Expansion, restructuring and characterization of the *Legionellaceae* family

Marco Gabrielli[1,*], Alessio Cavallaro[1], Frederik Hammes[1]

[1] Department of Environmental Microbiology, Eawag, Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland

* Corresponding author:

Name: Marco Gabrielli

Tel.: +41 58 765 5960

Email: marco.gabrielli@eawag.ch

Abstract

Legionellaceae are a large and widespread family of facultative intracellular bacteria with high clinical relevance. While *Legionella pneumophila*, the most clinically relevant species, is relatively well studied, very limited information is available regarding the other species belonging to this family. Here, we analysed all publicly-available *Legionellaceae* genomes and metagenomic-assembled genomes, nearly doubling the number of recognized species in this family and finding evidence for the need to restructure the family's taxonomy including multiple genera. In addition, we characterize the diversity of secretion systems and traits linked to host invasion encoded by the different species, finding evidence of a widespread horizontal gene transfer. While genus clusters were found to encode different metabolic capabilities, species more related to *L. pneumophila* were generically characterized by a more complete metabolism and a lower amount of auxotrophies, calling for the redesign of cultivation strategies to account for less studied taxa. Finally, a genome-informed analysis of the species environmental distribution revealed that, despite species with similar metabolic capabilities tending to co-occur in the same niches, most species have distinct environmental distributions. Together, our results shed light on the ecology of these microorganisms and highlight the diversity of traits which can occur in closely related facultative intracellular bacteria.

1 Introduction

Microbial ecology has to date largely relied on conventional short-read 16S rRNA sequencing and specific model organisms to explore microorganisms' environmental distribution and phenotypic characteristics. While this combination has enabled important insights into the ecology of several microorganisms and environments[1,2], the limits imposed by amplicon sequencing taxonomic resolution[3] and the restricted number of model organisms available[4] hamper our understanding of the ecological differences which can be present at lower taxonomic levels (i.e., species, strains)[5,6]. This often resulted in a blurred picture of the ecology of several species of interest. For example, assuming that only a fraction of the species within a clade are of biotechnological or clinical interest, assessing the response at the clade level provides only an "average" response which, following Simpson's paradox, might be contrasting to those of the individual species[7]. As a result, the data obtained might not properly inform legislation and management/biotechnological practices.

A relevant example of this problem lies in the family *Legionellaceae* and its only recognised genus, *Legionella,* a clade of gram-negative facultative intracellular bacteria which are responsible for legionellosis. This disease is responsible for a significant number of deaths, productivity losses and medical expenses exceeding 800 million USD per year in the USA alone[8] and which incidence is drastically rising worldwide[9]. While 73 species are currently recognized in this family[10], *L. pneumophila* is alone responsible for more than 90% of infections in Europe and the USA[11]. This contrast between the large number of species and the clinical importance of *L. pneumophila* resulted in different legislative actions, with countries differing both regarding the regulatory target (e.g., only *L. pneumophila*, the entire genus, a combination of the two)[12,13], as well as the target of regulation (e.g., shower water, drinking water, cooling towers)[14]. Additionally, most scientific studies have focused on *L. pneumophila* and its biphasic lifestyle[15] and the knowledge regarding other species lags far behind, especially in case of those with low or absent clinical importance. In fact, most of the information available on non-*L. pneumophila* species has been obtained from comparative genomic approaches which have uncovered a significant heterogeneity in terms of genome organization[16], the range of encoded effectors[17] and secondary metabolites[18]. However, information on the factors which affect their interactions with the environment, such as motility, secretion systems and metabolism, is still lacking. Moreover, the above-mentioned studies have focused only on isolated species, while recent ecological surveys have uncovered a significant diversity of *Legionellaceae* bacteria[19] highlighting the need to include uncultivated species to properly discern the evolutionary paths and the derived ecological differences. Understanding the ecological differences among different *Legionellaceae* species is a question of primary importance given its direct impact on the management and legislation of these species[12].

Here we retrieved over 5000 publicly available *Legionellaceae* genomes and metagenomic-assembled genomes and compared the functional potential of the different species included in this family, focusing specifically on evaluating its genetic diversity and different traits important for facultative intracellular bacteria, including secretion systems, motility and host invasion, and metabolism.

2 Results and discussion

2.1 Analysis of public genomes substantially restructures *Legionellaceae* family

Given the known difficulties in culturing bacteria[20], especially in the case of intracellular ones, culture-independent data provides an indispensable value to characterize microbial diversity. Species delineation of the publicly available *Legionellaceae* isolated genomes and metagenomic-assembled genomes (MAGs) using an average nucleotide identity (ANI) of 95%[21], resulted in the delineation of 128 species clusters (Table S1). This is nearly double the number of the species currently recognised in the family *Legionellaceae* by the International Code of Nomenclature of Prokaryotes[22,23]. Importantly, such delineation was highly concordant with the genomes' metadata: excluding genomes without species information, each ANI-delineated species clusters included genomes which presented the same species label in their metadata. The newly delineated species clusters substantially expand and restructure the *Legionellaceae* family, resulting in several clades with no previously known species. At the same time, several newly identified species clusters, phylogenetically related to known ones, were identified (Figure 1). As an example, ANI similarity supports the fact the *L. pneumophila* subspecies *L. pneumophila subsp. pascullei* and *subsp. fraseri* constitute, in fact, different species (i.e., *L. pascullei* and *L. fraseri*) characterized, by an average ANI similarity to *L. pneumophila* of respectively 91.2% and 92% and present of unique genes linked to biocide resistance and pathogenicity[24]. Similarly, the retrieved genomes indicated the presence of three other species clusters highly related to *L. pneumophila* for which no established scientific name exist (ANI compared to *L. pneumophila*: 91.0% - 91.8%). Given the limited number of genomes belonging to these species clusters, as well as the typing performed in the original publication[25], this result highlights the presence of rare *Legionellaceae* species which are difficult to discern from *L. pneumophila* using common analytical techniques and could induce false positive results (e.g., 16S rRNA sequencing, MALDI-TOF).

Only a limited number (i.e., 1 - 4) of genomovars (i.e., genomic variants, ANI threshold: 99.5%)[21] was identified for most species clusters (Table S2) likely due to the limited number of available genomes (max. 17 genomes per species). On the other hand, 213 and 33 genomovars were observed, respectively, for *L. pneumophila* and *L. fraseri* due to larger number of genomes present in public databases. Nonetheless, several species clusters presented a higher ratio between the number of genomovars and genomes than *L. pneumophila* and *L. fraseri*. While the interplay between the mechanisms of genetic variation (i.e., mutations, gene flow), selective pressure and genetic drift can vary among species and influence the number of intra-species diversity[6], this observation suggests that number of genomovars of non-*L. pneumophila* species is likely to be severely underestimated. Notably, even though the ANI threshold of 99.5% captures with sufficient accuracy the boundary between genomes belonging to the same or different sequence types (Figure S1), as previously proposed[21], the grouping obtained by the two strategies is only mildly similar (adjusted rand index = 0.58). In fact, up to 53 different sequence types belong to a single genomovar, indicating that different evolutionary processes affect the genes used to define sequence types compared to the whole genome[26], which results in confounding source attribution[27].

Despite no consensus definition exists, a genus is generically considered a restricted monophylic clade of organisms with distinct characteristics, which, consequently, present significant genome similarity[28]. The family *Legionellaceae* currently includes only a single genus, i.e. *Legionella*[22,23]. However, a comparison of the average amino acids identity (AAI) across proteins from the retrieved genomes returned values reaching as low as 45%, a value below the suggested boundary for genera occurring between 65 and 72% AAI[29]. Since AAI genus thresholds are not widely used , we further used relative evolutionary distance values[30] in combination with manual core genome phylogeny refinement to define genus clusters and remove polyphyly. This resulted in a total of 17 different

genus clusters present within the *Legionellaceae* family, of which only five include currently isolated representatives. This is substantially different from what is currently accepted, and highlights the need to re-structure the taxonomy of this family. Interestingly, most clinically relevant species cluster[31] belong to the same clade as *L. pneumophila.* However, clinically-associated species are present in three of the five genus clusters with isolated representatives. These three genus clusters with clinically-associated species encompass phylogenetically all clusters with previously known species suggesting that all isolated species could potentially pose a risk to immunocompromised subjects[32]. Interestingly, two of the most clinically relevant species, *L. micdadei*[32] and *L. feeleii*[33], were found to belong to a separate genus cluster than *L. pneumophila*. While such distinction was previously suggested[34], it was later criticized based on 16S rRNA analyses[35], a method now known to provide limited phylogenetic resolution[36].

Previous analyses based on isolated genomes showed how *Legionellaceae* bacteria exhibit large variations in estimated genome sizes[17]. Despite the high estimated completeness, the MAGs in two of the three genus clusters sharing both isolate genomes and MAGs presented significantly lower sizes (t-test, p-vals < 0.006). Even when accounting for this underestimation through a Monte-Carlo simulation, the genome size was found to be significantly different across genus clusters (ANOVA, p-vals < 0.001 on all simulations). In fact, species from the genus cluster including *L. pneumophila* are characterized by larger genomes than most other genus clusters (Figure S2). Noticeably, we observed a strong correlation between genome size and the number of putative horizontal gene transfer (HGT) events (Pearson correlation, r = 0.95, p-val < 0.001; Figure S3), which confirms previous claims regarding the genome expansions in *Legionellaceae* species through HGT[17]. This aspect is highly relevant for the *Legionellaceae* ecology, as genome size is negatively correlated to host-specificity[37], suggesting the wider host-range of the species in certain genus clusters. The combination of genome size variability, together with the proteome diversity across the family, result in a core pangenome (i.e., present in ≥95% of genomes) of only 697 gene families out of 64796 (1.1%). To put this finding in perspective: even though a detailed comparison is not possible due to the differences in pangenome estimation, a similar percentage of core genes was found in the phylum *Chlamydiae*[38], a fact which stresses the high variability of the genome content of the *Legionellaceae* family. In fact, pangenome size varies significantly among taxa in response to lifestyle traits and motile facultative intracellular bacteria, such as *Legionellaceae*, present greater pangenome sizes than non-motile strictly intracellular bacteria such as *Chlamydiae*[39]. The size of the core pangenome and total pangenome size for the entire *Legionellaceae* family are, respectively, -41.7% and +34.7% different than what extrapolated using the Heaps' Law[40] curves obtained from only previously-known species (core families: 1081, total families: 37071), a fact which stresses the diversity of the newly delineated species compared to previously-known ones. In addition, core pangenome genes make up less than half of total number of genes per genome (i.e., genomes from isolated species: 17% - 32%; MAGs: 21% - 46%) with between 2 and 834 gene families present exclusively in a single species and up to 161 additional unique genes present in genomovars. While the combination of genome size and relative phylogenetic distance (i.e., the phylogenetic distance to the closest other *Legionellaceae* genome) provide the best explanation of the number of unique gene families present in each genome (R^2 = 0.78, Figure S4), the genome size provides only a minor contribution compared to the relative phylogenetic distance (ΔAIC values compared to the full model equal to 227.3 and 6.4, respectively) highlighting the importance of the shared evolutionary trajectory of different species on their proteome content.

Finally, despite the lack of metadata from several of the downloaded genomes, *Legionellaceae* bacteria have been retrieved from multiple environments, ranging from human tissues to different water matrices and soils, confirming the widespread occurrence of these bacteria. It is notable that the genomes of many species have only been retrieved from clinical settings (Figure 1). Given the rarity of direct transmission of Legionnaires' disease, the lack of environmental isolates underlines not only the limits and biases of current isolation methods and the difficulty to identify environmental sources of clinical cases[41], but also the lack of knowledge regarding the environmental niches of *Legionellaceae* bacteria. While *Legionellaceae* bacteria have been commonly associated with freshwater and/or engineered environments[42], several species have been retrieved from soil and marine samples, a finding which aligns with recent evidence of the presence of uncultivated species in this environment[19] and the ability of *Legionellaceae* bacteria to grow within marine protists[43]. In fact, the overall environmental niche of facultative intracellular bacteria is not only due the bacteria's direct survival in the environment, but also to the one of their hosts[44], extending the possible environments in which such bacteria can be found. This is particularly relevant for bacteria with broad host ranges, as the case of *Legionellaceae* bacteria[45]. A better understanding of the host range of the different *Legionellaceae* species, as well as a definition of the hosts' environmental niches, is thus needed to understand the environmental prevalence of these bacteria.

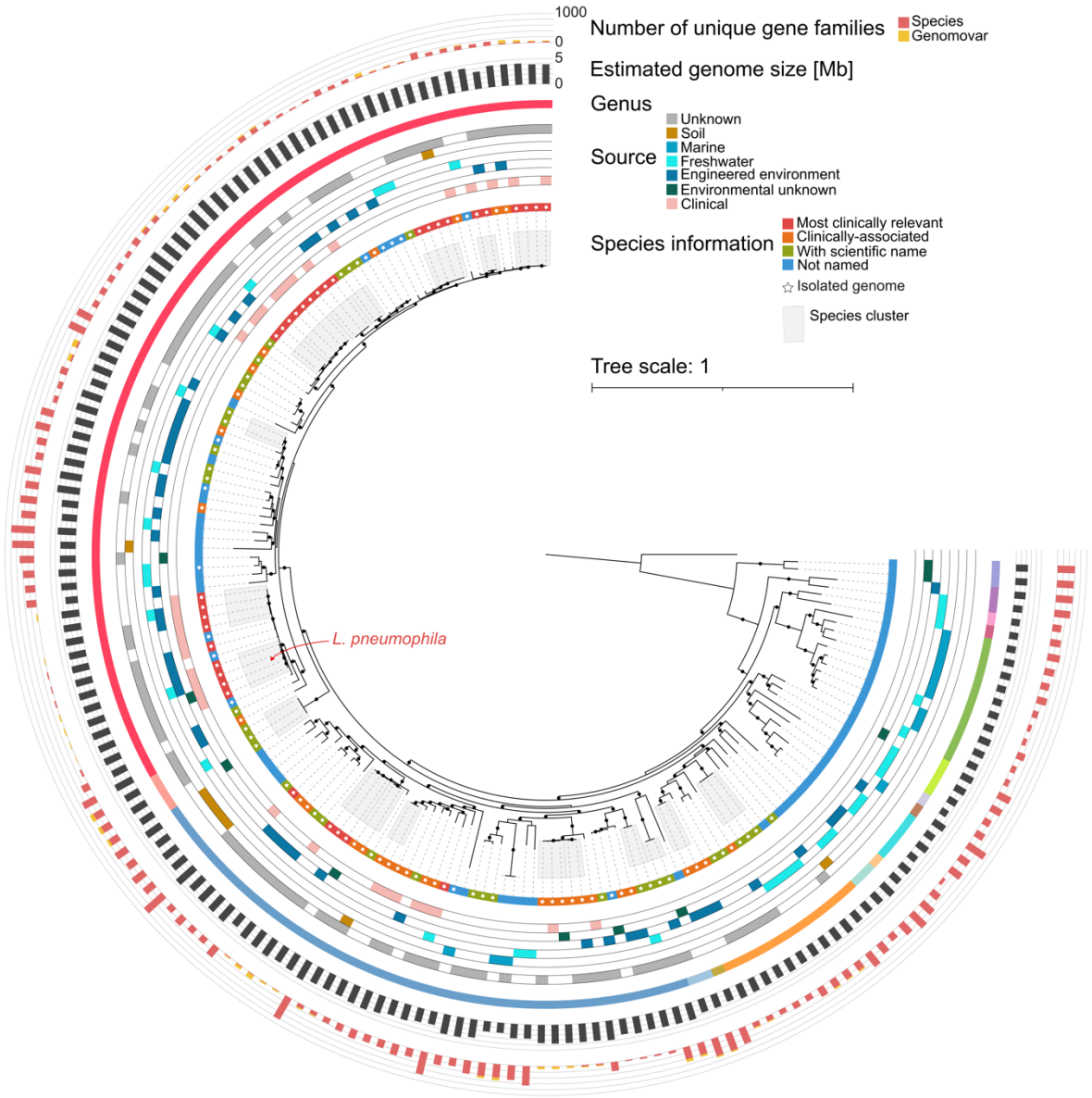


Figure 1. Maximum likelihood phylogenetic tree of the family *Legionellaceae* based on the identified 80% identity core genes families which presented also homologs within the *Coxiella burnetii*, the chosen outgroup (54 protein families, 19141 sites). Nodes with optimized ultrafast support ≥85% are marked with black circles. Leaves enclosed within grey areas indicate genomes of different genomovars belonging to the same species. Outer tree layers indicate whether the delineated species have currently valid scientific names and genomes from isolated specimens are available, the source of all the genomes belonging to the specific species/genomovar, genus cluster membership, the estimated genome size and the number of unique gene families present throughout a species or unique to specific genomovars.

2.2 Secretion systems presence is highly variable within the *Legionellaceae* family

The ability of secretion systems to translocate effector proteins to host cells is a key trait of intracellular bacteria, which allows them to suppress the hosts' defence mechanisms and hijack the

metabolism to their advantage[46]. Once in the host cell, *L. pneumophila* is known to secrete a variety of effectors through the Dot/Icm Type IVB (T4BSS) and Type II (T2SS) secretion systems[17,47,48]. The proteomes of all delineated genomovars included homologs of proteins belonging to the T4BSS of *L. pneumophila,* despite not all proteins always being present. The protein IcmR[49] is encoded in only five highly related species (i.e., *L. fraseri*, *L. pascullei* and three unnamed species) besides *L. pneumophila*, given that its functionality can be fulfilled by several non-homologous proteins[50]. All other proteins were predicted in about 62% of *Legionellaceae* species. No distinct pattern was found in the remaining species with missing genes likely due to genome incompleteness, as suggested by the correlation between the number of T4BSS proteins identified and estimated genome completeness (Spearman correlation, $\rho = 0.46$, p-val < 0.001). Besides the T4BSS secretion system, also the T2SS was detected in nearly all *Legionellaceae* species, confirming previous analyses of its wide distribution[51]. The vast majority (83%) of species harbours homologs of all *L. pneumophila* T2SS components with missing proteins likely attributable to genome incompleteness (Spearman correlation, $\rho = 0.59$, p-val < 0.001). Notably, the presence of a T2SS was not confirmed in the MAGs from two species forming their own genus cluster. While this result could be influenced by genome incompleteness (estimated completeness: 89% and 99%), the observed pattern suggests the occurrence of gene loss events by the clade's ancestor. Given that the T2SS favours survival of *L. pneumophila* at low temperatures[52], it is possible that such species might not be able to survive for long periods of time in such conditions. Interestingly, such two MAGs were, in fact, retrieved from soils incubated at 25 °C[53] and mine drainage[54] which are known to reach high temperatures[55], supporting this hypothesis.

A variety of other systems is present throughout the family, with some species presenting up to five different types (Table S3). However, despite the potential role of these systems in *Legionellaceae* ecology (e.g., host range, motility, HGT), little is known regarding their prevalence and diversity[56]. A Type I secretion system (T1SS), namely lss, is known to be harboured by *L. pneumophila* and to be involved in toxin secretion and host cell entry[57], while novel T1SSs were identified in *L. taurinensis* and later confirmed in other species[58]. T1SSs from *L. pneumophila* and *L. taurinensis* have been found in a total of 42 *Legionellaceae* species, expanding previous estimates of their distribution to several uncultivated clades[58]. The presence throughout the family of small phylogenetically-related groups of species clusters encoding T1SSs suggests that this trait was acquired by multiple ancestors through independent acquisition events. Notably, several genomes present additional novel T1SS which, following a phylogenetic analysis, have been found to be divided in two major clades with further finer clusters present (Figure S5). Both T1SSs clades are present in species from different genus clusters, indicating the circulation of these systems within the family. While the Dot/Icm T4BSS is the main genetic feature of the family *Legionellaceae*, Type IVA secretion systems (T4ASS) are present in 91% of the species. Multiple types of T4ASSs are present four classes (i.e., Types F, G, I and T) with several species encoding for all of them. Notably, T4ASSs presence/absence pattern is only limitedly influenced by phylogenetic relatedness (Mantel correlation between phylogenetic distance and binary Jaccard T4ASS diversity, $r = 0.19$, p-val < 0.001). In fact, while less than 5% of Type G and I T4ASS were found to be plasmid associated, 78% and 43% of, respectively Type F and T systems were found carried by plasmids, supporting previous results indicating their pervasive circulation among different species[17,59]. The *lvh* T4ASS carried by *L. pneumophila* is not essential for intracellular replication[60], but has been shown to support it in specific conditions by interacting with T4BSS effectors[61]. Given that multiple types of T4ASS are present throughout the family, it is possible that this diversity further contributes to the heterogeneity of effectors used to hijack hosts metabolism. Type V secretions

systems (T5SS) mediate several factors related to the survival of intracellular bacteria (e.g., adhesion to host cells) and biofilm formation[62]. T5SS are distributed throughout the family in small phylogenetically-related clusters, suggesting their recent acquisition leading to within-species diversity. For example, within the four *L. pneumophila* genomes analysed, only one of them encodes a T5SS, as reported previously[56]. While the most common type is Type 5a, three species were found to harbour Type 5c systems, a type of T5SS known for their function as adhesins[63]. While cells encoding a T5SS might be favoured in host adhesion and resisting to stresses by biofilm formation, no information is present on T5SS functionality in *Legionellaceae* bacteria[56]. In contrast to other secretion systems, including the pilus-lacking Dot/Icm T4BSS[47], the Type VI secretion systems (T6SS) provides the ability to secrete effector proteins directly into neighbouring cells[64]. A total of 15 species were found to encode such machinery, expanding previous reports of T6SS occurrence[51] and suggesting the occurrence of independent acquisition events. Notably, the tail sheath tssC and the spike evpJ proteins, which are part of standard T6SS architecture, were not detected in any genome analysed. This result suggests potential structural differences of the *Legionellaceae* T6SS or the presence of significant amino acids diversity compared to common model systems. T6SS's ability to inject proteins within nearby cells could potentially be used by *Legionellaceae* bacteria as a defence mechanism against either competing prokaryotes or non-infectable eukaryotes[65] or as a mean to obtain amino acids from neighbouring cells and warrants more detailed investigations.

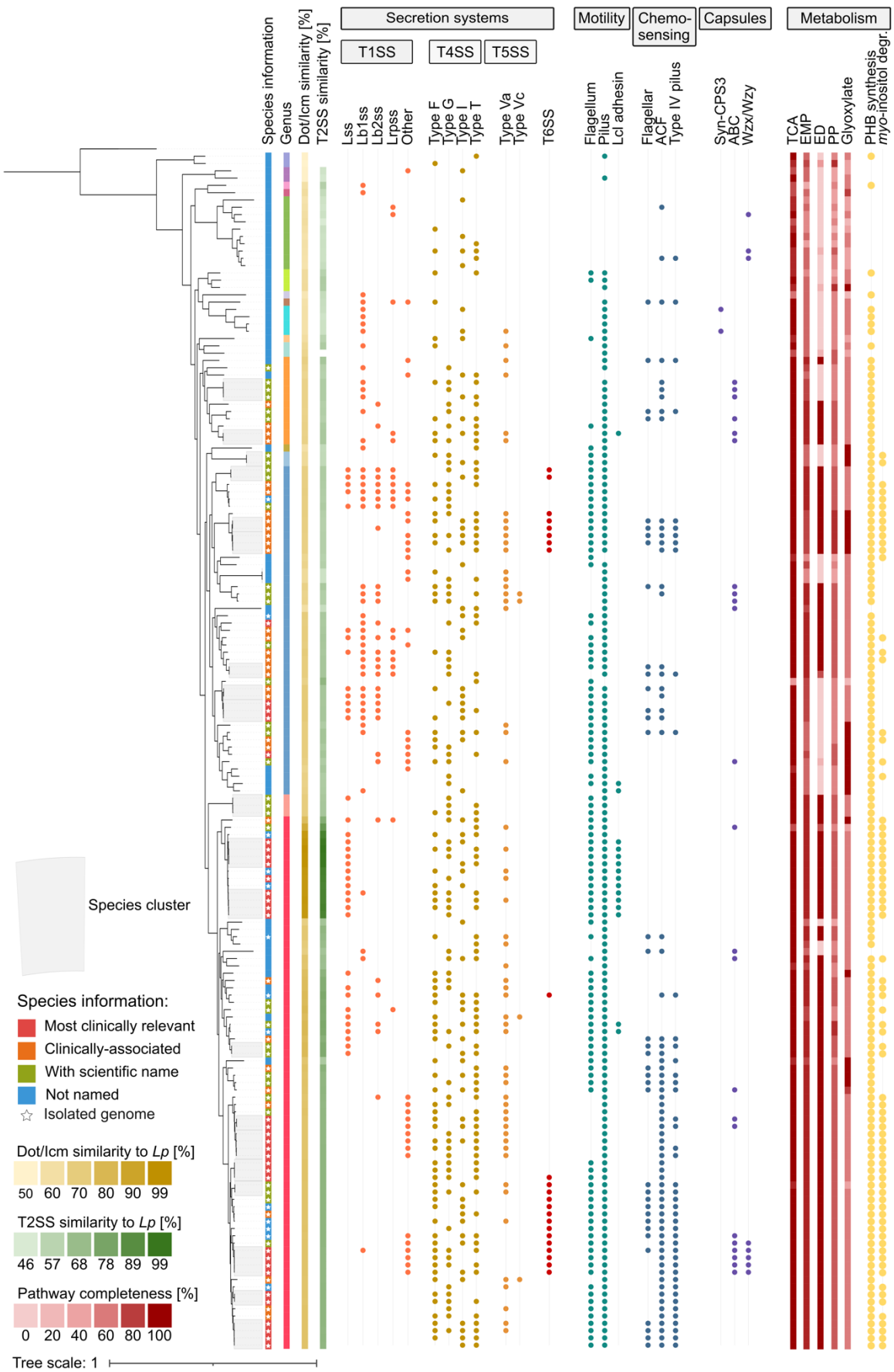

Species information
Genus
Dot/Icm similarity [%]
T2SS similarity [%]
Secretion systems
T1SS
T4SS
T5SS
Lss
Lb1ss
Lb2ss
Lrpss
Other
Type F
Type G
Type I
Type T
Type Va
Type Vc
T6SS
Motility
Flagellum
Pilus
Lcl adhesin
Chemo-sensing
Flagellar
ACF
Type IV pilus
Capsules
Syn-CPS3
ABC
Wzx/Wzy
Metabolism
TCA
EMP
ED
PP
Glyoxylate
PHB synthesis
myo-inositol degr.
Species cluster
Species information:
Most clinically relevant
Clinically-associated
With scientific name
Not named
Isolated genome
Dot/Icm similarity to Lp [%]
50 60 70 80 90 99
T2SS similarity to Lp [%]
46 57 68 78 89 99
Pathway completeness [%]
0 20 40 60 80 100
Tree scale: 1

Figure 2. Maximum likelihood phylogenetic tree of the family *Leginonellaceae* based on the identified 80% identity core genes families which presented also homologs within the *C. burnetii*, the chosen outgroup (54 protein families, 19141 sites). Leaves enclosed within grey areas indicate genomes of different genomovars belonging to the same species. The average Dot/Icm and T2SS protein similarity was estimated against *L. pneumophila* strain *Paris*, while rounded markers indicate the presence of the specified traits.

2.3 Factors linked to host invasion are not conserved among species

Several additional factors play a role in the successful host invasion by *Legionellaceae* bacteria and their persistence in different environments. In fact, not only is the likelihood of an intracellular bacterium encountering its host very low due the diversity of microorganisms in the environment[66], but also amoebae have been shown to actively avoid *L. pneumophila* predation when other preys are available in unstructured environments with low cell densities[67]. While the high densities of biofilms could potentially provide favourable conditions for such encounters, "active" mechanisms of host invasion are likely to increase the chances of an interaction. In fact, most *Legionellaceae* species were found to encode the genes for the expression of Type IV pili, responsible for twitching motility, or flagella, responsible for swimming and swarming (Figure 2, Table S3). The genes required for pili assembly are present in all cultivated species apart from *L. tunisiensis*, but absent in several novel species. While part of this pattern could be due to genome incompleteness, it is notable that no pili were detected in 10 species that constitute their own genus cluster, a factor which suggests the loss of such function by a common ancestor and suggests their limited efficiency in host colonization[68]. On the contrary, diverse flagella assembly genes were not detected among several species, confirming previous experimental observations (i.e., *L. longbeacheae*)[69] and suggesting the loss of this trait multiple times within the family. The lack of flagella in *L. longbeacheae*, a species often linked to soil and compost[31], is likely due to the fact that surface motility does not necessarily require such machinery[66]. Since also closely related species lack this functionality (i.e., *L. cincinnatiensis*, *L. gratiana*, *L. santheilensis*, *L. santicrucis*), it is possible that those species are also adapted to not aquatic environments (i.e., soil). Finally, a form of sliding motility was reported for several *Legionellaceae* species due to a surfactant linked to T2SS[70]. However, while almost all species encode a T2SS, not all were reported to perform surface translocation, such as the case of *L. midcadei*. In fact, effectors non related to intracellular replication tend to have lower prevalence across the genus[71]. Beyond motility mechanisms, adhesins are deemed key determinants of host-pathogen interactions and biofilm formation[72], such as the Lcl protein encoded by *L. pneumophila*[73]. Lcl homologs characterized by a collagen-like domain were detected not only in *L. pneumophila* and highly related species, but also in *L. quateirensis*, *L. oakridgensis* and two other novel species, indicating that this protein is not uniquely present in *L. pneumophila,* as previously suggested[73]. In addition, other 10 genomes were found to encode proteins with collagen domains characterized by "unknown" or "membrane" COG categories without further characterization, possibly highlighting additional adhesins absent in *L. pneumophila*. Besides an easier attachment to host cells and surfaces, the enhanced biofilm formation abilities provided by this protein might also favour resistance to stresses and favour the persistence of these species in the environment. Characterizing the importance of such traits is complicated by hosts diversity. For example, *L. pneumophila* was shown to require a flagellum to efficiently infect human macrophages, but amoebal infection was not drastically affected by this trait[74]. Given the importance

of hosts in *Legionellaceae* ecology, developing a framework to understand the factors characterizing host interactions is key and could potentially create predictive insights which could be used for the management of these bacteria.

Beyond random motility, chemosensing, the ability to sense chemical gradients, can be used to direct movement (i.e., chemotaxis) and improve the contact between motile intracellular bacteria and hosts by more than 10-fold[66]. However, the majority of *Legionellaceae* species lacked both CheW and methyl-accepting chemosensing proteins, two of the essential components of the chemotaxis Che system[75], indicating the absence of motility regulation through the Che pathway (Figure 2, Table S3). The presence of a Che system does not seem to be a trait linked to flagella and pili, given that the CheA proteins predicted in most genomes were linked to cellular functions not related to motility[75]. Still, even when including only CheA proteins predicted to be related to pili and flagellar chemotaxis, only a weak statistical association was only found for flagella (binomial generalized linear model, p-value = 0.02) and not pili (p-value = 0.07), further stressing the unrelatedness of these traits. Interestingly, previous research has highlighted how only around 50% of animal and human pathogens are known to encode Che systems[72] and that the majority of respiratory pathogens are non-chemotactic, a pattern which was attributed to the relative simplicity of the respiratory system compared to the digestive tract[66]. However, while this hypothesis might be considered valid for pathogens with a direct transmission route (i.e., human to human), it does not take into consideration the need to survive in the environment and find suitable hosts of facultative intracellular bacteria. As a matter of fact, while the Che system is the most common and studied system, other regulatory pathways to modulate chemotaxis exist[76]. For example, twitching motility in *L. pneumophila* was reported to be inhibited by the presence of other bacteria likely as a result to lower nutrients concentrations[77], indicating either that rich nutrient conditions are needed for twitching motility or the ability of this bacterium to sense chemical gradients despite not possessing chemotaxis genes. In fact, up to 80 signal transduction and 170 response regulator proteins were predicted per genome. Interestingly, while the genome origin (i.e., isolate/MAG) was found to significantly affect the number of genes detected (t-test, p-vals < 0.05 in 2/3 and 3/3 genus clusters with both genomes origin), a significant difference in the number of these proteins was also detected across the isolated bacteria in different genus clusters (ANOVA, Tukey HSD post-hoc test, p-values < 0.002; Figure S6) with the one including *L. pneumophila* presenting the highest number. In total, 726 and 1363 gene families were identified for signal transduction and response regulation proteins, of which only 2 and 13 were detected across the whole family (≥95% species), respectively. The identified signal transduction proteins are a phosphotransferase (ptsP) and a universal stress protein (UspA family), while the transcriptional regulator identified included several sigma factors (RpoD, RpoH, RpoS, FliA), the RNAase R protein, Fis-like and Omp-like DNA-binding proteins, the proteins CsrA and LetA, two iron-related regulator proteins and a NADPH-dependent 7-cyano-7-deazaguanine reductase. Most identified genes have already been described to affect several aspects of *L. pneumophila* lifestyle. In fact, UspA proteins have been involved in the survival in nutrient poor conditions such as drinking water[78,79], while ptsP favours adaptations to intracellular environments affecting the transcription levels of several effectors[80]. Several regulators identified have been linked to several aspects such as biofilm formation, metabolism regulation and responses to environmental conditions[15,81–84] and the influence of iron on lifestyle[85]. The conservation of such proteins suggests that similar regulatory networks might be conserved across the family, but around 50% of all signal transduction and response regulators protein families have been identified in only one species, while others show a

clear clade membership (e.g., ptsN)[80]. These results suggests that while a "core" regulatory network might be conserved across species, significant diversity is present within the family, which might significantly affect phenotypic responses to environmental cues and contribute defining the environmental niche of different species.

Extracellular capsules are another factor which can affect bacteria-host interactions and enhance survival in unfavourable environments by, for example, protecting cells from phagocytosis and antimicrobial compounds, and preventing desiccation[86]. Among *Legionellaceae* species, only *L. longebeacheae* has been reported to include a unique capsule cluster, which has been recognized as a crucial virulence factor towards *Acanthamoeba castellani*[87]. In fact, capsules systems are more common in bacteria found in soil than in water environments[86]. Despite not presenting notable homology to *L. longbeacheae* capsule proteins[87], biosynthetic gene clusters of different types were detected across the family. Even though only one capsule system was detected in most cases, *L. beliardensis*, *L. gresiliensis* and *L. busaniensis p*resented two ABC capsules systems per genome, while *L. gormanii* and *L. qingyii* encoded two different capsule types. Given that capsules do not necessarily protect against all stresses[87], each capsule might provide resistance to different stressors, extending the niche breadth of these species[88]. Within species, the similarity of the capsule systems varied widely across genomovars (Figure S7). For example, while *L. gormanii* and *L. longbeacheae* genomovars presented mostly conserved gene clusters, this was not the same for *L. oakridgensis* or *L. septentrionalis* where, besides a restricted conserved fraction, the majority of the protein differed substantially among genomovars. The high genetic turnover of capsules has been suggested to stem from the adaptation to environmental conditions or lifestyle changes[86], suggesting the high diversity of environments inhabited by different genomovars of *L. oakridgensis* and *L. septentrionalis.*

2.4 Genus clusters encode different metabolic potentials

Several laboratory studies have highlighted the presence of a complex biphasic metabolism in *L. pneumophila,* in which the main metabolic pathways used depend on the growth phase[15]. However, the metabolism of other species has been seldomly examined, leading to a major knowledge gap regarding the ecology of these species. *L. pneumophila* has been shown to rely on the tricarboxylic acid (TCA) cycle in *in vitro* experiments, converting serine into acetyl-CoA[89]. The TCA cycle showed completeness values above 75% in nearly all *Legionellaceae* species with the exception of *L. tunisiensis,* which presented a completeness of only 50% (Figure 2, Table S4), likely due to the genome's limited completeness (i.e., 85.8%) and high fragmentation (i.e., 340 contigs). Besides serine biotransformation, other pathways, such as the Embden-Meyerhof-Parnas (EMP) and the Entner-Doudoroff (ED) pathways can provide pyruvate required by the TCA cycle[90]. Tracer studies have shown that only minimal amounts of glucose are metabolized through the EMP pathway by *L. pneumophila,* as the majority was shown to undergo transformations through the ED pathway[89]. While the ED pathway functionality did not affect *L. pneumophila* growth *in vitro*, its defect led to a reduction in intracellular growth, strongly indicating its requirement for the establishment within the host and the reliance on glucose during the transmissive phase[15,91]. Interestingly, the ED pathway was either detected in its entirety, or otherwise completely absent. Notably, while no genomes with an absent ED pathway showed high EMP pathway completeness percentages (i.e., >85%), high EMP completeness values were observed in most genomes with a complete ED pathway, a fact that suggests the reciprocal dependency of the two pathways. In addition, the pentose phosphate (PP)

pathway has been also shown to have an active role in glucose and glycerol metabolism[92]. The known lack of the oxidative branch in *L. pneumophila*'s PP pathway and its compensation by the ED pathway[92] is concordant with the observed completeness of 85%. A higher PP pathway completeness was observed in genomes presenting a complete ED pathway (72% vs 56%; t-test, p-val < 0.001) corroborating this observation throughout the whole family. PP pathway completeness correlated positively with the AAI similarity with *L. pneumophila* (Sperman's ρ = 0.45, p-val < 0.001), suggesting that this pathway might have a more prevalent role in species phylogenetically related from this model organism. Coincidentally, the incompleteness of common glycolysis pathways highlights the possibility that the species belonging to the newly defined genus clusters have limited ability to convert carbohydrates. On the contrary, while the glyoxylate cycle is highly incomplete in *L. pneumophila*[89], its complete set of genes was observed in several species which, apart from *L. waltersii* and a restricted clade including *L. drancourtii* and *L. lytica*, mostly belong to species not in the *L. pneumophila* genus cluster, where they potentially contribute to the biosynthesis of cell carbon compounds. Studies on *Escherichia coli* highlighted the importance of this pathway for low growth rate regimes and in the case of nutrients switches[93], however future studies are needed to verify whether this applies also for *Legionellaceae* bacteria.

Besides such central metabolic pathways, *L. pneumophila* is known to be able to biosynthesize poly-3-hydroxybutyrate (PHB) and metabolize *myo*-inositol[94,95]. PHB acts as a storage compound to provide both energy and carbon in stationary phase[94]. A PHB polymerase is needed to transform (R)-OH-Butanoyl-CoA into PHB, but this enzyme is present almost uniquely in a specific clade composed of multiple genus clusters (Figure 2). Although the lack of PHB polymerases has virtually no effect on intracellular growth of *L. pneumophila*, PHB accumulation promotes long-term survival under starvation[96] and its absence from several *Legionellaceae* species indicates a substantial difference in the adaptation and/or survival to the potentially nutrient-poor extracellular conditions. Conversely, the ability to metabolize extracellular myo-inositol, a widely distributed class of compounds[97], by *L. pneumophila* was shown to favour its intracellular replication[95]. Homologs of the transporter and dehydrogenase involved in this pathway were found in most species belonging to the same genus cluster as *L. pneumophila* and few other species related to *L. rubrilucens* and *L. quinlivanii*, but were otherwise absent from the species in most genus clusters, suggesting the broader potential niche of selected *Legionellaceae* species.

The diversity of *Legionellaceae* metabolism is further confirmed by the fact that, even excluding MAGs to remove the bias derived from genome incompleteness[98], the completeness of the metabolic pathways shows significant differences among the genomes belonging to different genus clusters (Figure 3; PERMANOVA R^2 = 24.6%, p-val < 0.001). This metabolic clustering arises from the presence of specific enzymes in each genus cluster, as shown in Figure S8. Most notably, the genus clusters including *L. pneumophila* and *L. micdadei* show a larger number of enzymes, which are predominantly found only in such clades, stressing their richer metabolic capabilities. Nonetheless, such capabilities differ between the two clades as indicated by the differential presence of 59 enzymes (Table S5) which leads to more complete pathways for phosphatidylcholine biosynthesis and acylglycerol degradation in the genus cluster of *L. pneumophila*, while higher completeness of the glycogen biosynthesis pathway in the one of *L. micdadei*. Still, within each genus clusters, further smaller clusters of genomes are present (Figure S9, Table S6) indicating the presence of phylogenetically related clusters with similar metabolic potential (Spearman correlation between metabolic pathways completeness and phylogenetic distance, ρ = 0.5, p-val < 0.001). Several of such clusters included both species clinically-

associated or not, remarking the importance of effectors of the specific species in their relationship with hosts[99]. This also suggests that, given the similarity of their metabolism, species with different associations to disease can have overlapping environmental niches.

The greater metabolic capabilities of the species belonging to selected *Legionellaceae* genus clusters extends not only to traditional metabolic pathways, but also to carbohydrate-active enzymes (CAZymes) and peptidases. Even excluding the bias connected to MAGs incompleteness, statistically significant differences were found in the number of genes encoding for both secreted and intracellular CAZymes and peptidases (Figure 3b). To extend the analysis of the diversity of the classes of such enzymes to MAGs, we rarefied the number of predicted enzymes to a common value before estimating their diversity. This procedure reduced the effect of genome source (PERMANOVA, R^2 = 1.1 - 4.9%, p-val < 0.04) and allowed to discern the much greater effect of genus cluster membership for both intracellular and secreted CAZymes and peptidases (PERMANOVA, R^2 = 9.2 - 17.6%, p-val < 0.001), a fact also reflected in the number of predicted CAZYmes families and peptidases species per genome (Figure S10). Such results indicate that different genus clusters are not only characterized by different levels of functional diversity, but also encode the enzymes to rely on different complex organic matter substrates[100,101].

Overall, the more restricted metabolic potential of several of the novel delineated species clusters, even though potentially affected by MAGs completeness, suggests their higher dependence on amino acids than other species. While it is possible that such amino acids could be retrieved freely in the environment due to the leakage from other microorganisms[102], it is also a possible signal of a stronger host-association. Such hypothesis could explain the difficulties in their isolation and warrants further experimental investigations.

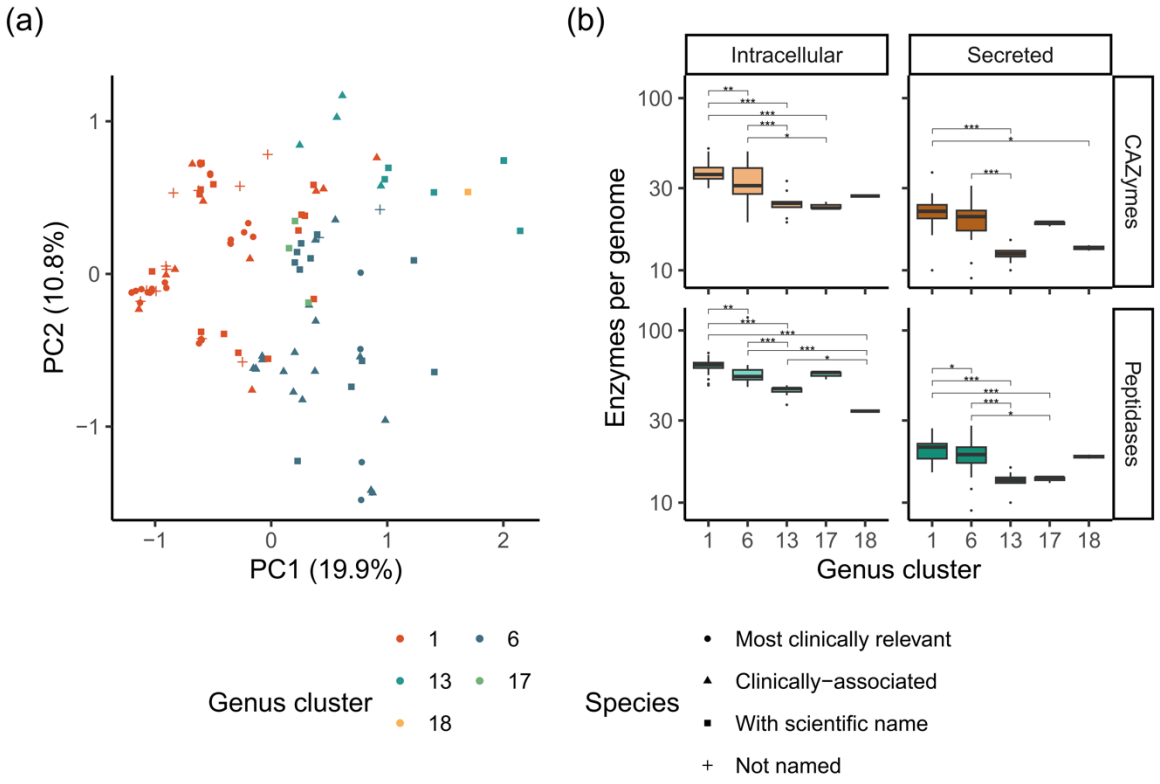


Figure 3. (a) Metabolic differences among *Legionellaceae* isolated genomes and clustering following genus cluster membership. (b) Number of CAZYmes and peptidades per isolated genomes as a function of their secretion and genus cluster membership. Statistically significant differences among genus clusters are marked by their p-values: ”*” = 0.05 - 0.01, ”**” = 0.01 - 0.001, ”***” < 0.001.

2.5 Auxotrophy predictions suggest redesign of cultivation media

Intracellular bacteria have evolved auxotrophy for diverse amino acids[103], gaining a fitness advantage in their host-associated niche, but becoming dependent on external resources for survival[104]. This also holds true for *L. pneumophila*, for which several studies have confirmed its auxotrophy for several amino acids (Table 1). While certain auxotrophies have been detected in all studies, some discrepancies can be found, possibly due to differences in environmental conditions. *In silico* analysis of *L. pneumophila* auxotrophies was performed using GapMind, a pathway annotation tool[105], and gapseq, an automated tool for genome-scale metabolic model reconstruction[106]. Except for L-isoleucine and L-serine, the predictions of the two tools were mostly concordant and matched, respectively, eight and six of the 11 reported auxotrophies. To test whether such differences might be due to strain-level differences, the two tools were tested on the different *L. pneumophila* strains investigated by George and coworkers[107]. While a different requirement for L-proline and L-serine was detected experimentally, both computational approaches did not result in different auxotrophy profiles among the various strains. However, a lower flux to biomass (i.e., reduced growth) was reported by gapseq in the absence of L-serine, matching the experimental result obtained for one of the tested strains. It is worth to note that, besides being affected by the various knowledge gaps around microbial metabolism, genome annotations and metabolic models represent the maximum potential capabilities of microorganisms. Such capabilities might not match what is actually expressed

in a given condition, resulting inherently in a lower amount of predicted auxotrophies compared to what observed. The different results obtained for L-isoleucine and L-serine by two approaches lies in the fact that these compounds are predicted by gapseq to be synthesized following two pathways not considered by GapMind: L-isoleucine biosynthesis pathway III and part of the folate cycle (Table S7). The consistency of the reports regarding L-isoleucine auxotrophy indicates that the pathway identified by gapseq is likely not biologically relevant. Conversely, the variability of the reports regarding L-serine suggests that the identified pathway might be feasible in certain conditions and, thus, warrants more investigation. Interestingly, while L-cysteine is commonly used for culture confirmation of *L. pneumophila*, both tools did not predict its auxotrophy. However, GapMind predictions indicated that even though the genes required for the biosynthesis of L-cysteine through the reverse transsulfuration pathway[108] are likely present, this pathway required homocysteine, an intermediate of methionine biosynthesis, which cannot be synthetized by *L. pneumophila*. In addition, in agreement with the close link between L-cysteine and sulphur[109], gapseq highlighted the requirement of reduced and oxidized sulphur sources (e.g., SO_4^{2-}, H_2S) to perform L-cysteine biosynthesis. Such observations warrant dedicated experimental confirmations to clarify *L. pneumophila* metabolic capabilities.

Table 1. Reported and predicted auxotrophies for *L. pneumophila*. Experimental evidence retrieved from [89,107,110]. High and medium GapMind biosynthesis prediction confidence results indicate the detection of homologs for all biosynthetic steps for a given amino acid. Conversely, low confidence results highlight the lack of biosynthetic steps and imply auxotrophy[105]. Multiple results per amino acid indicate that distinct results were obtained with different genomovars. The asterisk indicates that while the genes for the L-cysteine biosynthesis have been annotated, the genes required to produce the required compound homocysteine are absent.

Amino acid	Experimental evidence	GapMind biosynthesis prediction confidence	gapseq prediction
L-alanine	-	Not assessed	-
L-arginine	Auxotrophy consistently reported	Low	Auxotrophy
L-asparagine	-	High	-
L-aspartate	-	Not assessed	-
L-cysteine	Auxotrophy reported	Medium*	-
L-glutamate	-	Not assessed	-
L-glutamine	-	High	-
Glycine	-	High	-
L-histidine	-	Medium/High	-
L-isoleucine	Auxotrophy consistently reported	Low	-
L-leucine	Auxotrophy consistently reported	Low	Auxotrophy
L-lysine	-	High	-
L-methionine	Auxotrophy consistently reported	Low	Auxotrophy
L-phenylalanine	Auxotrophy reported	Low	Auxotrophy
L-proline	Auxotrophy reported	High	-
L-serine	Auxotrophy reported	Low	-
L-threonine	Auxotrophy reported	Low	Auxotrophy

L-tryptophan	-	Medium	-
L-tyrosine	Auxotrophy reported	High	-
L-valine	Auxotrophy consistently reported	Low	Auxotrophy

A similar concordance (again, with the exception of L-serine and L-isoleucine) was also observed for the auxotrophy predictions of the other species for which an isolated genome was available. Given the susceptibility of annotation-based tools to genome incompleteness and the presence of several genus clusters without isolated representatives, it was decided to continue the analysis only using gapseq's results. The gapfilling carried out by this tool, in fact, removed the differences in the number of predicted auxotrophies between isolates and MAGs from the same genus cluster (t-test, p-vals > 0.1), while highlighting significant differences among genus clusters (ANOVA, p-val < 0.001). In fact, different genus clusters presented distinct auxotrophy profiles with no amino acid auxotrophy being conserved across the entire family, despite several being highly prevalent in most genus clusters (Figure 4, Table S8). No genomovar showed auxotrophy towards five amino acids (L-alanine, L-aspartate, L-glutamate, L-glutamine and glycine). Interestingly, while L-cysteine auxotrophy was not detected in *L. pneumophila*, this was still detected in a few genomes belonging to the same genus cluster of *L. micdadei* and in several others without isolated representatives. With the exception of the genus cluster composed exclusively of *L. geestiana*, the genus clusters with isolated representatives showed low prevalence of L-histidine auxotrophy. Given that not all *Legionellaceae* species grow equally well on standard medial[111], this result suggests the need for media supplementation of additional amino acids such as L-histidine for the isolation of novel species.

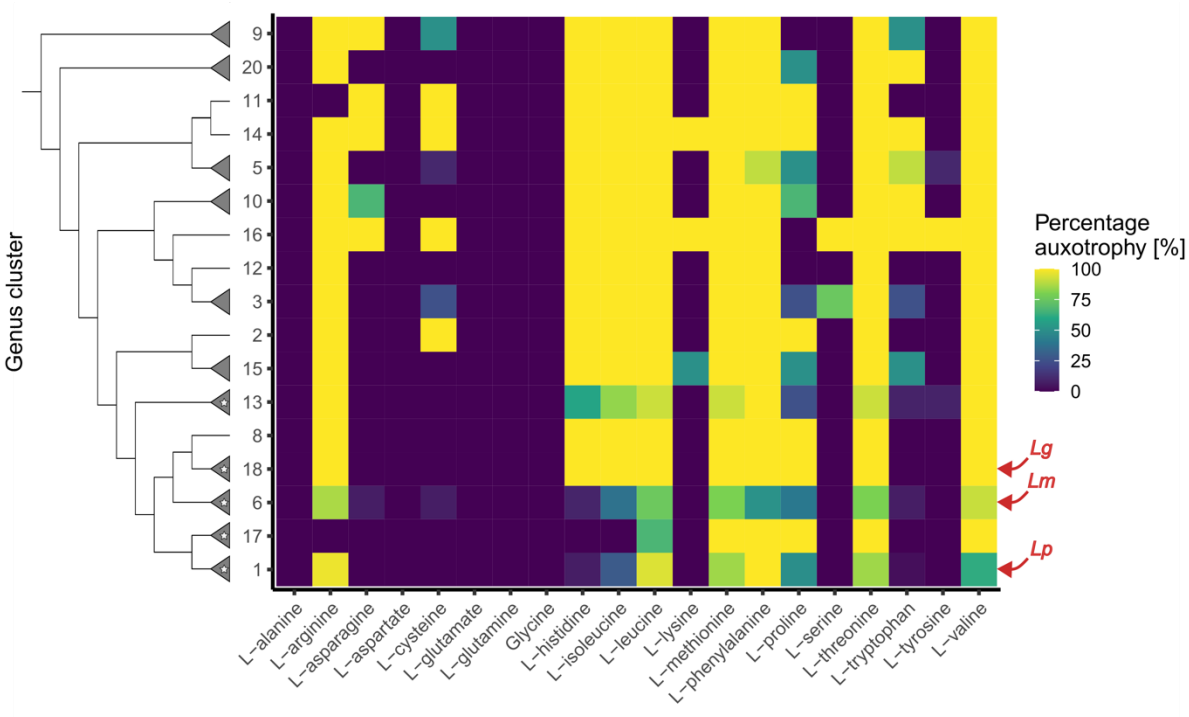


Figure 4. Prevalence of auxotrophies predicted in each genus cluster. The phylogenetic tree on the left represents the relationships among genus clusters, indicating with a star the ones for which

isolated representatives are available. The labels "*Lg*", "*Lm*" and "*Lp*" represent the genus clusters including *L. geestiana*, *L. micdadei* and *L. pneumophila*, respectively.

2.6 *Legionellaceae* species have distinct environmental distributions

Given the genotypic diversity among *Legionellaceae* species, it is expected that they will have distinct, but partly overlapping, ecological niches and environmental distributions. To assess this, we used the OTUs provided by MicrobeAtlas[112], linking them to the full-length 16S rRNA genes of 38 species clusters through reciprocal best hits (Table S9) and enabling us to assess their environmental preferences, as well as their co-occurrence. Such analysis detected up to 13 OTUs in the same sample, confirming previous reports of species co-occurrence[113]. However, the OTUs distribution over the environmental classes varied greatly (Figure 4a). Most OTUs were found in all classes, with a surprisingly high number of samples belonging to animal-associated environments. A closer inspection of the tags related to the animal-associated hosts (Figure S11) revealed how many samples were retrieved from the gastrointestinal tract or faeces, suggesting that these bacteria are shed by infected individuals. In fact, gastrointestinal *Legionella* infections have been reported, even though often overlooked[114]. While in general the OTUs were mostly found in aquatic samples, especially in the case of the ones linked to *L. antarctica* and two uncultivated species, several OTUs showed a drastically higher prevalence of soil samples, such as the ones linked to *L. busanensis*, *L. clemsonensis*, *L. saoudiensis*, *L. jamestowniensis* and *L. fallonii*, or animal-associated ones, as *L. beliardensis* and *L. maceachernii*. Despite being present in samples from the same environmental classes, the *Legionellaceae* OTUs presented weak correlations among their relative abundances (Figure 4b), highlighting their different environmental prevalence. Yet, OTUs which were found to belong to the same metabolic cluster (Table S6) showed significantly higher correlations among themselves than with others (t-test, p-val = 0.04), corroborating the importance of the link between metabolic potential and ecological niche. In fact, the shared evolutionary history plays a role in defining bacterial environmental niches, with stronger negative correlations were observed with growing phylogenetic distances (Spearman correlation, $\rho = -0.26$, p-val < 0.001). Besides this, the OTUs linked to *L. micdadei*, *L. shakespearei* and two uncultivated species presented significant negative correlations with at least other 19 OTUs, possibly arising either due to the interaction between different *Legionellaceae* species, other microorganisms or mutually exclusive environmental niches which should be verified experimentally. Future research identifying the genes responsible for the adaptation to diverse environments could help in source tracking by providing insights on the environmental origin of clinical specimens.

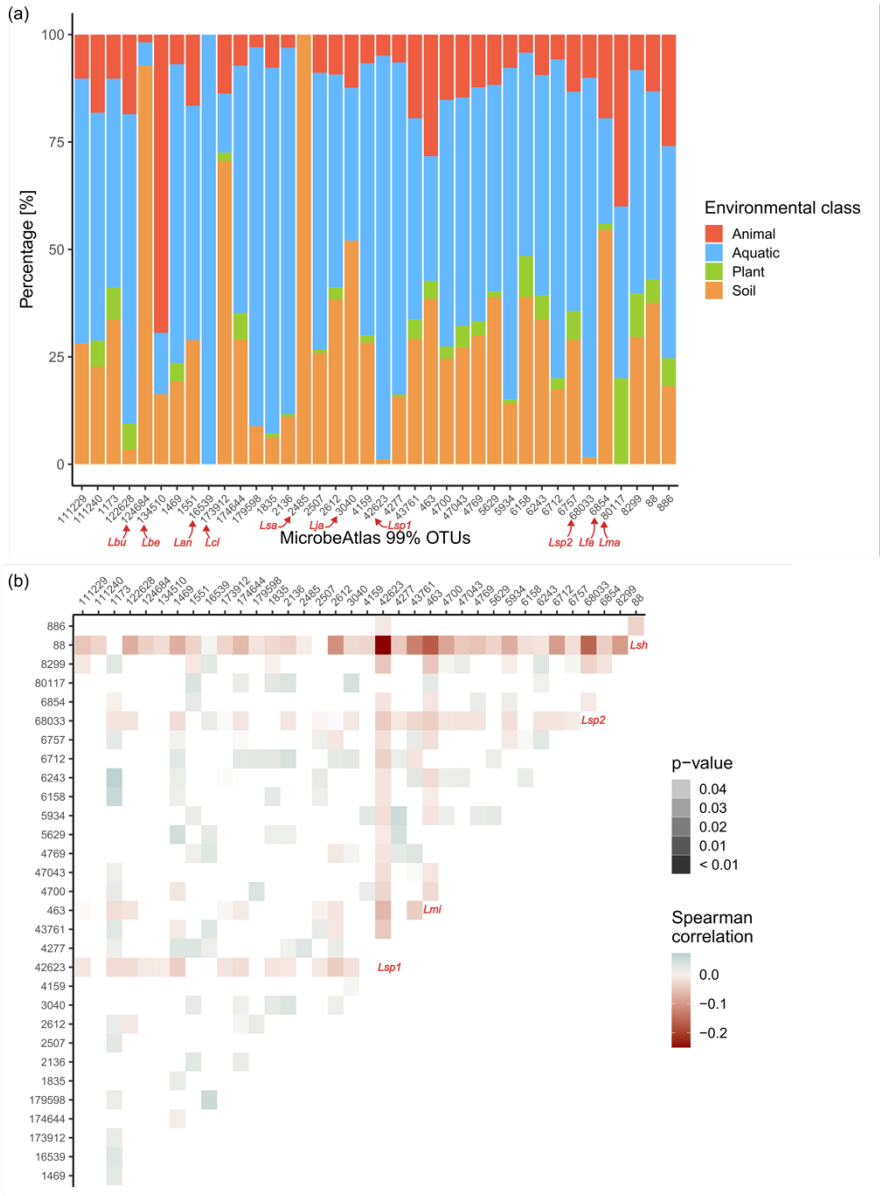


Figure 4. (a) Distribution among environmental classes of the samples in which each MicrobeAtlas OTU (99% similarity) was found. (b) Correlation matrix among the relative abundance of the MicrobeAtlas OTUs. Species labels are the following: *L. antarctica (Lan), L. beliardensis (Lbe), L. busanensis (Lbu), L. clemsoniensis (Lcl), L. fallonii (Lfa), L. jamestowniensis (Lja), L. maceachernii (Lma), L. micdadei (Lmi), L. saoudiensis (Lsa), L. shakespearei (Lsh), L. spp cluster 94_0 (Lsp1), L. spp cluster 85_0 (Lsp2).*

2.7 Implications for future research

This study highlights the importance of genomic data to expand the phylogenetic diversity and characterize the different traits linked to host invasion and metabolism in intracellular bacteria such as *Legionellaceae*. Given the variety of traits and environmental preferences that can be observed even in closely related microorganisms, future research should aim to discern between different genus clusters and species and avoid conflating microorganisms with distinct niches. This is especially

important in cases where closely related species vary in their clinical or biotechnological importance, as the case of *Legionellaceae*. Methods to achieve this in environmental surveys already exist (e.g., clade-specific primers, full-length 16S rRNA sequencing, strain-resolved metagenomics)[115–117], but their adoption is still limited. Such methods could be applied to achieve a detailed characterization of the environmental prevalence of single species, as well as linking co-occurrence pattern across environments to similarities in functional traits[118]. In parallel to environmental surveys, given the known difficulty in the isolation of host-associated microorganisms, future research should also focus on exploiting the metabolic information contained in MAGs for the isolation of the novel identified species and more (e.g., [119]). Achieving this would allow to better understand the evolutionary processes related to a diverse facultative intracellular clade such as *Legionellaceae.* Moreover, this would also enable to test the traits predicted from genomic data and their relevance for the persistence of these bacteria in the environment and their interactions with other microorganisms. With respect to this latter point, the environmental interactions between protistan host and facultative intracellular constitute a major research gap for our understanding of their ecology. This is especially relevant for microorganisms with wide host ranges, such as *Legionellaceae* bacteria[120], for which it is necessary to identify their and the hosts traits affecting the outcome of the interaction. Such trait-based approach would allow building a mechanistic understanding of the host-colonization process and providing the predictive power to estimate interaction of yet-to-be-cultivated species. Such knowledge will not only have scientific value, but would also allow for better optimization of biotechnological processes and aid in the safeguard public health by allowing the improvement of relevant legislation.

3 Methods

3.1 Genomes retrieval, quality assessment and comparison

A total of 5023 genomes classified as belonging to the family *Legionellaceae* were identified from publicly-available sources (date last accession: 15/05/2023, Table S9). Sources included NCBI Assembly database, ENA, JGI and other several large-scale metagenomics survey[121–129]. In addition, several genomes from Wee and coworkers[124] were assembled de-novo using spades v3.15.4[130] after performing read trimming using fastp v0.23.2[131] and the removal of known UniVec contaminant[122] using BWA-MEM2 v2.2.1[132] and samtools v1.16.1[133]. The quality of all the genomes and their taxonomic affiliation was assessed using CheckM2 v1.0.1[134], and GUNC v1.0.5[135] and GTDB-Tk v2.3.0[136] in order to remove genomes with more than 5% contamination, less than 85% complete and not belonging to the investigated family. Nineteen genomes belonging to species without a scientific name which failed GUNC clade separation score or CheckM2 contamination thresholds were manually refined using Anvi'o v8[137] eventually removing contigs with divergent k-mers profiles allowing to recover 10 of them. The sequence type of all *L. pneumophila* genomes was obtained using ONTmompS v2.0.0[138]. Genomes dereplication was performed similarly to previous published strategies[139] performing first an initial clustering at 90% ANI using MASH v1.1[140], which was then followed by a secondary one within each primary cluster using skani v0.2.1[141] at 95% and 99.5% ANI. Representative genomes for species (i.e., 95% ANI) and genomovars (i.e., 99.5% ANI) were chosen as the ones with the highest value derived from the following formula: *Completeness* – 5**Contamination* + 0.5*log(*N50*) + *Centrality*/100, using CheckM2 *Completeness* and *Contamination* estimates and where *N50* is the minimum contig length for which at least half of the genome nucleotides belongs to contigs

with the *N50* length or longer, while *Centrality* is the average ANI with respect to genomes within the same secondary cluster. Due to the high number of genomovars found for *L. pneumophila* and *L. fraseri,* only the species representative and the three representatives of the genomovar with the highest number of genomes were selected. Species' clinical-associations were retrieved from published literature[142].

3.2 Proteins predictions and functional annotations

Proteins were predicted using Bakta v1.8.1[143] (v5.0, full database) and annotated using eggNOG-mapper v2.1.12[144]. The similarity of the protein content across selected genomes was assessed using FastAAI v0.1.20[145]. T4SS were detected using CONJScan v2.0.1[146], while other secretion systems, flagella and pili using TXSScan v1.1.3[147]. Differently from TXSScan default, T6SSs were considered as present in case more than nine proteins were detected in order to match previous reports of the presence of this machinery[51]. Plasmid-associated contigs were detected using Platon v1.7[148]. The presence of capsule systems was assessed using CapsuleFinder v1.0.5[86]. Sensory and regulatory proteins were selected based on eggNOG-mapper annotations relying on KEGG profiles collected in MiST v4.0 database[149] marked as signal transduction specific. HMMER v3.4[150] was further used to identify chemotaxis and methyl-accepting proteins based on the hidden markov models developed for the MiST database and considering a bitscore threshold of 25. KEGG metabolic pathways completeness was estimated using DRAM v1.4.6[151], while the amino acids biosynthetic potential was evaluated using GapMind and gapseq. CAZYmes were annotated against the dbCAN3 database using dbcan_run v4.0.0[152], retaining only proteins identified by at least two detection methods and classifying CAZYmes families from annotation majority voting and relying on dbCAN_sub substrates prediction. Peptidases were identified searching the predicted proteins against the MEROPS v12.4 database[153] using MMseqs2 v14-7e284[154] retaining proteins showing at least 30% identity over 65% query coverage and displaying a bitscore ≥50. Secreted proteins were identified using SignalP v6.0[155]. Clusters of genomes presenting similar metabolic capabilities were obtained through HDBSCAN clustering based on the euclidean distance of metabolic pathways completeness. Homologs of selected *L. pneumophila Paris* proteins were detected using BlastP v2.12.0[156] using the same thresholds used for MMseqs2 and inspecting Bakta and eggNOG-mapper annotations (e.g., PFAMs detected). Type 1 secretion systems were considered as present if both LssD/lrpssD/lb1ssD/lb2ssD and LssB/lrpssB/lb1ssB/lb2ssB proteins were detected in a genome within a five proteins neighbourhood[58]. Eventual duplicated matches with TXSScan results were assessed using a custom script. Analysis of putative HGT events was performed using HGTector v2.0b3[157] using the proteins included in GTDB[158] release 2014 and a e-value threshold of 1^{-20}. The *Legionellaceae* family as self-group and Gammaproteobacteria as close-group.

3.3 Pangenome and phylogeny construction

Predicted proteins were used to perform calculations using PPanGGOLiN v1.2.105[159]. A pangenome employing a protein clustering identity percentage threshold of 80% was used to identify core proteins (i.e., present in ≥95% of genomes) to reconstruct phylogeny. Identified proteins were searched for homologs in the genome of *C. burnetii* (RefSeq accession: GCF_005280755.1). Proteins with detected homologs were then aligned using MAFFT G-INS-i as implemented in Concatenator v0.2.1[160,161] and

used to calculate maximum likelihood phylogenetic trees using IQ-TREE v2.2.2.6[162] under the LG model and automated model selection provided by the ModelFinder utility[163]. Support values were calculated based on 1000 ultrafast bootstraps with bootstrap tree optimization[164]. The resulting phylogenetic tree was visualized and annotated in iTOL v6[165]. An initial genus cluster membership was obtained using GTDB-Tk de-novo workflow using calibrated RED values. Successively, cluster membership was refined manually to remove polyphyly. A second pangenome employing an identity percentage threshold of 50% was used to identify gene families among *Legionellaceae* genomes and identify shared proteins, as suggested by PPanGGOLiN authors.

3.4 Assessment of species environmental distribution

Full length 16S rRNA genes were retrieved from the Bakta annotation of the selected species representatives or, in case, from other genomes belonging to the same species cluster. One random 16S rRNA copy was selected per genome, after removing the ones with an excessive number of ambiguous bases. Such sequences were matched to the full-length OTUs (99% similarity) retrieved from MicrobeAtlas v3.0 (downloaded on: 16/9/2024) through MMseqs2's reciprocal best hit workflow and removing OTUs which showed best hits to multiple species clusters. Major environmental classes were obtained from MicrobeAtlas metadata, selecting only samples for which only a single class was listed. Spearman correlations were used to estimate the concordance of the environmental distribution among selected OTUs.

3.5 Statistical testing and data analysis

Analysis of the bioinformatic results was performed in R v4.3.0[166] leveraging the functions provided by libraries Tidyverse v2.0.0[167], APE v5.7-1[168], dbscan v1.2-0[169] and vegan v2.6-4[170].

4 Acknowledgements

This research was funded through the LeCo project (Legionella Control in Buildings; Aramis nr.: 4.20.01) by the Federal Food Safety and Veterinary Office (FSVO), in partnership with the Federal Offices of Public Health (FOPH) and Energy (SFOE) of Switzerland, as well as Eawag discretionary funding. The authors would also like to acknowledge Alyssa Henderson for her constructive feedback on genomic-scale metabolic models and Valeria Gaia for providing additional *Legionellaceae* isolates.

5 References

1. Thompson, L. R. *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457–463 (2017).
2. Blount, Z. D. The unexhausted potential of E. coli. *eLife* **4**, e05826 (2015).
3. Hrovat, K., Dutilh, B. E., Medema, M. H. & Melkonian, C. Taxonomic resolution of different 16S rRNA variable regions varies strongly across plant-associated bacteria. *ISME Commun.* **4**, ycae034 (2024).

4. Leigh, J. A., Albers, S.-V., Atomi, H. & Allers, T. Model organisms for genetics in the domain Archaea: methanogens, halophiles, *Thermococcales* and *Sulfolobales*. *FEMS Microbiol. Rev.* **35**, 577–608 (2011).
5. Beukes, C. W. *et al.* Genome Data Provides High Support for Generic Boundaries in Burkholderia Sensu Lato. *Front. Microbiol.* **8**, 1154 (2017).
6. Van Rossum, T., Ferretti, P., Maistrenko, O. M. & Bork, P. Diversity within species: interpreting strains in microbiomes. *Nat. Rev. Microbiol.* **18**, 491–506 (2020).
7. Crosby, K. C. *et al.* Genomic delineation and description of species and within-species lineages in the genus Pantoea. *Front. Microbiol.* **14**, 1254999 (2023).
8. Baker-Goering, M., Roy, K., Edens, C. & Collier, S. Economic Burden of Legionnaires' Disease, United States, 2014. *Emerg. Infect. Dis.* **27**, 255–257 (2021).
9. Moffa, M. A. *et al.* Legionellosis on the rise: A scoping review of sporadic, community-acquired incidence in the United States. *Epidemiol. Infect.* **151**, e133 (2023).
10. LPSN. Family Legionellaceae. https://lpsn.dsmz.de/family/legionellaceae.
11. Yu, V. L. *et al.* Distribution of *Legionella* Species and Serogroups Isolated by Culture in Patients with Sporadic Community-Acquired Legionellosis: An International Collaborative Survey. *J. Infect. Dis.* **186**, 127–128 (2002).
12. van der Wielen, P. W. J. J., Wierenga, W., Oesterholt, F. I. H. M., Oostdijk, A. & van der Werff, A. *Rightly to an effective Legionella prevention (in Dutch)*. (2021).
13. National Academies of Sciences, Engineering, and Medicine. *Management of Legionella in Water Systems*. 25474 (National Academies Press, Washington, D.C., 2020). doi:10.17226/25474.
14. Van Kenhove, E., Dinne, K., Janssens, A. & Laverge, J. Overview and comparison of Legionella regulations worldwide. *Am. J. Infect. Control* **47**, 968–978 (2019).
15. Oliva, G., Sahr, T. & Buchrieser, C. The Life Cycle of L. pneumophila: Cellular Differentiation Is Linked to Virulence and Metabolism. *Front. Cell. Infect. Microbiol.* **8**, 3 (2018).
16. Joseph, S. J. *et al.* Dynamics of genome change among Legionella species. *Sci. Rep.* **6**, 33442 (2016).
17. Gomez-Valero, L. *et al.* More than 18,000 effectors in the *Legionella* genus genome provide multiple, independent combinations for replication in human cells. *Proc. Natl. Acad. Sci.* **116**, 2265–2273 (2019).
18. Tobias, N. J. *et al. Legionella* shows a diverse secondary metabolism dependent on a broad spectrum Sfp-type phosphopantetheinyl transferase. *PeerJ* **4**, e2720 (2016).
19. Graells, T., Ishak, H., Larsson, M. & Guy, L. The all-intracellular order Legionellales is unexpectedly diverse, globally distributed and lowly abundant. *FEMS Microbiol. Ecol.* **94**, (2018).
20. Vartoukian, S. R., Palmer, R. M. & Wade, W. G. Strategies for culture of 'unculturable' bacteria: Culturing the unculturable. *FEMS Microbiol. Lett.* **309**, 1–7 (2010).
21. Rodriguez-R, L. M. *et al.* An ANI gap within bacterial species that advances the definitions of intra-species units. *mBio* **15**, e02696-23 (2024).
22. Oren, A. & Garrity, G. Notification of changes in taxonomic opinion previously published outside the IJSEM. List of changes in taxonomic opinion no. 35. *Int. J. Syst. Evol. Microbiol.* **72**, (2022).
23. Saini, N. & Gupta, R. S. A robust phylogenetic framework for members of the order Legionellales and its main genera (Legionella, Aquicella, Coxiella and Rickettsiella) based on phylogenomic analyses and identification of molecular markers demarcating different clades. *Antonie Van Leeuwenhoek* **114**, 957–982 (2021).
24. Kozak-Muiznieks, N. A. *et al.* Comparative genome analysis reveals a complex population structure of Legionella pneumophila subspecies. *Infect. Genet. Evol.* **59**, 172–185 (2018).

25. Svetlicic, E. *et al.* Genomic characterization and assessment of pathogenic potential of Legionella spp. isolates from environmental monitoring. *Front. Microbiol.* **13**, 1091964 (2023).
26. Taboada, E. N. *et al.* Comparative genomic assessment of Multi-Locus Sequence Typing: rapid accumulation of genomic heterogeneity among clonal isolates of Campylobacter jejuni. *BMC Evol. Biol.* **8**, 229 (2008).
27. Prussin, A. J., Schwake, D. O. & Marr, L. C. Ten questions concerning the aerosolization and transmission of Legionella in the built environment. *Build. Environ.* **123**, 684–695 (2017).
28. Philippot, L. *et al.* The ecological coherence of high bacterial taxonomic ranks. *Nat. Rev. Microbiol.* **8**, 523–529 (2010).
29. Konstantinidis, K. T. & Tiedje, J. M. Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. *Curr. Opin. Microbiol.* **10**, 504–509 (2007).
30. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
31. Chambers, S. T., Slow, S., Scott-Thomas, A. & Murdoch, D. R. Legionellosis Caused by Non-Legionella pneumophila Species, with a Focus on Legionella longbeachae. *Microorganisms* **9**, 291 (2021).
32. Muder, R. R. & Yu, V. L. Infection Due to *Legionella* Species Other Than *L. pneumophila*. *Clin. Infect. Dis.* **35**, 990–998 (2002).
33. Wang, C., Chuai, X. & Liang, M. Legionella feeleii: pneumonia or Pontiac fever? Bacterial virulence traits and host immune response. *Med. Microbiol. Immunol. (Berl.)* **208**, 25–32 (2019).
34. Garrity, G. M., Brown, A. & Vickers, R. M. Tatlockia and Fluoribacter: Two New Genera of Organisms Resembling Legionella pneumophila. *Int. J. Syst. Bacteriol.* **30**, 609–614 (1980).
35. Fry, N. K., Warwick, S., Saunders, N. A. & Embley, T. M. The use of 16S ribosomal RNA analyses to investigate the phylogeny of the family Legionellaceae. *J. Gen. Microbiol.* **137**, 1215–1222 (1991).
36. Hassler, H. B. *et al.* Phylogenies of the 16S rRNA gene and its hypervariable regions lack concordance with core genome phylogenies. *Microbiome* **10**, 104 (2022).
37. Wang, Z. *et al.* Host specificity of plant-associated bacteria is negatively associated with genome size and host abundance along a latitudinal gradient. *Ecol. Lett.* **27**, e14447 (2024).
38. Köstlbacher, S. *et al.* Pangenomics reveals alternative environmental lifestyles among chlamydiae. *Nat. Commun.* **12**, 4021 (2021).
39. Dewar, A. E., Hao, C., Belcher, L. J., Ghoul, M. & West, S. A. Bacterial lifestyle shapes pangenomes. *Proc. Natl. Acad. Sci.* **121**, e2320170121 (2024).
40. Tettelin, H., Riley, D., Cattuto, C. & Medini, D. Comparative genomics: the bacterial pan-genome. *Curr. Opin. Microbiol.* **11**, 472–477 (2008).
41. Orkis, L. T. *et al.* Environmental sources of community-acquired legionnaires' disease: A review. *Int. J. Hyg. Environ. Health* **221**, 764–774 (2018).
42. Kanarek, P., Bogiel, T. & Breza-Boruta, B. Legionellosis risk—an overview of Legionella spp. habitats in Europe. *Environ. Sci. Pollut. Res.* **29**, 76532–76542 (2022).
43. Gast, R. J., Moran, D. M., Dennett, M. R., Wurtsbaugh, W. A. & Amaral-Zettler, L. A. Amoebae and Legionella pneumophila in saline environments. *J. Water Health* **9**, 37–52 (2011).
44. Casadevall, A. Evolution of Intracellular Pathogens. *Annu. Rev. Microbiol.* **62**, 19–33 (2008).
45. Boamah, D. K., Zhou, G., Ensminger, A. W. & O'Connor, T. J. From Many Hosts, One Accidental Pathogen: The Diverse Protozoan Hosts of Legionella. *Front. Cell. Infect. Microbiol.* **7**, 477 (2017).
46. Poirier, V. & Av-Gay, Y. Intracellular Growth of Bacterial Pathogens: The Role of Secreted Effector Proteins in the Control of Phagocytosed Microorganisms. *Microbiol. Spectr.* **3**, 3.6.12 (2015).

47. Wallden, K., Rivera-Calzada, A. & Waksman, G. Microreview: Type IV secretion systems: versatility and diversity in function: Diversity and versatility of the type IV secretion system. *Cell. Microbiol.* **12**, 1203–1212 (2010).
48. White, R. C. & Cianciotto, N. P. Assessing the impact, genomics and evolution of type II secretion across a large, medically important genus: the Legionella type II secretion paradigm. *Microb. Genomics* **5**, (2019).
49. Duménil, G. & Isberg, R. R. The *Legionella pneumophila* IcmR protein exhibits chaperone activity for IcmQ by preventing its participation in high-molecular-weight complexes. *Mol. Microbiol.* **40**, 1113–1127 (2001).
50. Feldman, M., Zusman, T., Hagag, S. & Segal, G. Coevolution between nonhomologous but functionally similar proteins and their conserved partners in the *Legionella* pathogenesis system. *Proc. Natl. Acad. Sci.* **102**, 12206–12211 (2005).
51. Qin, T., Zhou, H., Ren, H. & Liu, W. Distribution of Secretion Systems in the Genus Legionella and Its Correlation with Pathogenicity. *Front. Microbiol.* **08**, (2017).
52. Söderberg, M. A., Dao, J., Starkenburg, S. R. & Cianciotto, N. P. Importance of Type II Secretion for Survival of *Legionella pneumophila* in Tap Water and in Amoebae at Low Temperatures. *Appl. Environ. Microbiol.* **74**, 5583–5588 (2008).
53. Carrión, V. J. *et al.* Pathogen-induced activation of disease-suppressive functions in the endophytic root microbiome. *Science* **366**, 606–612 (2019).
54. Wang, L. *et al.* Mining of novel secondary metabolite biosynthetic gene clusters from acid mine drainage. *Sci. Data* **9**, 760 (2022).
55. Banks, D., Athresh, A., Al-Habaibeh, A. & Burnside, N. Water from abandoned mines as a heat source: practical experiences of open- and closed-loop strategies, United Kingdom. *Sustain. Water Resour. Manag.* **5**, 29–50 (2019).
56. De Buck, E., Anné, J. & Lammertyn, E. The role of protein secretion systems in the virulence of the intracellular pathogen Legionella pneumophila. *Microbiology* **153**, 3948–3953 (2007).
57. Fuche, F., Vianney, A., Andrea, C., Doublet, P. & Gilbert, C. Functional Type 1 Secretion System Involved in Legionella pneumophila Virulence. *J. Bacteriol.* **197**, 563–571 (2015).
58. Brown, C. L. *et al.* Whole genome sequence analysis reveals the broad distribution of the RtxA type 1 secretion system and four novel putative type 1 secretion systems throughout the Legionella genus. *PLOS ONE* **15**, e0223033 (2020).
59. Gomez-Valero, L. *et al.* Comparative analyses of Legionella species identifies genetic features of strains causing Legionnaires disease. *Genome Biol.* **15**, (2014).
60. Joseph, S. J. *et al.* Dynamics of genome change among Legionella species. *Sci. Rep.* **6**, 33442 (2016).
61. Bandyopadhyay, P., Lang, E. A. S., Rasaputra, K. S. & Steinman, H. M. Implication of the VirD4 Coupling Protein of the Lvh Type 4 Secretion System in Virulence Phenotypes of Legionella pneumophila. *J. Bacteriol.* **195**, 3468–3475 (2013).
62. Meuskens, I., Saragliadis, A., Leo, J. C. & Linke, D. Type V Secretion Systems: An Overview of Passenger Domain Functions. *Front. Microbiol.* **10**, 1163 (2019).
63. Fan, E., Chauhan, N., Udatha, D. B. R. K. G., Leo, J. C. & Linke, D. Type V Secretion Systems in Bacteria. *Microbiol. Spectr.* **4**, 4.1.10 (2016).
64. Singh, R. P. & Kumari, K. Bacterial type VI secretion system (T6SS): an evolved molecular weapon with diverse functionality. *Biotechnol. Lett.* **45**, 309–331 (2023).
65. Gallegos-Monterrosa, R. & Coulthurst, S. J. The ecological impact of a bacterial weapon: microbial interactions and the Type VI secretion system. *FEMS Microbiol. Rev.* **45**, fuab033 (2021).

66. Raina, J.-B., Fernandez, V., Lambert, B., Stocker, R. & Seymour, J. R. The role of microbial motility and chemotaxis in symbiosis. *Nat. Rev. Microbiol.* **17**, 284–294 (2019).
67. Shaheen, M. & Ashbolt, N. J. Differential Bacterial Predation by Free-Living Amoebae May Result in Blooms of Legionella in Drinking Water Systems. *Microorganisms* **9**, 174 (2021).
68. Hoppe, J. *et al.* PilY1 Promotes Legionella pneumophila Infection of Human Lung Tissue Explants and Contributes to Bacterial Adhesion, Host Cell Invasion, and Twitching Motility. *Front. Cell. Infect. Microbiol.* **7**, (2017).
69. Cazalet, C. *et al.* Analysis of the Legionella longbeachae Genome and Transcriptome Uncovers Unique Strategies to Cause Legionnaires' Disease. *PLoS Genet.* **6**, e1000851 (2010).
70. Stewart, C. R., Rossier, O. & Cianciotto, N. P. Surface Translocation by *Legionella pneumophila* : a Form of Sliding Motility That Is Dependent upon Type II Protein Secretion. *J. Bacteriol.* **191**, 1537–1546 (2009).
71. White, R. C. & Cianciotto, N. P. Assessing the impact, genomics and evolution of type II secretion across a large, medically important genus: the Legionella type II secretion paradigm. *Microb. Genomics* **5**, (2019).
72. Matilla, M. A. & Krell, T. The effect of bacterial chemotaxis on host infection and pathogenicity. *FEMS Microbiol. Rev.* **42**, (2018).
73. Duncan, C. *et al.* Lcl of Legionella pneumophila Is an Immunogenic GAG Binding Adhesin That Promotes Interactions with Lung Epithelial Cells and Plays a Crucial Role in Biofilm Formation. *Infect. Immun.* **79**, 2168–2181 (2011).
74. Dietrich, C., Heuner, K., Brand, B. C., Hacker, J. & Steinert, M. Flagellum of *Legionella pneumophila* Positively Affects the Early Phase of Infection of Eukaryotic Host Cells. *Infect. Immun.* **69**, 2116–2122 (2001).
75. Wuichet, K. & Zhulin, I. B. Origins and Diversification of a Complex Signal Transduction System in Prokaryotes. *Sci. Signal.* **3**, (2010).
76. Chaban, B., Hughes, H. V. & Beeby, M. The flagellum in bacterial pathogens: For motility and a whole lot more. *Semin. Cell Dev. Biol.* **46**, 91–103 (2015).
77. Coil, D. A. & Anné, J. Twitching motility in *Legionella pneumophila*. *FEMS Microbiol. Lett.* **293**, 271–277 (2009).
78. Li, L., Mendis, N., Trigui, H. & Faucher, S. P. Transcriptomic changes of Legionella pneumophila in water. *BMC Genomics* **16**, 637 (2015).
79. Aurass, P., Kim, S., Pinedo, V., Cava, F. & Isberg, R. R. Identification of Genes Required for Long-Term Survival of *Legionella pneumophila* in Water. *mSphere* **8**, e00454-22 (2023).
80. Speiser, Y., Zusman, T., Pasechnek, A. & Segal, G. The Legionella pneumophila Incomplete Phosphotransferase System Is Required for Optimal Intracellular Growth and Maximal Expression of PmrA-Regulated Effectors. *Infect. Immun.* **85**, e00121-17 (2017).
81. Lynch, D., Fieser, N., GlÃ¶ggler, K., Forsbach-Birk, V. & Marre, R. The response regulator LetA regulates the stationary-phase stress response in Legionella pneumophila and is required for efficient infection of Acanthamoeba castellanii. *FEMS Microbiol. Lett.* **219**, 241–248 (2003).
82. Appelt, S. & Heuner, K. The Flagellar Regulon of Legionella—A Review. *Front. Cell. Infect. Microbiol.* **7**, 454 (2017).
83. Zusman, T., Speiser, Y. & Segal, G. Two Fis Regulators Directly Repress the Expression of Numerous Effector-Encoding Genes in Legionella pneumophila. *J. Bacteriol.* **196**, 4172–4183 (2014).
84. Charpentier, X., Faucher, S. P., Kalachikov, S. & Shuman, H. A. Loss of RNase R Induces Competence Development in *Legionella pneumophila*. *J. Bacteriol.* **190**, 8126–8136 (2008).

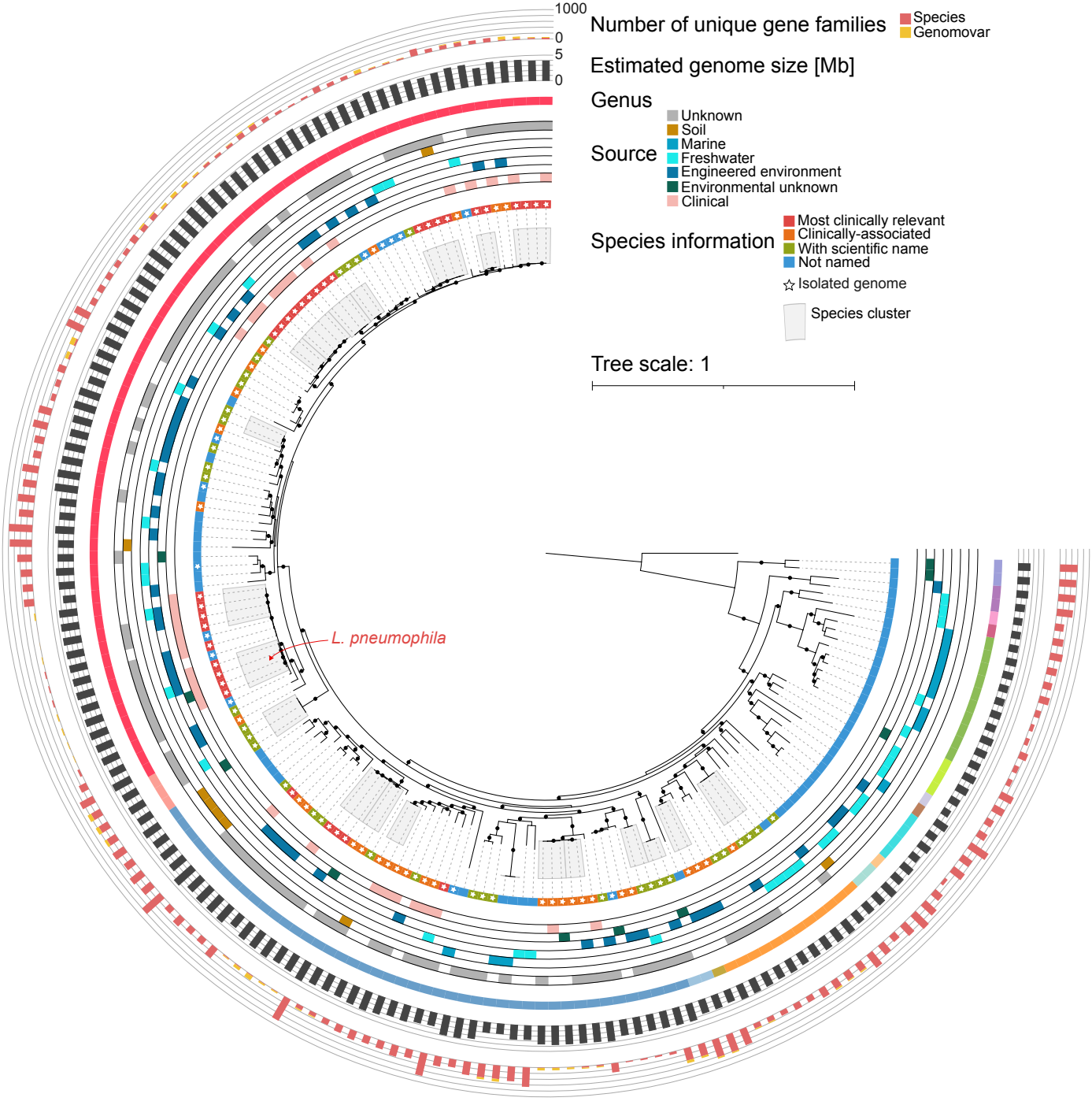
85. Hickey, E. K. & Cianciotto, N. P. An iron- and fur-repressed Legionella pneumophila gene that promotes intracellular infection and encodes a protein with similarity to the Escherichia coli aerobactin synthetases. *Infect. Immun.* **65**, 133–143 (1997).
86. Rendueles, O., Garcia-Garcerà, M., Néron, B., Touchon, M. & Rocha, E. P. C. Abundance and co-occurrence of extracellular capsules increase environmental breadth: Implications for the emergence of pathogens. *PLOS Pathog.* **13**, e1006525 (2017).
87. Schmidt, S. *et al.* The unique Legionella longbeachae capsule favors intracellular replication and immune evasion. *PLOS Pathog.* **20**, e1012534 (2024).
88. Hsieh, S. A., Donermeyer, D. L., Horvath, S. C. & Allen, P. M. Phase-variable bacteria simultaneously express multiple capsules: Read the story behind the paper on the Microbe Post here. *Microbiology* **167**, (2021).
89. Eylert, E. *et al.* Isotopologue Profiling of Legionella pneumophila. *J. Biol. Chem.* **285**, 22232–22243 (2010).
90. Eisenreich, W. & Heuner, K. The life stage-specific pathometabolism of *Legionella pneumophila*. *FEBS Lett.* **590**, 3868–3886 (2016).
91. Harada, E., Iida, K.-I., Shiota, S., Nakayama, H. & Yoshida, S.-I. Glucose Metabolism in *Legionella pneumophila* : Dependence on the Entner-Doudoroff Pathway and Connection with Intracellular Bacterial Growth. *J. Bacteriol.* **192**, 2892–2899 (2010).
92. Häuslein, I., Manske, C., Goebel, W., Eisenreich, W. & Hilbi, H. Pathway analysis using ^{13}C-glycerol and other carbon tracers reveals a bipartite metabolism of *Legionella pneumophila*. *Mol. Microbiol.* **100**, 229–246 (2016).
93. Hu, X.-P., Schroeder, S. & Lercher, M. J. Proteome efficiency of metabolic pathways in *Escherichia coli* increases along the nutrient flow. *mSystems* **8**, e00760-23 (2023).
94. Gillmaier, N. *et al.* Growth-related Metabolism of the Carbon Storage Poly-3-hydroxybutyrate in Legionella pneumophila. *J. Biol. Chem.* **291**, 6471–6482 (2016).
95. Manske, C., Schell, U. & Hilbi, H. Metabolism of *myo* -Inositol by Legionella pneumophila Promotes Infection of Amoebae and Macrophages. *Appl. Environ. Microbiol.* **82**, 5000–5014 (2016).
96. James, B. W., Mauchline, W. S., Dennis, P. J., Keevil, C. W. & Wait, R. Poly-3-Hydroxybutyrate in *Legionella pneumophila* , an Energy Source for Survival in Low-Nutrient Environments. *Appl. Environ. Microbiol.* **65**, 822–827 (1999).
97. Turner, B. L., Papházy, M. J., Haygarth, P. M. & Mckelvie, I. D. Inositol phosphates in the environment. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **357**, 449–469 (2002).
98. Eisenhofer, R., Odriozola, I. & Alberdi, A. Impact of microbial genome completeness on metagenomic functional inference. *ISME Commun.* **3**, 12 (2023).
99. Gomez-Valero, L. *et al.* More than 18,000 effectors in the *Legionella* genus genome provide multiple, independent combinations for replication in human cells. *Proc. Natl. Acad. Sci.* **116**, 2265–2273 (2019).
100. Lopez-Vazquez, C. M. *et al.* Modeling the PAO–GAO competition: Effects of carbon source, pH and temperature. *Water Res.* **43**, 450–462 (2009).
101. Nguyen, T. T. H., Myrold, D. D. & Mueller, R. S. Distributions of Extracellular Peptidases Across Prokaryotic Genomes Reflect Phylogeny and Habitat. *Front. Microbiol.* **10**, 413 (2019).
102. McKinlay, J. B. Are Bacteria Leaky? Mechanisms of Metabolite Externalization in Bacterial Cross-Feeding. *Annu. Rev. Microbiol.* **77**, 277–297 (2023).
103. Ramoneda, J., Jensen, T. B. N., Price, M. N., Casamayor, E. O. & Fierer, N. Taxonomic and environmental distribution of bacterial amino acid auxotrophies. *Nat. Commun.* **14**, 7608 (2023).

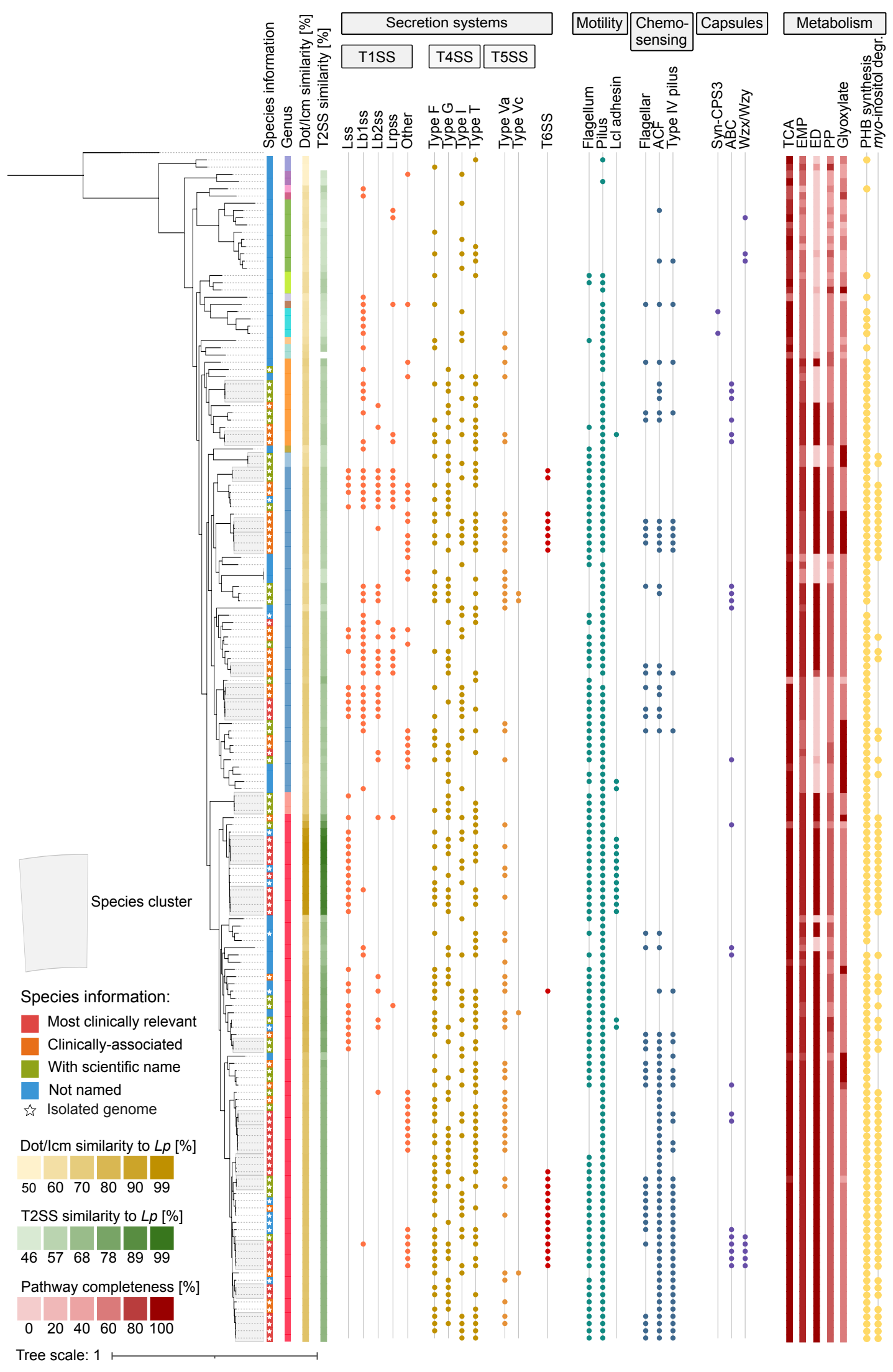
104. D'Souza, G. *et al.* Less is more: selective advantages can explain the prevalent loss of biosynthetic genes in bacteria: adaptive loss of biosynthetic genes in bacteria. *Evolution* **68**, 2559–2570 (2014).
105. Price, M. N., Deutschbauer, A. M. & Arkin, A. P. GapMind: Automated Annotation of Amino Acid Biosynthesis. *mSystems* **5**, 10.1128/msystems.00291-20 (2020).
106. Zimmermann, J., Kaleta, C. & Waschina, S. gapseq: informed prediction of bacterial metabolic pathways and reconstruction of accurate metabolic models. *Genome Biol.* **22**, 81 (2021).
107. George, J. R., Pine, L., Reeves, M. W. & Harrell, W. K. Amino acid requirements of Legionella pneumophila. *J. Clin. Microbiol.* **11**, 286–291 (1980).
108. Stipanuk, M. H. SULFUR AMINO ACID METABOLISM: Pathways for Production and Removal of Homocysteine and Cysteine. *Annu. Rev. Nutr.* **24**, 539–577 (2004).
109. Kredich, N. M. Biosynthesis of Cysteine. *EcoSal Plus* **3**, 10.1128/ecosalplus.3.6.1.11 (2008).
110. Tesh, M. J. & Miller, R. D. Amino acid requirements for Legionella pneumophila growth. *J. Clin. Microbiol.* **13**, 865–869 (1981).
111. Lee, T. C., Vickers, R. M., Yu, V. L. & Wagener, M. M. Growth of 28 Legionella species on selective culture media: a comparative study. *J. Clin. Microbiol.* **31**, 2764–2768 (1993).
112. Matias Rodrigues, J. F., Schmidt, T. S. B., Tackmann, J. & Von Mering, C. MAPseq: highly efficient k-mer search with confidence estimates, for rRNA sequence analysis. *Bioinformatics* **33**, 3808–3810 (2017).
113. Lesnik, R., Brettar, I. & Höfle, M. G. *Legionella* species diversity and dynamics from surface reservoir to tap water: from cold adaptation to thermophily. *ISME J.* **10**, 1064–1080 (2016).
114. Emuron, D. & Senadhi, V. Gastrointestinal Legionella: An Underecognized Entity and Cause of Secretory Diarrhea: A Case and Review of the Literature: 889. *Off. J. Am. Coll. Gastroenterol. ACG* **106**, (2011).
115. Segata, N. On the Road to Strain-Resolved Comparative Metagenomics. *mSystems* **3**, e00190-17 (2018).
116. Johnson, J. S. *et al.* Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat. Commun.* **10**, 5029 (2019).
117. Pereira, R. P. A., Peplies, J., Brettar, I. & Höfle, M. G. Development of a genus-specific next generation sequencing approach for sensitive and quantitative determination of the Legionella microbiome in freshwater systems. *BMC Microbiol.* **17**, 79 (2017).
118. Ramoneda, J., Hoffert, M., Stallard-Olivera, E., Casamayor, E. O. & Fierer, N. Leveraging genomic information to predict environmental preferences of bacteria. *ISME J.* wrae195 (2024) doi:10.1093/ismejo/wrae195.
119. Omsland, A., Hackstadt, T. & Heinzen, R. A. Bringing Culture to the Uncultured: Coxiella burnetii and Lessons for Obligate Intracellular Bacterial Pathogens. *PLoS Pathog.* **9**, e1003540 (2013).
120. Boamah, D. K., Zhou, G., Ensminger, A. W. & O'Connor, T. J. From Many Hosts, One Accidental Pathogen: The Diverse Protozoan Hosts of Legionella. *Front. Cell. Infect. Microbiol.* **7**, 477 (2017).
121. Chen, I.-M. A. *et al.* IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Res.* **47**, D666–D677 (2019).
122. Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. *Nucleic Acids Res.* **44**, D67–D72 (2016).
123. Leinonen, R., Sugawara, H., Shumway, M., & on behalf of the International Nucleotide Sequence Database Collaboration. The Sequence Read Archive. *Nucleic Acids Res.* **39**, D19–D21 (2011).

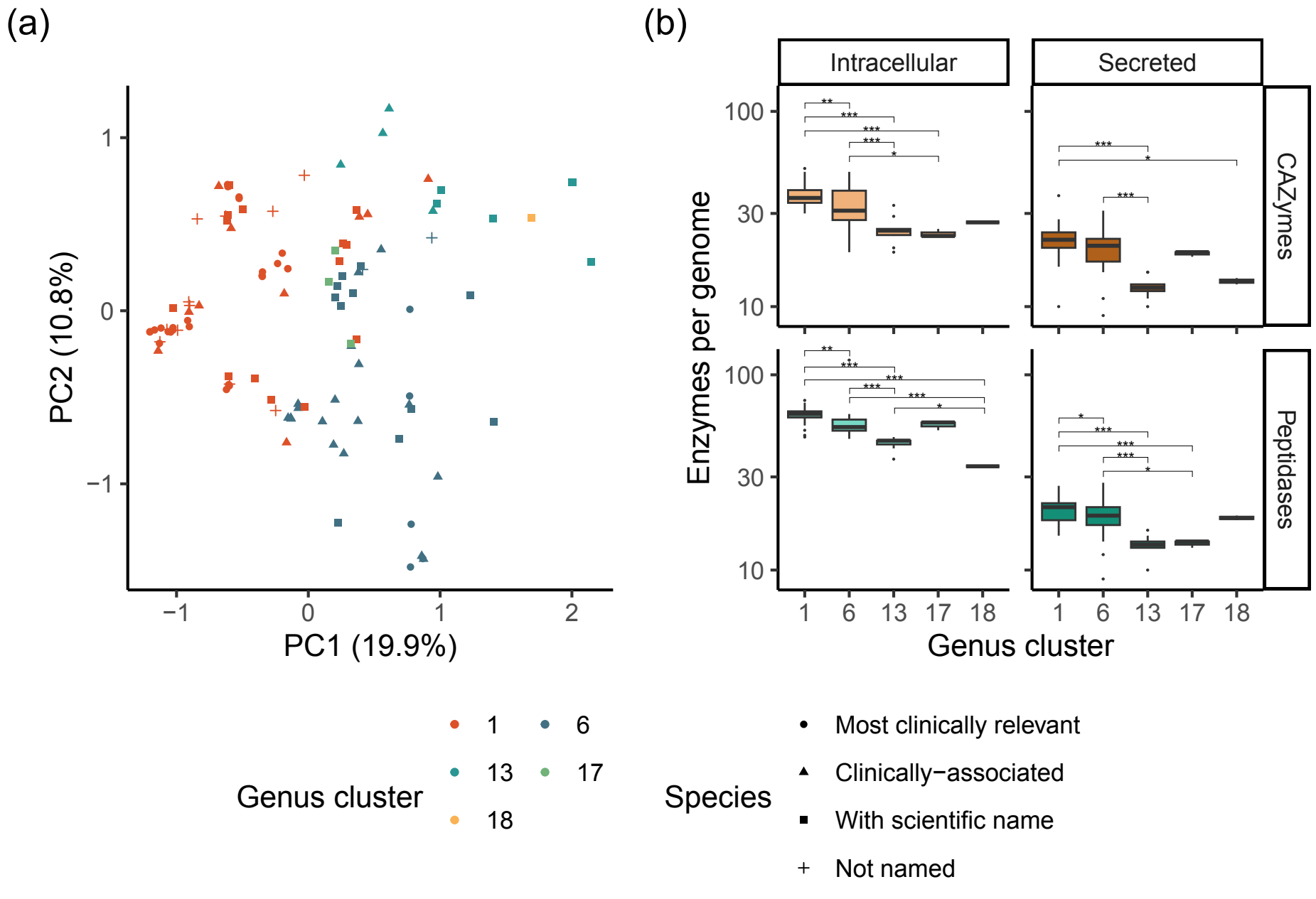
124. Wee, B. A. *et al.* Population analysis of Legionella pneumophila reveals a basis for resistance to complement-mediated killing. *Nat. Commun.* **12**, 7165 (2021).
125. Buck, M. *et al.* Comprehensive dataset of shotgun metagenomes from oxygen stratified freshwater lakes and ponds. *Sci. Data* **8**, 131 (2021).
126. Gurbich, T. A. *et al.* MGnify Genomes: A Resource for Biome-specific Microbial Genome Catalogues. *J. Mol. Biol.* 168016 (2023) doi:10.1016/j.jmb.2023.168016.
127. Nayfach, S. *et al.* A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* **39**, 499–509 (2021).
128. Nishimura, Y. & Yoshizawa, S. The OceanDNA MAG catalog contains over 50,000 prokaryotic genomes originated from various marine environments. *Sci. Data* **9**, 305 (2022).
129. Sudarshan, A. S. *et al.* New Drinking Water Genome Catalog Identifies a Globally Distributed Bacterial Genus Adapted to Disinfected Drinking Water Systems. *Environ. Sci. Technol.* **58**, 16475–16487 (2024).
130. Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A. & Korobeynikov, A. Using SPAdes De Novo Assembler. *Curr. Protoc. Bioinforma.* **70**, e102 (2020).
131. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
132. Vasimuddin, Md., Misra, S., Li, H. & Aluru, S. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. in *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)* 314–324 (IEEE, Rio de Janeiro, Brazil, 2019). doi:10.1109/IPDPS.2019.00041.
133. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
134. Chklovski, A., Parks, D. H., Woodcroft, B. J. & Tyson, G. W. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat. Methods* **20**, 1203–1212 (2023).
135. Orakov, A. *et al.* GUNC: detection of chimerism and contamination in prokaryotic genomes. *Genome Biol.* **22**, 178 (2021).
136. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* **38**, 5315–5316 (2022).
137. Eren, A. M. *et al.* Community-led, integrated, reproducible multi-omics with anvi'o. *Nat. Microbiol.* **6**, 3–6 (2020).
138. Krøvel, A. V. *et al.* Long-read sequencing for reliably calling the mompS allele in Legionella pneumophila sequence-based typing. *Front. Cell. Infect. Microbiol.* **13**, 1176182 (2023).
139. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).
140. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).
141. Shaw, J. & Yu, Y. W. Fast and robust metagenomic sequence comparison through sparse chaining with skani. *Nat. Methods* **20**, 1661–1665 (2023).
142. LeCo Project. LegioSpecies - an updated and referenced list of Legionellaceae species. https://doi.org/10.5281/ZENODO.11072744 (2024).
143. Schwengers, O. *et al.* Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification: Find out more about Bakta, the motivation, challenges and applications, here. *Microb. Genomics* **7**, (2021).

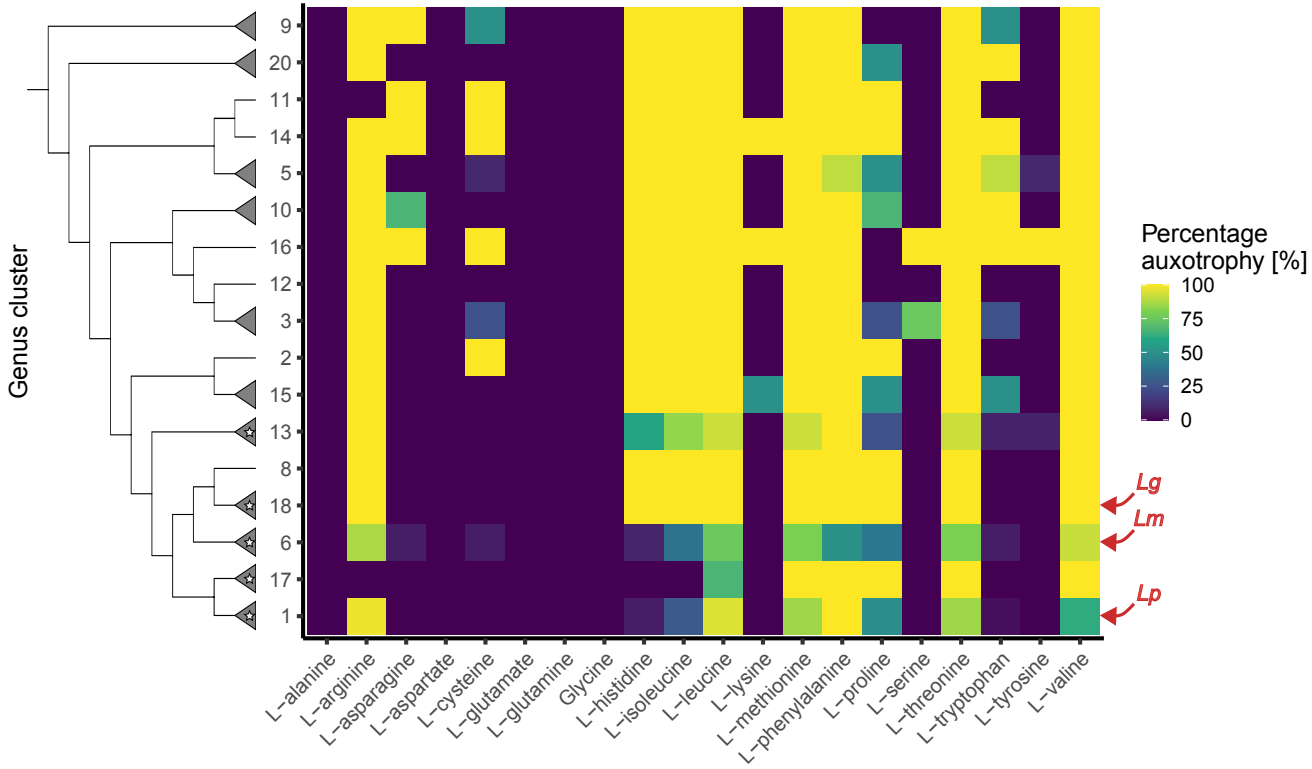
144. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
145. Konstantinidis, K. *et al.* FastAAI: Efficient Estimation of Genome Average Amino Acid Identity and Phylum-level relationships using Tetramers of Universal Proteins. Preprint at https://doi.org/10.21203/rs.3.rs-1459378/v1 (2022).
146. Cury, J., Abby, S. S., Doppelt-Azeroual, O., Néron, B. & Rocha, E. P. C. Identifying Conjugative Plasmids and Integrative Conjugative Elements with CONJscan. in *Horizontal Gene Transfer* (ed. De La Cruz, F.) vol. 2075 265–283 (Springer US, New York, NY, 2020).
147. Abby, S. S. *et al.* Identification of protein secretion systems in bacterial genomes. *Sci. Rep.* **6**, 23080 (2016).
148. Schwengers, O. *et al.* Platon: identification and characterization of bacterial plasmid contigs in short-read draft assemblies exploiting protein sequence-based replicon distribution scores. *Microb. Genomics* **6**, (2020).
149. Gumerov, V. M., Ulrich, L. E. & Zhulin, I. B. MiST 4.0: a new release of the microbial signal transduction database, now with a metagenomic component. *Nucleic Acids Res.* **52**, D647–D653 (2024).
150. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
151. Shaffer, M. *et al.* DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Res.* **48**, 8883–8900 (2020).
152. Zheng, J. *et al.* dbCAN3: automated carbohydrate-active enzyme and substrate annotation. *Nucleic Acids Res.* **51**, W115–W121 (2023).
153. Rawlings, N. D. *et al.* The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Res.* **46**, D624–D632 (2018).
154. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
155. Teufel, F. *et al.* SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat. Biotechnol.* **40**, 1023–1025 (2022).
156. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
157. Zhu, Q., Kosoy, M. & Dittmar, K. HGTector: an automated method facilitating genome-wide discovery of putative horizontal gene transfers. *BMC Genomics* **15**, 717 (2014).
158. Parks, D. H. *et al.* GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* **50**, D785–D794 (2022).
159. Gautreau, G. *et al.* PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph. *PLOS Comput. Biol.* **16**, e1007732 (2020).
160. Vences, M. *et al.* iTaxoTools 0.1: Kickstarting a specimen-based software toolkit for taxonomists. *Megataxa* **6**, (2021).
161. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
162. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
163. 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 (2017).

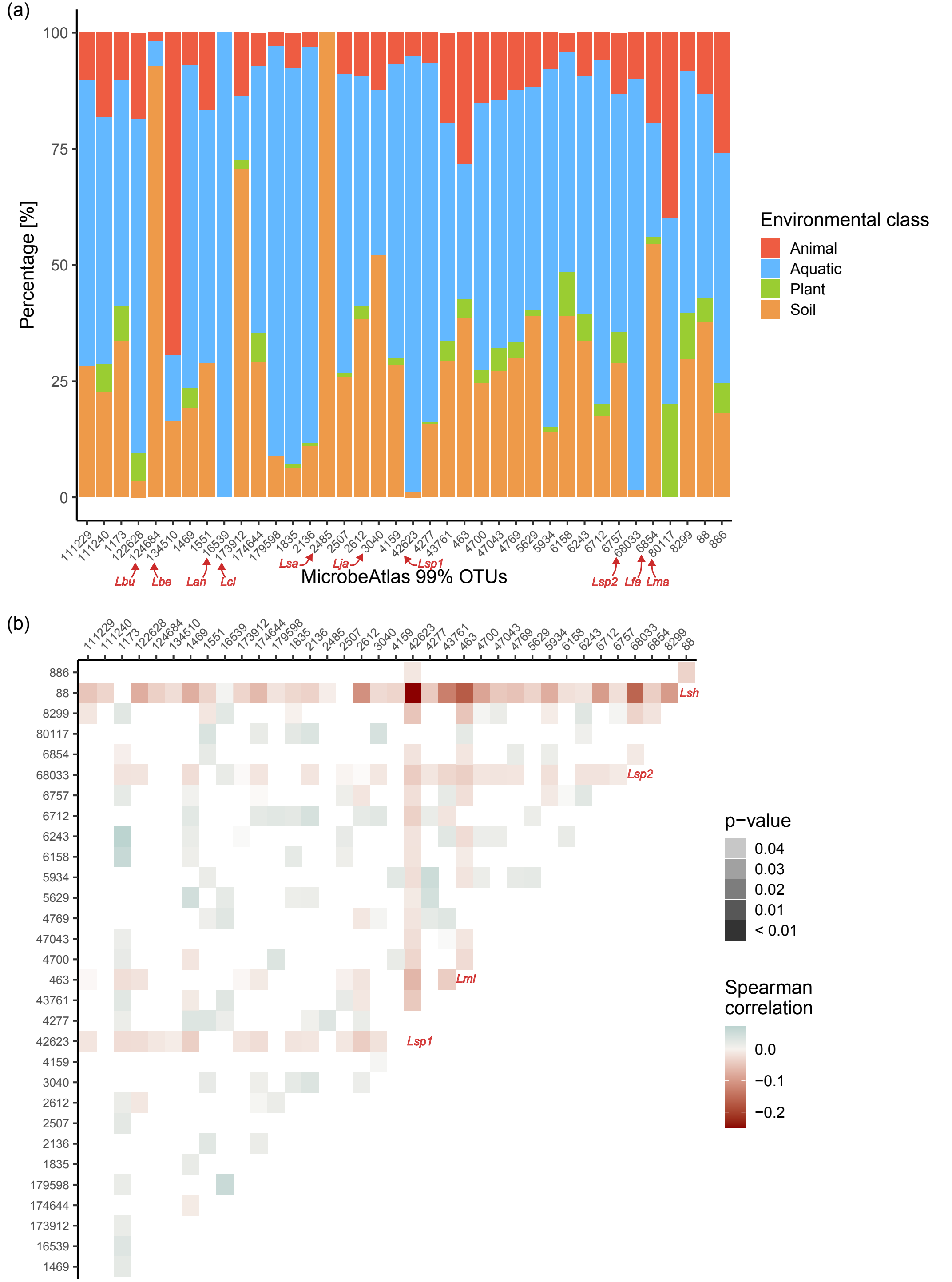
164. Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).
165. Letunic, I. & Bork, P. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Res.* **52**, W78–W82 (2024).
166. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, Vienna, Austria, 2023).
167. Wickham, H. *et al.* Welcome to the Tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).
168. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289–290 (2004).
169. Hahsler, M., Piekenbrock, M. & Doran, D. **dbscan** : Fast Density-Based Clustering with *R*. *J. Stat. Softw.* **91**, (2019).
170. Oksanen, J. *et al. Vegan: Community Ecology Package*. (2022).


1000
0
5
0
Number of unique gene families
Species
Genomovar
Estimated genome size [Mb]
Genus
Source
Unknown
Soil
Marine
Freshwater
Engineered environment
Environmental unknown
Clinical
Species information
Most clinically relevant
Clinically-associated
With scientific name
Not named
Isolated genome
Species cluster
Tree scale: 1
L. pneumophila


Species information
Genus
Dot/Icm similarity [%]
T2SS similarity [%]
Secretion systems
T1SS
T4SS
T5SS
Lss
Lb1ss
Lb2ss
Lrpss
Other
Type F
Type G
Type I
Type T
Type Va
Type Vc
T6SS
Motility
Flagellum
Pilus
Lcl adhesin
Chemo-sensing
Flagellar
ACF
Type IV pilus
Capsules
Syn-CPS3
ABC
Wzx/Wzy
Metabolism
TCA
EMP
ED
PP
Glyoxylate
PHB synthesis
myo-inositol degr.
Species cluster
Species information:
Most clinically relevant
Clinically-associated
With scientific name
Not named
Isolated genome
Dot/Icm similarity to Lp [%]
50 60 70 80 90 99
T2SS similarity to Lp [%]
46 57 68 78 89 99
Pathway completeness [%]
0 20 40 60 80 100
Tree scale: 1


(a)
PC2 (10.8%)
PC1 (19.9%)
1
0
−1
−1
0
1
2
Genus cluster
1
6
13
17
18
(b)
Intracellular
Secreted
CAZymes
Peptidases
Enzymes per genome
100
30
10
Genus cluster
1
6
13
17
18
Species
Most clinically relevant
Clinically−associated
With scientific name
Not named


Genus cluster
9
20
11
14
5
10
16
12
3
2
15
13
8
18
6
17
1
L−alanine
L−arginine
L−asparagine
L−aspartate
L−cysteine
L−glutamate
L−glutamine
Glycine
L−histidine
L−isoleucine
L−leucine
L−lysine
L−methionine
L−phenylalanine
L−proline
L−serine
L−threonine
L−tryptophan
L−tyrosine
L−valine
Percentage auxotrophy [%]
100
75
50
25
0
Lg
Lm
Lp


(a)
Percentage [%]
100
75
50
25
0
111229
111240
1173
122628
124684
134510
1469
1551
16539
173912
174644
179598
1835
2136
2485
2507
2612
3040
4159
42623
4277
43761
463
4700
47043
4769
5629
5934
6158
6243
6712
6757
68033
6854
80117
8299
88
886
Lbu
Lbe
Lan
Lcl
Lsa
Lja
Lsp1
Lsp2
Lfa
Lma
MicrobeAtlas 99% OTUs
Environmental class
Animal
Aquatic
Plant
Soil
(b)
Lsh
Lsp2
Lmi
Lsp1
p−value
0.04
0.03
0.02
0.01
< 0.01
Spearman correlation
0.0
−0.1
−0.2