA analyses to investigate the phylogeny of the family Legionellaceae. *J Gen Microbiol* 1991;**137**:1215–22.

Garcia-Vidal C, Labori M, Viasus D *et al*. Rainfall is a risk factor for sporadic cases of Legionella pneumophila pneumonia. *PLoS One* 2013;**8**:e61036.

Garrity GM, Brown A, Vickers RM. Tatlockia and fluoribacter: Two new genera of organisms resembling legionella pneumophila. *Int J Syst Bacteriol* 1980;**30**:609–14.

Garrity GM, Bell JA, Lilburn T. Order VI. Legionellales ord. nov. *Bergey's Manual of Systematic Bacteriology*, 2nd edition, Vol. **2** (The Proteobacteria), part B (The Gammaproteobacteria), In: Brenner DJ, Krieg NR, Staley JT, Garrity GM (eds). pp. 210–48. Springer: New York; 2005.

Gleason JA, Kratz NR, Greeley RD *et al*. Under the weather: legionellosis and meteorological factors. *EcoHealth* 2016;**13**:293–302.

Gottlieb Y, Lalzar I, Klasson L. Distinctive genome reduction rates revealed by genomic analyses of two coxiella-like endosymbionts in ticks. *Genome biology and evolution* 2015;**7**:1779–96.

Hosen JD, Febria CM, Crump BC *et al*. Watershed urbanization linked to differences in stream bacterial community composition. *Front Microbiol* 2017;**8**:1452.

Hunter S, Corbett M, Denise H *et al*. EBI metagenomics–a new resource for the analysis and archiving of metagenomic data. *Nucleic Acids Res* 2014;**42**:D600–606.

Katoh K, Toh H. Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform* 2008;**9**:286–98.

Klindworth A, Pruesse E, Schweer T *et al*. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res* 2013;**41**:e1.

Lamoth F, Greub G. Amoebal pathogens as emerging causal agents of pneumonia. *FEMS Microbiol Rev* 2010;**34**:260–80.

Lawler SN, Kellogg CA, France SC *et al*. Coral-associated bacterial diversity is conserved across two deep-sea anthothela species. *Front Microbiol* 2016;**7**:458.

Leclerque A. Whole genome-based assessment of the taxonomic position of the arthropod pathogenic bacterium Rickettsiella grylli. *FEMS Microbiol Lett* 2008;**283**:117–27.

Leclerque A, Kleespies RG. Type IV secretion system components as phylogenetic markers of entomopathogenic bacteria of the genus Rickettsiella. *FEMS Microbiol Lett* 2008;**279**:167–73.

Lesnik R, Brettar I, Hofle MG. *Legionella* species diversity and dynamics from surface reservoir to tap water: from cold adaptation to thermophily. *ISME J* 2016;**10**:1064–80.

Mathew MJ, Subramanian G, Nguyen TT *et al*. Genome sequence of Diplorickettsia massiliensis, an emerging Ixodes ricinus-associated human pathogen. *J Bacteriol* 2012;**194**:3287.

McDonald D, Price MN, Goodrich J *et al*. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J* 2012;**6**:610–8.

McKinney W. Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference*. In: Walt S, Millman J , (eds). SciPy, Austin, Texas, USA 2010, pp. 51–56. http://conference.scipy.org/proceedings/scipy2010/pdfs/proceedings.pdf

Mediannikov O, Sekeyova Z, Birg ML *et al*. A novel obligate intracellular gamma-proteobacterium associated with ixodid ticks, Diplorickettsia massiliensis, Gen. Nov., Sp. Nov. *PLoS One* 2010;**5**:e11478.

Mehari YT, Arivett BA, Farone AL *et al*. Draft genome sequences of two novel Amoeba-Resistant intranuclear bacteria, ‘Candidatus Berkiella cookevillensis’ and ‘Candidatus Berkiella aquae’. *Genome announcements* 2016;**4**:e01732–01715.

Mitchell A, Bucchini F, Cochrane G *et al.* EBI metagenomics in 2016–an expanding and evolving resource for the analysis and archiving of metagenomic data. *Nucleic Acids Res* 2016;**44**:D595–603.

Mitchell AL, Scheremetjew M, Denise H *et al.* EBI Metagenomics in 2017: Enriching the analysis of microbial communities, from sequence reads to assemblies. *Nucleic Acids Res* 2018;**46**:D726–35.

Mukherjee S, Stamatis D, Bertsch J *et al.* Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements. *Nucleic Acids Res* 2017;**45**:D446–56.

Nagai H, Kubori T. Type IVB secretion systems of legionella and other gram-negative bacteria. *Front Microbiol* 2011;**2**:136.

Naghoni A, Emtiazi G, Amoozegar MA *et al.* Microbial diversity in the hypersaline Lake Meyghan, Iran. *Sci Rep* 2017;**7**:11522.

Nguyen LT, Schmidt HA, von Haeseler A *et al.* IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 2015;**32**:268–74.

Ntougias S, Polkowska Z, Nikolaki S *et al.* Bacterial community structures in freshwater polar environments of svalbard. *Microbes Environ* 2016;**31**:401–9.

Omsland A. Axenic growth of Coxiella burnetii. *Adv Exp Med Biol* 2012;**984**:215–29.

Parte AC. LPSN - List of Prokaryotic names with Standing in Nomenclature (bacterio.net), 20 years on. *Int J Syst Evol Microbiol* 2018;**68**:1825–9.

Peabody MA, Caravas JA, Morrison SS *et al.* Characterization of legionella species from watersheds in British Columbia, Canada. *mSphere* 2017;**2**:e00246–00217.

Price MN, Dehal PS, Arkin AP. FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One* 2010;**5**:e9490.

Proctor CR, Dai D, Edwards MA *et al.* Interactive effects of temperature, organic carbon, and pipe material on microbiota composition and Legionella pneumophila in hot water plumbing systems. *Microbiome* 2017;**5**:130.

Qiu J, Luo ZQ. Legionella and Coxiella effectors: strength in diversity and activity. *Nat Rev Microbiol* 2017;**15**:591–605.

R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria; 2017.

Richards AM, Von Dwingelo JE, Price CT *et al.* Cellular microbiology and molecular ecology of Legionella-amoeba interaction. *Virulence* 2013;**4**:307–14.

Rihova J, Novakova E, Husnik F *et al.* Legionella becoming a mutualist: adaptive processes shaping the genome of symbiont in the louse polyplax serrata. *Genome biol evol* 2017;**9**:2946–57.

Rognes T, Flouri T, Nichols B *et al.* VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 2016;**4**:e2584.

Rosenberg R, Lindsey NP, Fischer M *et al.* Vital Signs: trends in reported vectorborne disease cases - united states and territories, 2004–2016. *MMWR Morb Mortal Wkly Rep* 2018;**67**:496–501.

Rossi-Tamisier M, Benamar S, Raoult D *et al.* Cautionary tale of using 16S rRNA gene sequence similarity values in identification of human-associated bacterial species. *Int J Syst Evol Microbiol* 2015;**65**:1929–34.

Sakamoto R. Legionnaire’s disease, weather and climate. *Bull World Health Organ* 2015;**93**:435–6.

Santos P, Pinhal I, Rainey FA *et al.* Gamma-proteobacteria Aquicella lusitana gen. nov., sp. nov., and Aquicella siphonis sp. nov. infect protozoa and require activated charcoal for growth in laboratory media. *Appl Environ Microbiol* 2003;**69**:6533–40.

Scheikl U, Sommer R, Kirschner A *et al.* Free-living amoebae (FLA) co-occurring with legionellae in industrial waters. *Eur J Protistol* 2014;**50**:422–9.

Schloss PD, Westcott SL, Ryabin T *et al.* Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 2009;**75**:7537–41.

Segal G, Feldman M, Zusman T. The Icm/Dot type-IV secretion systems of Legionella pneumophila and Coxiella burnetii. *FEMS Microbiol Rev* 2005;**29**:65–81.

Semenza JC, Suk JE. Vector-borne diseases and climate change: a European perspective. *FEMS Microbiol Lett* 2018;**365**:fnx244.

St. John J. SeqPrep. https://github.com/jstjohn/SeqPrep. 2011.

Stackebrandt E, Goebel BM. A Place for DNA-DNA reassociation and 16S Ribosomal-RNA Sequence-Analysis in the present species definition in bacteriology. *Int J Syst Bacteriol* 1994;**44**:846–9.

Stackebrandt E, Ebers J. Taxonomic parameters revisited: Tarnished gold standards. *Microbiol Today* 2006;**6**:152–5.

Subramanian G, Mediannikov O, Angelakis E *et al.* Diplorickettsia massiliensis as a human pathogen. *Eur J Clin Microbiol Infect Dis* 2012;**31**:365–9.

Taylor M, Mediannikov O, Raoult D *et al.* Endosymbiotic bacteria associated with nematodes, ticks and amoebae. *FEMS Immunol Med Microbiol* 2012;**64**:21–31.

Toft C, Andersson SGE. Evolutionary microbial genomics: insights into bacterial host adaptation. *Nat Rev Genet* 2010;**11**:465–75.

Tsao H-F, Scheikl U, Volland J-M *et al.* ’Candidatus Cochliophilus cryoturris’ (Coxiellaceae), a symbiont of the testate amoeba Cochliopodium minus. *Sci Rep* 2017;**7**:3394.

Tsuchida T, Koga R, Horikawa M *et al.* Symbiotic bacterium modifies aphid body color. *Science* 2010;**330**:1102–4.

van Heijnsbergen E, Schalk JA, Euser SM *et al.* Confirmed and Potential Sources of Legionella Reviewed. *Environ Sci Technol* 2015;**49**:4797–815.

van Schaik EJ, Chen C, Mertens K *et al.* Molecular pathogenesis of the obligate intracellular bacterium *Coxiella burnetii*. *Nat Rev Microbiol* 2013;**11**:561–73.

Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag: New York; 2009.

Williams KP, Gillespie JJ, Sobral BW *et al.* Phylogeny of gammaproteobacteria. *J Bacteriol* 2010;**192**:2305–14.

Wullings BA, van der Kooij D. Occurrence and genetic diversity of uncultured Legionella spp. in drinking water treated at temperatures below 15 degrees C. *Appl Environ Microbiol* 2006;**72**:157–66.

Yilmaz P, Parfrey LW, Yarza P *et al.* The SILVA and ”All-species Living Tree Project (LTP)” taxonomic frameworks. *Nucleic Acids Res* 2014;**42**:D643–648.