Expansion, restructuring and characterization of the *Legionellaceae* family

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

[1] Department of Environmental Microbiology, Eawag, 8600 Dübendorf, Switzerland

*Corresponding author.

Name: Marco Gabrielli

Tel: +41 58 765 5960

Email: marco.gabrielli@eawag.ch

Keywords: *Legionellaceae*; diversity; host-invasion; metabolism; environmental distribution.

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 analyzed all publicly-available *Legionellaceae* genomes and metagenome-assembled genomes, nearly doubling the number of recognized species and highlighting the need to restructure the family's taxonomy to include multiple genera that present distinct clinical relevance. We found that secretion systems and traits linked to host invasion vary significant across genus clusters and species, likely influencing their survival in their environment. While genus clusters presented different metabolic capabilities, species closely related to *L. pneumophila* were characterized by a more complete metabolism with fewer auxotrophies. Finally, we showed that while species have distinct environmental distributions, niche overlap increases at higher metabolic similarity. Together, our results shed light on the ecology of these microorganisms, highlight the diversity of traits that can occur in closely related facultative intracellular bacteria, and call for the restructuring of this clade to align with genomic information and clinical relevance to aid the management of *Legionellaceae* bacteria.

Significance Statement

Legionellaceae bacteria are responsible for Legionnaires' disease, the leading waterborne disease in many industrialized countries. Historically, attention was focused on *Legionella pneumophila* and the ecology of other species remained largely unknown and unexplored. However, several water safety regulations cover the entire *Legionella* genus, currently the only recognized within this large family, including microorganisms with unknown health relevance. Using a comparative genomics approach and leveraging public genomes and amplicon data, we shed light on the scale, diversity and ecology of this family, finding a surprisingly diversity of the traits linked to host invasion and metabolism, likely driving the species' environmental distributions. This study proposes to

56 restructure the family *Legionellaceae* in multiple genera to align with known clinical associations
57 and predicted phenotypes.
58
59 **Main Text**
60 **Introduction**
61 Microbial ecology largely relies on conventional short-read 16S rRNA sequencing and specific
62 model organisms to infer microorganisms' environmental distribution and phenotypic
63 characteristics. While this combination has enabled important insights into the ecology of several
64 microorganisms and environments(1, 2), the limits imposed by amplicon sequencing taxonomic
65 resolution(3) and the restricted number of model organisms available(4) hamper our understanding
66 of the ecological differences at lower taxonomic levels (i.e., species, strains)(5, 6). This results in a
67 blurred picture of the ecology of several species, including those of clinical relevance. When only
68 some of the species within a clade are of interest, the average response of the clade may contrast
69 with the response of the individual species, as in Simpson's paradox(7). As a consequence, the
70 data obtained would not properly inform legislation and management practices.
71 A relevant example of this problem lies in the family *Legionellaceae* and its only recognised genus,
72 *Legionella,* a clade of gram-negative facultative intracellular bacteria. Multiple *Legionella* species
73 can cause legionellosis, a disease responsible for a significant number of deaths, productivity
74 losses and medical expenses exceeding 800 million USD per year in the USA alone(8) and for
75 which the reported incidence is rising worldwide(9). While 73 species are currently recognized(10),
76 *L. pneumophila* is alone responsible for more than 90% of infections in Europe and the USA(11).
77 This contrast between the large number of species and the clinical importance of *L. pneumophila*
78 results in different legislative actions, with countries differing both regarding the regulatory target
79 (e.g., only *L. pneumophila*, the entire genus, a combination of the two)(12, 13), as well as the
80 regulated exposure source (e.g., showers, drinking water, cooling towers)(14). While the choice of
81 a broader regulatory target follows the precautionary principle, an overly broad target can conflate
82 species with different phenotypes and hamper identification of truly critical situations. This is likely

83 the case for *Legionellaceae*, given that most scientific studies focus on *L. pneumophila* and the
84 knowledge regarding other species lags far behind. In fact, most of the information available on
85 non-*L. pneumophila* species has been obtained from comparative genomic approaches. These
86 studies uncovered significant heterogeneity in terms of genome organization(15), encoded
87 effectors(16) and secondary metabolites(17). However, information on the factors which affect their
88 interactions with the environment, such as motility, secretion systems and metabolism, is still
89 lacking. Moreover, the aforementioned studies focus only on cultured isolates, while recent
90 molecular-based ecological surveys have uncovered a significant uncultured diversity of
91 *Legionellaceae* bacteria(18), a fact that highlights the need to include uncultivated species to
92 properly discern the evolutionary paths and the derived ecological differences. Understanding the
93 ecological differences among different *Legionellaceae* species has direct impact on the
94 management and legislation of these species(12).
95 Here we retrieved over 5000 publicly available *Legionellaceae* genomes and metagenome-
96 assembled genomes and compared the functional potential of the different species included in this
97 family, focusing specifically on evaluating its genetic diversity and different traits important for
98 facultative intracellular bacteria, including secretion systems, motility and host invasion, and
99 metabolism. We complemented this analysis with a genome-informed mining of 16S rRNA surveys,
100 assessing species' distribution across environments. Collectively, this work provides key insights
101 regarding the diversity and the ecology of *Legionellaceae* bacteria which will allow refining
102 legislation and management strategies towards the most clinically relevant species and safeguard
103 public health.
104

105 **Results and discussion**

106 **Analysis of public genomes substantially restructures *Legionellaceae* family**

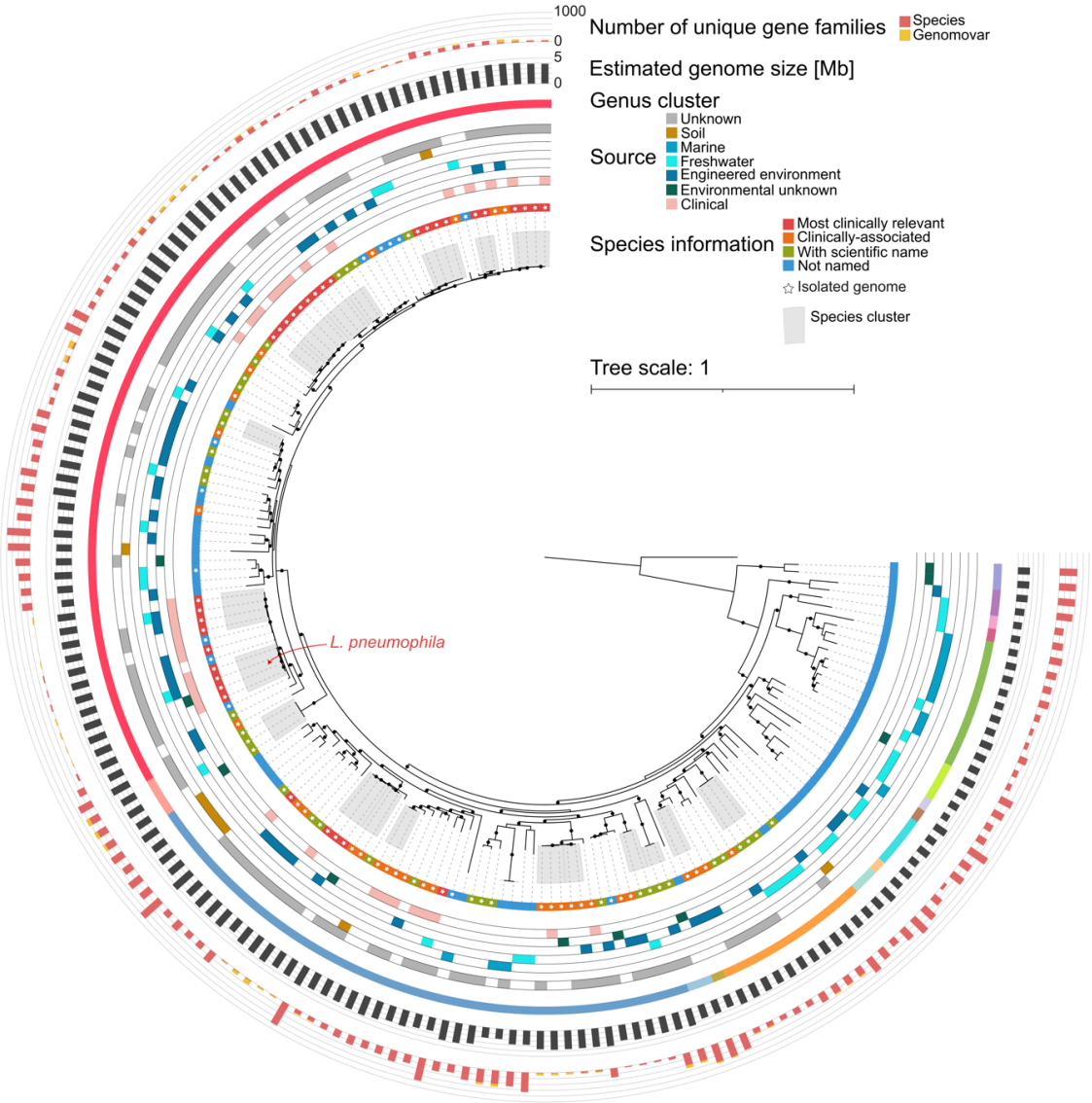
107 Given the known difficulties in culturing bacteria(19), culture-independent data is necessary to
108 characterize the full range of microbial diversity. We delineated 128 species clusters (Table S1,
109 see Data availability section) from publicly available *Legionellaceae* isolated and metagenomic-

110 assembled genomes (MAGs) using an average nucleotide identity (ANI) of 95%(20). Importantly,
111 each ANI-delineated species cluster included genomes that presented the same species label in
112 their metadata, or were not annotated to the species-level. The newly delineated species clusters
113 substantially expand the *Legionellaceae* family, nearly doubling the number of the species currently
114 recognised(21, 22) and resulting in several novel phylogenetically clades (Figure 1). At the same
115 time, we identified several new species clusters, phylogenetically related to known ones. As an
116 example, ANI similarity supports the differentiation of *L. pneumophila* subspecies *subsp. pascullei*
117 and *subsp. fraseri* into two distinct species (i.e., *L. pascullei* and *L. fraseri*) based on an average
118 ANI similarity to *L. pneumophila* of, respectively, 91.2% and 92%. Notably, given that most of the
119 unique genes are of unknown function(23), the phenotypic differences linked with this differentiation
120 are unknown. Also, the retrieved genomes indicate the presence of three other species clusters
121 previously classified as *L. pneumophila,* which represent instead distinct species (ANI compared
122 to *L. pneumophila*: 91.0% - 91.8%). Within each species, we found between 1 and 213 genomovars
123 (genomic variants, ANI threshold: 99.5%). While the mechanisms of genetic variation and selective
124 pressure can have different influences on the intra-species diversity of different species(6), the ratio
125 between the number of genomovars and the number of genomes in each species suggests that
126 the intra-species diversity of non-*L. pneumophila* species is likely to still be underestimated (Table
127 S2).
128 Recognizing the lack of a consensus definition, a genus is generically considered a restricted
129 monophylic clade of organisms with conserved characteristics, which arise due to significant
130 genome similarity(24). The family *Legionellaceae* currently includes only a single genus, i.e.
131 *Legionella*(21, 22). However, a comparison of the average amino acid identity (AAI) returned values
132 reaching as low as 45%, a value well below the suggested boundary for genera occurring between
133 65 and 72% AAI(25). We further investigated relative evolutionary distance values(26) in
134 combination with manual core genome phylogeny refinement to define genus clusters and remove
135 polyphyly. This resulted in defining 17 different genus clusters present within the *Legionellaceae*
136 family, of which only five include cultured isolates. This is substantially different from what is

137 currently accepted, and highlights the need to re-structure the taxonomy of this family. Interestingly,
138 most clinically relevant species clusters(27) belong to the same clade as *L. pneumophila.* However,
139 clinically-associated species are present in three of the five genus clusters with isolated
140 representatives. Interestingly, two of the most clinically relevant species, *L. micdadei*(28) and *L.*
141 *feeleii*(29), were found to belong to a separate genus cluster than *L. pneumophila*. While such
142 distinction was previously suggested(30), it was later criticized based on 16S rRNA analyses(31),
143 a method now known to provide limited phylogenetic resolution(32). These three genus clusters
144 with clinically-associated species encompass phylogenetically all clusters with previously known
145 species suggesting that all isolated species could potentially pose a risk to immunocompromised
146 subjects(28). On the other hand, the absence of clinical isolates from a large clade of the
147 *Legionellaceae* tree suggests that the current definition of the *Legionella* genus (i.e., coinciding to
148 the family one) includes distinct phenotypes, of which there is a clear distinction in clinical
149 relevance. As a result, in case legislation is applied at the genus level, the risk of *L. pneumophila*
150 and species with no indication of clinical relevance is equated, not allowing legislators/operators to
151 prioritize remediation interventions.
152 We found the genus clusters to be characterized by different genome sizes, even when accounting
153 for differences caused by genome origin (i.e., isolates, MAGs) (two-way ANOVA, p-vals < 0.001).
154 The combination of genome size variability, together with the proteome diversity across the family,
155 results in a core pangenome (i.e., present in ≥95% of genomes) of only 697 gene families out of
156 64796 (1.1%). To put this finding in perspective: even though a detailed comparison is not possible
157 due to methodological differences, a similar percentage of core genes was found in the phylum
158 *Chlamydiae*(33), a fact that emphasizes the high variability of the genome content of the
159 *Legionellaceae* family. In fact, motile facultative intracellular bacteria, such as *Legionellaceae*,
160 present greater pangenome sizes than non-motile strictly intracellular bacteria(34), such as
161 *Chlamydiae*. The size of the core and total pangenome for the entire *Legionellaceae* family are,
162 respectively, -41.7% and +34.7% different than what is estimated through the Heaps' Law(35)
163 curves obtained from only previously-known species (core families: 1081, total families: 37071).

164 This highlights the diversity of the newly delineated species compared to previously known ones.
165 In addition, core pangenome genes make up less than half of total number of genes per genome
166 (i.e., isolates: 17% - 32%; MAGs: 21% - 46%) with between 2 and 834 gene families present
167 exclusively in a single species and up to 161 additional unique genes present in genomovars. The
168 number of unique gene families is best explained by the combination of genome size and relative
169 phylogenetic distance (i.e., the phylogenetic distance to the closest other *Legionellaceae* genome)
170 (R^2 = 0.78, Figure S1). Yet, most of the explanatory power is due to the relative phylogenetic
171 distance rather than genome size (ΔAIC values compared to the full model equal to 6.4 and 227.3,
172 respectively), highlighting that genome size increase was accomplished in *Legionellaceae* bacteria
173 through different evolutionary trajectories which led to a diversification of proteome content.

174

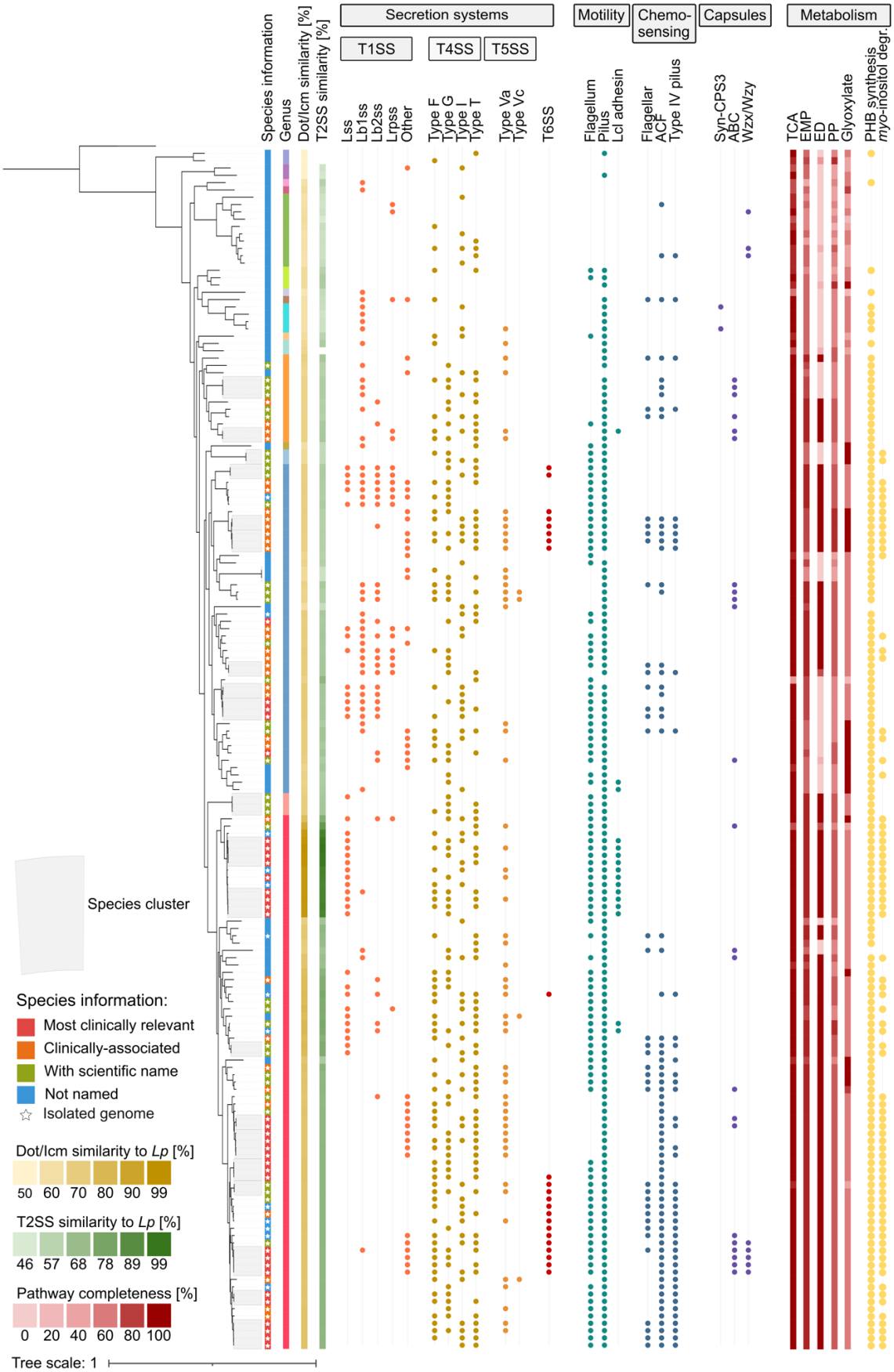


175
176 **Figure 1.** Maximum likelihood phylogenetic tree of the family *Legionellaceae* based on 80% identity
177 of core genes families which presented also homologs within the *Coxiella burnetii*, the chosen
178 outgroup (54 protein families, 19141 sites, model LG+F+I+R7). Nodes with optimized ultrafast
179 support ≥85% are marked with black circles. Leaves enclosed within grey areas indicate genomes
180 of different genomovars belonging to the same species. Outer tree layers indicate whether the
181 delineated species have currently valid scientific names and genomes from isolated specimens are
182 available, the source of all the genomes belonging to the specific species/genomovar, genus cluster

183 membership, the estimated genome size and the number of unique gene families present
184 throughout a species or unique to specific genomovars.
185
186 **Diversity in *Legionellaceae* family highlights differences in putative phenotypes**
187 The extent of the proteome diversity across different genus clusters suggests the presence of
188 distinct phenotypes across this family. Given the importance of interactions facultative intracellular
189 bacteria with the host and the importance of metabolism bacterial niche definition, we assessed
190 the differences of secretion systems, factors linked to host invasion and the metabolic potential
191 across different species and genus clusters.
192
193 **Secretion systems presence is highly variable within the *Legionellaceae* family**
194 The ability of secretion systems to translocate effector proteins to host cells is a key trait of
195 intracellular bacteria, which allows them to manipulate hosts to their advantage(36). *L.*
196 *pneumophila* is known to secrete a variety of effectors through the Dot/Icm Type IVB (T4BSS) and
197 Type II (T2SS) secretion systems(16, 37, 38). We found homologs of proteins belonging to the
198 T4BSS of *L. pneumophila* in all genomovars, confirming its presence in the novel genus clusters.
199 Besides the T4BSS secretion system, the T2SS was detected in nearly all *Legionellaceae* species,
200 confirming previous analyses of its wide distribution(39). Notably, the presence of a T2SS was not
201 confirmed in two MAGs forming their own genus cluster. While this result could be influenced by
202 genome incompleteness (estimates: 11% and 1%), the observed pattern suggests the occurrence
203 of gene loss events by the clade's ancestor. Given that the T2SS favours survival of *L. pneumophila*
204 at temperatures below 17 °C(40), it is possible that species lacking a T2SS might not be able to
205 survive for long periods of time low temperatures. Interestingly, two of these MAGs were
206 retrieved from environments with warm temperatures, specifically soils incubated at 25 °C(41)
207 and mine drainage(42).

208 Despite the potential role of other secretion systems in *Legionellaceae* ecology (e.g., host range,
209 motility, HGT), little is known regarding their prevalence and diversity(43). The results of our
210 analysis highlight the presence of a wide range of secretion systems spread throughout the family
211 (Figure 2, Table S3). The different secretion systems have distinct distribution across genus
212 clusters. We detected Type IVA secretion systems (T4ASS) of various types (i.e., Types F, G, I and
213 T) throughout the family and found only a limited influence of phylogenetic relatedness on their
214 presence/absence pattern (Jaccard diversity-based Mantel correlation, $r = 0.19$, p-val < 0.001). In
215 fact, we determined 78% and 43% of, respectively, Type F and T systems to be carried by plasmids,
216 supporting their pervasive circulation across the family(16, 44). Conversely, we found other
217 secretion systems to be present predominantly in specific genus clusters. We found the lss and
218 novel Type I (T1SS) and Type V (T5SS) secretion system predominantly in the genus clusters that
219 includes *L. pneumophila*, *L. micdadei* and *L. israelensis*. Such machineries could be potentially
220 favour host cell adhesion and entry(45), as well as biofilm formation(46). In addition, we identified
221 Type VI secretion systems (T6SS) within restricted clades of such genus clusters. T6SSs could
222 potentially be used against either competing prokaryotes or non-infectable eukaryotes(47) or to
223 obtain nutrients from neighbouring cells(48), potentially favouring extracellular survival. On the
224 other hand, we noticed different types of T1SSs (i.e., lb1ss, lb2ss and lrpss)(49) to be largely
225 lacking in the *L. pneumophila* genus cluster, while being present in the ones of *L. micdadei*, *L.*
226 *oakridgensis* and the ones composed exclusively of MAGs.
227 While novel clades with no history of clinical association seem to be characterized by a lower
228 number of secretion systems, a direct connection between the reduced presence of secretion
229 systems and pathogenicity is unlikely. In fact, previous studies have shown a limited concordance
230 between *Legionellaceae* pathogenicity and secretion systems(50, 51). Such limited concordance

231 is due to several causes. First of all, pathogenicity, as well as secretion systems, can vary within
232 species, requiring to analyse all available genomes and not select only few representatives.
233 Furthermore, intracellular replication is connected to the release of adequate effectors within a
234 host(52). Given the known heterogeneity of effectors even within single species (e.g., (38, 53)),
235 encoding genes for a secretion system is not a sufficient proxy for released effectors. Finally,
236 human infection is also dependent on the environmental distribution of these bacteria, as well as
237 their ability to survive aerosolization and reach human lungs, factors that are influenced by other
238 cellular traits (e.g., metabolism, UV resistance).


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

240 **Figure 2.** Maximum likelihood phylogenetic tree of the family *Legionellaceae* based on the
241 identified 80% identity core genes families, which presented also homologs within the *C. burnetii*,
242 the chosen outgroup (54 protein families, 19141 sites, model LG+F+I+R7). Leaves enclosed within
243 grey areas indicate genomes of different genomovars belonging to the same species. The average
244 Dot/Icm and T2SS protein similarity was estimated against *L. pneumophila* strain *Paris*, while
245 rounded markers indicate the presence of the specified traits.
246
247 **Factors linked to host invasion are not conserved among species**
248 The likelihood of an intracellular bacterium encountering its host can be very low due the diversity
249 of microorganisms in the environment(54) and hosts feeding preference(55), suggesting the
250 importance "active" mechanisms of host invasion. We identified the genes required for Type IV pili
251 in throughout the family with the exception of a clade composed of three novel genus clusters
252 (Figure 2, Table S3). The lack of pili in these genus clusters suggests their limited efficiency in host
253 colonization(56). On the contrary, we found flagellar assembly genes to be present mostly in the
254 genus clusters of *L. pneumophila*, *L. micdadei*, *L. israelensis* and *L. geestiana*, while mostly absent
255 in others. Compared to pili, flagellated species show a lower prevalence in such genus clusters,
256 confirming previous experimental observations (i.e., *L. longbeacheae*)(57) and suggesting the
257 occurrence of multiple loss events of this trait. The lack of flagella in *L. longbeacheae*, a species
258 often linked to soil and compost(27), is likely due to the fact that surface motility does not
259 necessarily require such machinery(54). In fact, non-flagellated species make up the majority
260 (62.5%) of the ones that were isolated from soil, suggesting their adaptation to non-aquatic
261 environments. Beyond motility, adhesins, such as the Lcl protein encoded by *L. pneumophila*(58),
262 are also key determinants of host-pathogen interactions and biofilm formation(59). While we found
263 Lcl homologs characterized by a collagen-like domain in the genus cluster of *L. pneumophila*, we
264 detected other proteins with collagen domains characterized by "unknown" or "membrane" COG
265 categories in 11 species mainly in the genus clusters of *L. oakridgensis*, *L. geestiana* and *L.*
266 *micdadei*, highlighting the presence of additional adhesins absent in the *L. pneumophila* genus

267 cluster (Table S4). Besides an easier attachment to host cells and surfaces, the enhanced biofilm
268 formation abilities provided by these proteins might also favour resistance to stresses and, thus,
269 persistence in the environment. Yet, characterizing the importance of such traits is complicated by
270 hosts diversity. For example, *L. pneumophila* was shown to require a flagellum to efficiently infect
271 human macrophages, but not amoebae(60).
272 Chemosensing, the ability to sense chemical gradients, can be used to direct movements (i.e.,
273 chemotaxis) towards hosts(54). However, we noticed that, except for restricted clades with the *L.*
274 *pneumophila* and *L. micdadei* genus clusters, the majority of *Legionellaceae* species lacked both
275 CheW and methyl-accepting chemosensing proteins, two of the essential components of the
276 chemotaxis Che system(61). This indicates the absence of motility regulation through the Che
277 pathway (Figure 2, Table S3). In addition, we found a weak association between the Che pathway
278 and flagella presence (binomial GLM, p-value = 0.02) and pili (p-value = 0.07), further stressing the
279 unrelatedness of these traits. Interestingly, previous research highlighted that most respiratory
280 pathogens are non-chemotactic, supposedly due to the relative simplicity of the respiratory
281 system(54),(59). However, this hypothesis does not take into consideration the need to survive in
282 the environment and find suitable hosts of facultative intracellular bacteria, such as *Legionellaceae*
283 bacteria. As a matter of fact, other regulatory pathways to modulate chemotaxis exist(62). For
284 example, twitching motility in *L. pneumophila* is inhibited by the presence of other bacteria likely
285 due to lower nutrients concentrations(63), indicating either that rich nutrient conditions are needed
286 for twitching motility or the ability to perform chemotaxis despite lacking the Che pathway.
287 We predicted up to 80 signal transduction and 170 response regulator proteins per genome.
288 Interestingly, we found a different number of these proteins across the isolated bacteria in different
289 genus clusters (ANOVA, Tukey HSD post-hoc test, p-values < 0.002; Figure S2), with the one
290 including *L. pneumophila* presenting the highest number. In total, we identified 726 and 1363 gene
291 families for signal transduction and response regulation, of which only 2 and 13 were, respectively,
292 detected across the whole family (≥95% species). In most cases, the function of these proteins was
293 already reported for *L. pneumophila* (Table S5). Beyond those, we detected 71.6% and 66.9% of

294 signal transduction and response regulators to be genus cluster-specific, but that only 35.6% and
295 33.4% of them are encoded by multiple species within their cluster. These results suggests that
296 while a “core” regulatory network might be conserved across species, significant diversity is present
297 not only across, but also within genus clusters, which might significantly affect phenotypic
298 responses to environmental cues and contribute defining the environmental niche of different
299 species.
300 Finally, also extracellular capsules can affect bacteria-host interactions and enhance survival in
301 unfavourable environments(64). A capsule has been only reported for *L. longebeacheae* where it
302 constitutes a crucial virulence factor towards *Acanthamoeba castellani*(65) and potentially aids its
303 survival in soil(64). Despite not presenting notable homology to *L. longbeacheae* capsule proteins,
304 we detected capsular biosynthetic gene clusters of different types across the family, and especially
305 in the genus cluster of *L. oakridgensis*. *L. beliardensis*, *L. gresiliensis* and *L. busaniensis* presented
306 two ABC capsules systems, while *L. gormanii* and *L. qingyii* encoded two different capsule types,
307 a trait which might favour these bacteria against different stresses(65) and extend their niche
308 breadth(66). Interestingly, the similarity of the capsule systems varied widely within species (Figure
309 S3), possibly arising from the adaptation to different environmental conditions or lifestyles of distinct
310 genomovars(64).
311
312 **Genus clusters encode different metabolic potentials**
313 Several laboratory studies have highlighted the presence of a complex biphasic metabolism in *L.*
314 *pneumophila*(67). However, the metabolism of other species has been seldomly examined, leading
315 to a major knowledge gap regarding the ecology of these species. We found a complete
316 tricarboxylic acid (TCA) cycle in nearly all species, highlighting the reliance of the whole family on
317 this pathway to convert serine into acetyl-CoA(68) (Figure 2, Table S6). On the other hand, we
318 found other metabolic pathways to be not as conserved. As expected, we observed species
319 belonging to the genus cluster of *L. pneumophila* to present a similar metabolism to this model
320 species, with a complete Entner-Doudoroff (ED) and Embden-Meyerhof-Parnas (EMP), and an

321 incomplete pentose phosphate (PP) pathway to provide the pyruvate required by the TCA
322 cycle(69),(68),(70). However, we detected complete ED pathways only in subclades of the *L.*
323 *micdadei* and *L. oakridgensis* genus clusters and in none of the ones composed only of novel
324 species. Given that this pathway was either complete or completely absent, it is unlikely that this
325 result arises due to MAGs incompleteness, but is rather a true adaptation signal. Interestingly, we
326 found that both the EMP and the PP pathways presented higher completeness when the ED
327 pathway was present (t-test, p-val < 0.001), suggesting a reciprocal dependency of such pathways
328 within the whole family and not only *L. pneumophila*(70). As a result of this dependency, the
329 completeness of both these pathways varied significantly across genus clusters (Kruskal test, p-
330 val < 0.001) and presented a negative correlation with the phylogenetic distance to *L. pneumophila*
331 (Sperman correlation; ρ = -0.48 – -0.7, p-val < 0.001), suggesting a less prevalent role of these
332 pathways in phylogenetically unrelated species to *L. pneumophila* and their reliance on different
333 metabolic pathways. For example, while the glyoxylate cycle is highly incomplete in *L.*
334 *pneumophila*(68), we observed its complete set of genes in several species which mostly belong
335 to other genus clusters. This pathway could potentially contribute to the biosynthesis of cell carbon
336 compounds, but future studies are needed to verify this hypothesis.
337 The diversity of metabolism among *Legionellaceae* genus clusters is further confirmed by the fact
338 that, even excluding MAGs to remove genome incompleteness bias(71), we observed significant
339 differences in the KEGG metabolic pathways completeness profiles among genus clusters (Figure
340 3a; PERMANOVA, R^2 = 24.6%, p-val < 0.001) due to the different presence of several enzymes
341 across genus clusters, as shown in Figure S4. Most notably, the genus clusters including *L.*
342 *pneumophila* and *L. micdadei* show a larger number of unique enzymes indicating their richer, but
343 different, metabolic capabilities. In fact, we found more complete pathways for phosphatidylcholine
344 biosynthesis and acylglycerol degradation in the genus cluster of *L. pneumophila*, while higher
345 completeness of the glycogen biosynthesis pathway in the one of *L. micdadei* (Table S7).
346 Concordant with the metabolic differences observed at genus cluster-level, we detected a negative
347 correlation between metabolic differences and phylogenetic relatedness (Spearman correlation; ρ

348 = -0.5, p-val < 0.001) resulting in smaller subclusters of highly related genomes with similar
349 metabolic capabilities (Figure S5, Table S8). Several of such metabolic clusters included both
350 species clinically-associated and not, suggesting that species with different associations to disease
351 can have overlapping environmental niches.
352 The greater metabolic capabilities of the species belonging to selected *Legionellaceae* genus
353 clusters extends not only to central metabolic pathways, but involve also secondary metabolites
354 such as CAZYmes and peptidases, as well as other phenotypes, such as the biosynthesis of the
355 storage compound poly-3-hydroxybutyrate (PHB) and *myo*-inositol metabolism, which have been
356 shown to influence *L. pneumophila* growth(72),(73, 74). We observed statistically significant
357 differences in the number of genes encoding for both secreted and intracellular CAZymes and
358 peptidases, with the genus clusters of *L. pneumophila* and *L. micdadei* presenting a larger number
359 of enzymes as other genus clusters (Figure 3b). In addition, we identified that both the number
360 (Figure S6) and the repertoire of CAZYme families and peptidases classes is explained by genus
361 cluster membership (PERMANOVA, R^2 = 9.2 - 17.6%, p-val < 0.001), highlighting the difference in
362 the complex organic matter degradable among clusters(75, 76). Similarly, other phenotypes
363 reported for *L. pneumophila*, such as the ability to synthetize PHB and metabolize *myo*-inositol,
364 tended to be rarer in genus clusters phylogenetically distant from the *L. pneumophila* one (Figure
365 2), once again stressing the reduced metabolic capabilities of novel genus clusters and suggesting
366 substantial difference in their mechanisms of intracellular replication and survival to potentially
367 nutrient-poor extracellular conditions(77),(74).
368 Overall, the more restricted metabolic potential of genus clusters composed of novel species, even
369 though potentially affected by MAGs incompleteness, suggests their higher metabolic dependence
370 on other microorganisms. While it is possible that the required nutrients could be retrieved freely in
371 the environment due to the leakage from other microorganisms(78), it is also a possible signal of a
372 stronger metabolic reliance on their hosts. This could explain the difficulties in their isolation and
373 warrants further experimental investigations.
374

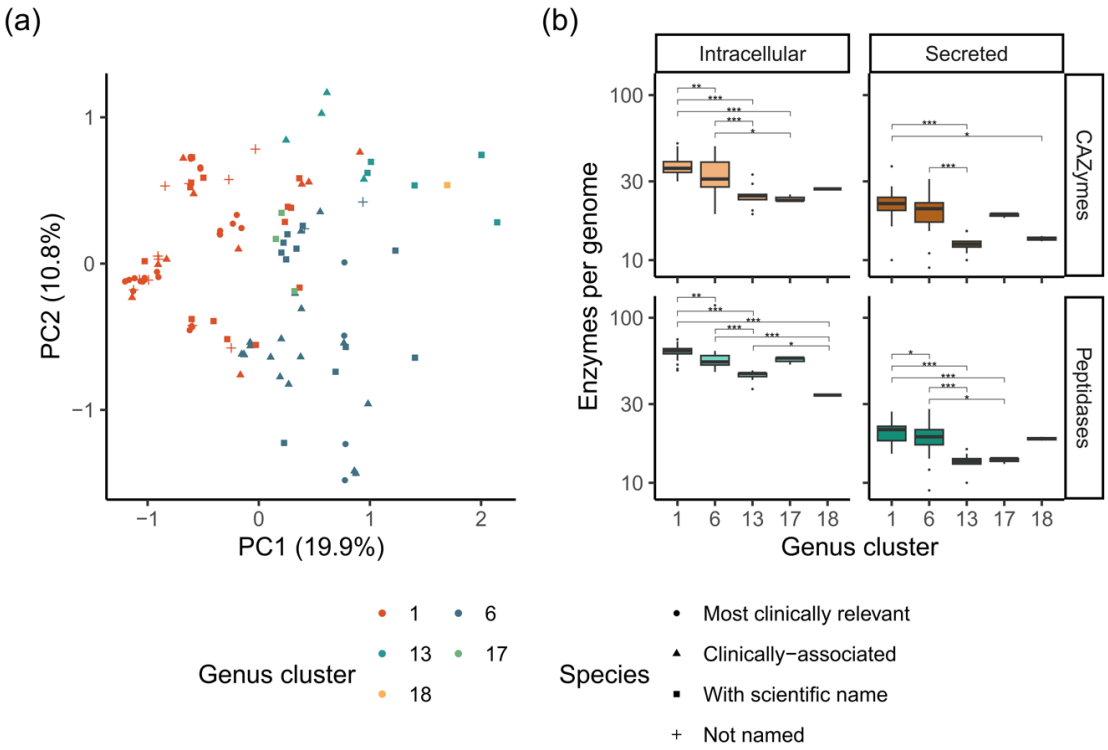


Figure 3. (a) Metabolic differences among *Legionellaceae* isolated genomes and clustering following genus cluster membership. (b) Number of CAZYmes and peptidades per isolated genomes as a function of their secretion and genus cluster membership. Statistically significant differences among genus clusters are marked by their p-values: ”*” = 0.05 - 0.01, ”**” = 0.01 - 0.001, ”***” < 0.001. Genus clusters representatives are: “1”: *L. pneumophila*, “6”: *L. micdadei*, “13”: *L. oakridgensis*, “17”: *L. israelensis*, “18”: *L. geestiana*.

Auxotrophy predictions suggest redesign of cultivation media

Intracellular bacteria have evolved auxotrophy for diverse amino acids(79), gaining a fitness advantage in their host-associated niche, but becoming dependent on external resources for survival(80). This also holds true for *L. pneumophila*, for which several studies have confirmed its auxotrophy for several amino acids (Table 1). We performed *in silico* analysis of *L. pneumophila* auxotrophies using GapMind, a pathway annotation tool(81), and gapseq, an automated tool for genome-scale metabolic model reconstruction(82). Except for L-isoleucine and L-serine, the

390 predictions of the two tools were mostly concordant and matched, respectively, eight and six of the
391 11 reported auxotrophies with limited indications that this might be attributable to strain-level
392 differences (Table S9). The different results obtained for L-isoleucine and L-serine by two
393 approaches lies in the fact that these compounds are predicted by gapseq to be synthesized
394 following two pathways not considered by GapMind (Table S10) and warrant more investigations
395 to verify their biological relevance. Interestingly, while L-cysteine is commonly used for culture
396 confirmation of *L. pneumophila*, both tools did not predict its auxotrophy. While we detected the
397 genes required for L-cysteine biosynthesis through the reverse transsulfuration pathway(83), we
398 noticed the lack of homocysteine biosynthesis, a required precursor for L-cysteine biosynthesis, as
399 well as the requirement for inorganic sulphur sources (e.g., SO_4^{2-}, H_2S). The latter, despite the
400 known link between L-cysteine biosynthesis and sulphur(84), is not controlled for in growth media
401 (e.g., BYEB). Such results indicate that L-cysteine might be an "indirect" auxotrophy and warrant
402 dedicated experimental confirmations to clarify *L. pneumophila* metabolic capabilities.
403
404 **Table 1.** Reported and predicted auxotrophies for *L. pneumophila*. Experimental evidence retrieved
405 from experimental studies(68, 143, 144). High and medium GapMind biosynthesis prediction
406 confidence results indicate the detection of homologs for all biosynthetic steps for a given amino
407 acid. Conversely, low confidence results highlight the lack of biosynthetic steps and imply
408 auxotrophy(81). Multiple results per amino acid indicate that distinct results were obtained with
409 different genomovars. The asterisk indicates that while the genes for the L-cysteine biosynthesis
410 have been annotated, the genes required to produce the required compound homocysteine are
411 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

412

413 Given the susceptibility of annotation-based tools to genome incompleteness and the tools' overall
414 concordance (Table S9), we continued the analysis only using gapseq's results. The gapfilling
415 carried out by this tool, in fact, removed the differences in the number of predicted auxotrophies
416 between isolates and MAGs from the same genus cluster (t-test, p-vals > 0.1). This allowed us to
417 highlight significant differences in the number of auxotrophies among genus clusters and the lower
418 amount in the ones of *L. pneumophila*, *L. micdadei* and *L. israelensis* (ANOVA, p-val < 0.001). We
419 did not find any amino acid auxotrophy to be conserved across the entire family, despite several
420 being highly prevalent in most genus clusters (Figure 4, Table S11). Conversely, no species
421 showed auxotrophy towards five amino acids (L-alanine, L-aspartate, L-glutamate, L-glutamine and
422 glycine). Interestingly, while we did not detect L-cysteine auxotrophy in the *L. pneumophila* genus
423 cluster, we found auxotrophic species towards this amino acid in the genus cluster of *L. micdadei*
424 and in several others without isolated representatives. Similarly, with the exception of the genus
425 cluster composed exclusively of *L. geestiana*, the genus clusters with isolated representatives
426 showed low prevalence of L-histidine auxotrophy. Such auxotrophic differences are likely partly
427 responsible for the differential growth ability of different *Legionellaceae* species on standard
428 medial(85), and call for media supplementation of additional amino acids such as L-histidine for the
429 isolation of novel species.
430

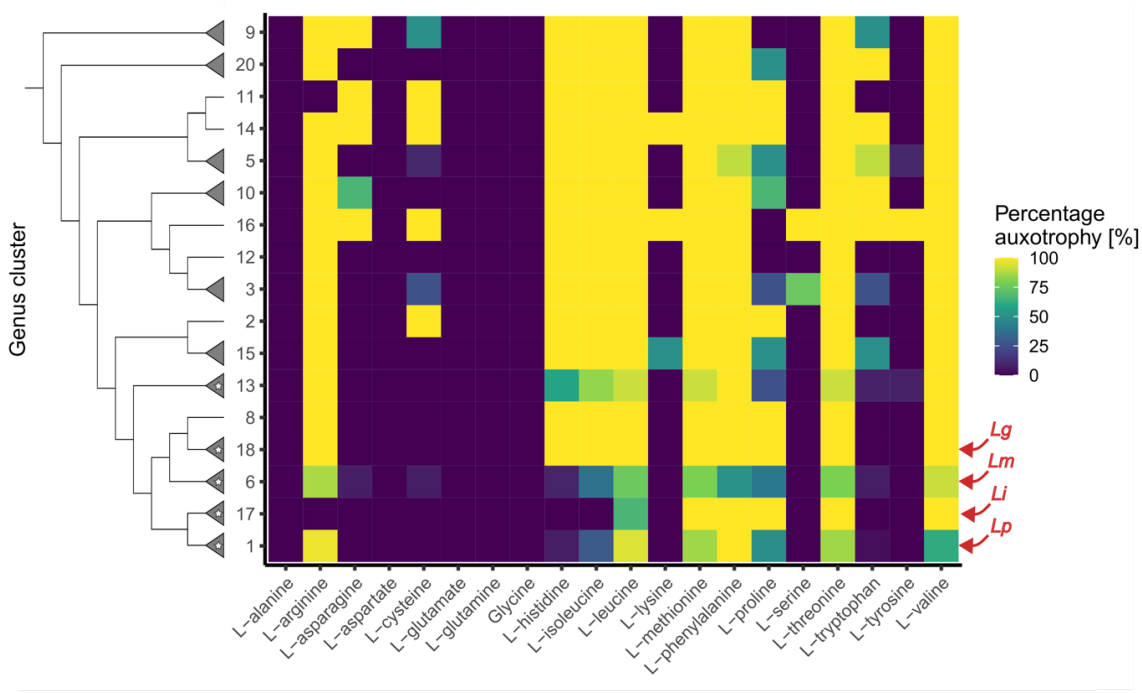


Figure 4. Prevalence of auxotrophies predicted in each genus cluster. The phylogenetic tree on the left represents the relationships among genus clusters, indicating with a star the ones for which isolated representatives are available. The labels "*Lg*", "Li", "*Lm*" and "*Lp*" represent the genus clusters including *L. geestiana*, *L. israelensis*, *L. micdadei* and *L. pneumophila*, respectively.

***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(86), linking them to the full-length 16S rRNA genes of 38 species clusters through reciprocal best hits (Table S12). We detected most *Legionellaceae* OTUs in all major environmental classes (Figure 5), confirming the variety of niches of these bacteria observable for the isolation and MAGs data (Figure 1). The distribution over the environmental classes varied greatly within genus clusters, with a surprisingly high fraction of animal-associated samples. A closer inspection (Figure S7) revealed several samples from non-human hosts, as well as the gastrointestinal tract or faeces, suggesting that these bacteria can

447 infect other animals and be shed by infected individuals, confirming previous reports(87, 88). While
448 in general we found OTUs mostly in aquatic samples, several OTUs showed a drastically higher
449 prevalence of soil samples or animal-associated ones (Figure 5). Comparing different genus
450 cluster, we did not observe a statistical difference in the occurrence of OTUs belonging to different
451 genus clusters (Kruskal-Wallis test, p-val = 0.76) and we detected only a weak statistical effect on
452 the distribution over environmental classes of genus cluster membership (PERMANOVA, R^2 =
453 7.6%, p-val = 0.07, only genus clusters with more than one OTU considered). These results
454 highlight that different genus clusters are similarly present in the environment and their species
455 present only a limited adaptation to specific major environmental classes, given also the
456 heterogeneity of these broad classifications. Yet, we found more similar environmental distributions
457 for OTUs belonging to the same metabolic cluster (Table S8) than with others (t-test, p-val = 0.04)
458 and at lower phylogenetic distance (Spearman correlation, ρ = -0.26, p-val < 0.001). This
459 corroborates the link between shared evolutionary history, metabolic capabilities and ecological
460 niches and remarking how these processes should be assessed at very fine taxonomic resolutions.
461 It is important to note that the environmental niche of facultative intracellular bacteria is not only
462 due the bacteria's direct survival in the environment, but also to the environmental survival of their
463 hosts(89), a fact which can extend the overall niche breath. Future research identifying the genes
464 responsible for the adaptation to diverse environments and understanding better the impact of
465 hosts in *Legionellaceae* environmental distribution could help in source tracking by providing
466 insights on the environmental origin of clinical specimens.
467

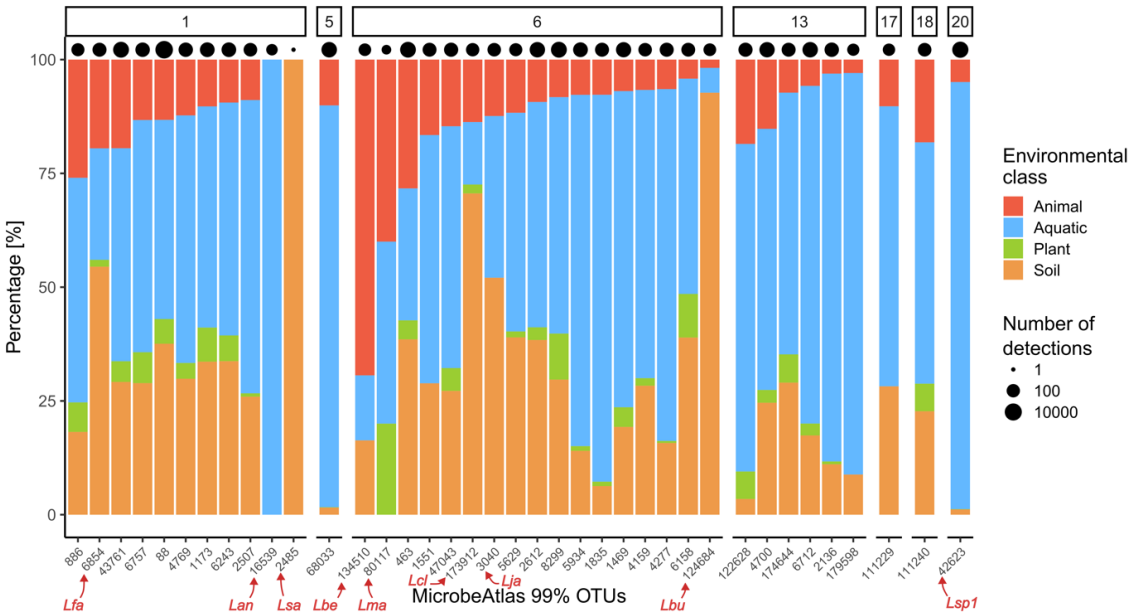


Figure 5. Number of samples in which the MicrobeAtlas OTU (99% similarity) were found (i.e., OTUs' occurrence) and their distribution among major environmental classes. Species labels are the following: *L. antarctica (Lan), L. beliardensis (Lbe), L. busanensis (Lbu), L. clemsoniensis (Lcl), L. fallonii (Lfa), L. jamestowniensis (Lja), L. maceachernii (Lma), saoudiensis (Lsa), spp cluster 94_0 (Lsp1).*

Implications

This study highlights the importance of genomic data to expand the phylogenetic diversity and characterize the different traits in intracellular bacteria such as *Legionellaceae*. Given the heterogeneity of traits and environmental preferences in closely related microorganisms, it is important to design surveys able to reach species-level, if not genomovar-level, characterization to define accurately microorganisms' environmental niches and lifestyles. This is especially significant when closely related species vary not in their clinical importance, as the case of *Legionellaceae*, but also their biotechnological relevance. Methods to achieve this in environmental surveys already exist (e.g., clade-specific primers, full-length 16S rRNA sequencing, strain-resolved metagenomics)(90–92), but their adoption is still limited. Such methods could be applied to achieve

485 a detailed characterization of the environmental prevalence of single species and link this to
486 functional traits(93). Moreover, given the known difficulty in the isolation of host-associated
487 microorganisms, the use of MAGs-derived metabolic information will allow to isolate currently
488 uncultured species and more (e.g., (94)) . The combination of greater specificity of environmental
489 surveys and the isolation of novel species will enable to test predicted phenotypes and the genes
490 responsible for the adaptation to environmental conditions, as well as achieving a greater
491 understanding of eco-evolutionary processes and their interactions with other microorganisms.
492 With respect to this latter point, the environmental interactions between protistan host and
493 facultative intracellular constitute a major research gap for our understanding of their ecology. This
494 is especially relevant for microorganisms with wide host ranges, such as *Legionellaceae*(95), for
495 which it is necessary to identify their and the hosts traits affecting the outcome of the interaction to
496 build a mechanistic understanding of host-colonization processes. Beyond scientific value, this
497 knowledge would also allow for better optimization of biotechnological processes and aid
498 safeguarding public health by allowing the improvement of relevant legislation.
499

500 **Materials and Methods**

501 **Genomes retrieval, quality assessment and comparison**

502 A total of 5023 genomes classified as belonging to the family *Legionellaceae* were identified from
503 publicly-available sources (date last accession: 15/05/2023, Table S13). Sources included NCBI
504 Assembly database, ENA, JGI and other several large-scale metagenomics survey(96–104). In
505 addition, reads from Wee and coworkers(99) were assembled using spades v3.15.4(105) after read
506 trimming using fastp v0.23.2(106) and the removal of known UniVec contaminant(97) using
507 BWA-MEM2 v2.2.1(107) and samtools v1.16.1(108). The quality of all the genomes and their
508 taxonomic affiliation was assessed using CheckM2 v1.0.1(109), and GUNC v1.0.5(110) and
509 GTDB-Tk v2.3.0(111) to remove genomes with more than 5% contamination, less than 85%
510 complete and not belonging to the investigated family, eventually after manual refinement using
511 Anvi'o v8(112). Genomes dereplication was performed with an initial clustering at 90% ANI using

512 MASH v1.1(113), followed by a secondary one using skani v0.2.1(114) at 95% and 99.5% ANI.
513 Representative genomes for species (i.e., 95% ANI) and genomovars (i.e., 99.5% ANI) were
514 chosen as the ones with the highest value derived from the following formula(115): *Completeness*
515 – 5**Contamination* + 0.5*log(*N50*) + *Centrality*/100, using CheckM2 quality estimates and where
516 *N50* is the minimum length for which half of the genome belongs to contigs longer than *N50*, while
517 *Centrality* is the average ANI to genomes within the same secondary cluster. Due to the high
518 number of genomovars found for *L. pneumophila* and *L. fraseri,* only the species representative
519 and the three representatives of the genomovar with the highest number of genomes were
520 selected. Species' clinical-associations were retrieved from published literature(116).
521
522 **Proteins predictions and functional annotations**
523 Proteins were predicted using Bakta v1.8.1(117) (v5.0, full database) and annotated using
524 eggNOG-mapper v2.1.12(118). Amino acid similarity was assessed using FastAAI v0.1.20(119).
525 T4SS were detected using CONJScan v2.0.1(120), while other secretion systems, flagella and pili
526 using TXSScan v1.1.3(121). T6SSs were considered as present in case more than nine proteins
527 were detected in order to match previous reports(39). Plasmid-associated contigs were detected
528 using Platon v1.7(122). The presence of capsule systems was assessed using CapsuleFinder
529 v1.0.5(64). Sensory and regulatory proteins were selected based on eggNOG-mapper annotations
530 and the signal transduction specific KEGG profiles from MiST v4.0 database(123). Chemotaxis and
531 methyl-accepting proteins were detected using HMMER v3.4(124) (minimum bitscore: 25) and the
532 models from the MiST database. KEGG metabolic pathways completeness was estimated using
533 DRAM v1.4.6(125), while the amino acids biosynthetic potential was evaluated using GapMind and
534 gapseq. Clusters of genomes presenting similar metabolic capabilities were obtained through
535 HDBSCAN clustering based on the euclidean distance of metabolic pathways completeness.
536 CAZYmes were annotated against the dbCAN3 database using dbcan_run v4.0.0(126), retaining
537 only proteins identified by at least two detection methods. Peptidases were identified searching the
538 predicted proteins against the MEROPS v12.4 database(127) using MMseqs2 v14-7e284(128)

539 retaining proteins showing at least 30% identity over 65% query coverage and displaying a bitscore
540 ≥50. Secreted proteins were identified using SignalP v6.0(129). Homologs of selected *L.*
541 *pneumophila* strain *Paris* proteins were detected using BlastP v2.12.0(130) using the same
542 thresholds used for MMseqs2 and inspecting Bakta and eggNOG-mapper annotations (e.g.,
543 PFAMs detected). Type 1 secretion systems were considered as present if both
544 LssD/lrpssD/lb1ssD/lb2ssD and LssB/lrpssB/lb1ssB/lb2ssB proteins were detected within a five
545 proteins neighbourhood(49). Eventual duplicated matches with TXSScan results were assessed
546 using a custom script.
547
548 **Pangenome and phylogeny construction**
549 Predicted proteins were used to perform calculations using PPanGGOLiN v1.2.105(131). A
550 pangenome employing a protein clustering identity percentage threshold of 80% was used to
551 identify core proteins (i.e., present in ≥95% of genomes) to reconstruct phylogeny. Identified
552 proteins were searched for homologs in the genome of *C. burnetii* (RefSeq accession:
553 GCF_005280755.1). Proteins were then aligned using MAFFT G-INS-i as implemented in
554 Concatenator v0.2.1(132, 133) and used to calculate maximum likelihood phylogenetic trees using
555 IQ-TREE v2.2.2.6(134) under the LG model and automated model selection(135). Support values
556 were calculated based on 1000 ultrafast bootstraps with bootstrap tree optimization(136). The
557 resulting phylogenetic tree was visualized and annotated in iTOL v6(137). An initial genus cluster
558 membership was obtained using GTDB-Tk de-novo workflow using calibrated RED values.
559 Successively, cluster membership was refined manually to remove polyphyly. A second
560 pangenome employing an identity percentage threshold of 50% was used to identify gene families
561 among *Legionellaceae* genomes and identify shared proteins, as suggested by PPanGGOLiN
562 authors.
563
564 **Assessment of species environmental distribution**

One random full length 16S rRNA gene per genome was retrieved from the Bakta annotation of the selected species representatives or, otherwise, from other genomes belonging to the same species cluster. 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.

Statistical testing and data analysis

Analysis of the bioinformatic results was performed in R v4.3.0(138) leveraging the functions provided by libraries Tidyverse v2.0.0(139), ape v5.7-1(140), dbscan v1.2-0(141) and vegan v2.6-4(142). Statistical significance was assessed using appropriate two-tailed tests and a threshold p-value of 0.05.

Data availability

Supplementary data and tables S1-S13 are available on Zenodo at the following link: https://doi.org/10.5281/zenodo.14748706.

Author contributions

M.G.: Conceptualization, Methodology, Formal Analysis, Investigation, Data Curation, Visualization, Writing - Original Draft; A.C.: Writing - Review & Editing; F.H.: Writing - Review & Editing, Funding acquisition.

Acknowledgments

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

592 Offices of Public Health (FOPH) and Energy (SFOE) of Switzerland, as well as Eawag discretionary
593 funding. The authors would also like to acknowledge Alyssa Henderson for her constructive
594 feedback on genomic-scale metabolic models, David Johnson and Tim Julian for the valuable
595 comments on an initial manuscript version and Valeria Gaia for providing additional *Legionellaceae*
596 isolates.
597

598 **Competing interest statement**

599 None to declare.

600

601 **References**

602 1. L. R. Thompson, *et al.*, A communal catalogue reveals Earth's multiscale microbial
603 diversity. *Nature* **551**, 457–463 (2017).

604 2. Z. D. Blount, The unexhausted potential of E. coli. *eLife* **4**, e05826 (2015).

605 3. K. Hrovat, B. E. Dutilh, M. H. Medema, C. Melkonian, Taxonomic resolution of different
606 16S rRNA variable regions varies strongly across plant-associated bacteria. *ISME*
607 *Communications* **4**, ycae034 (2024).

608 4. J. A. Leigh, S.-V. Albers, H. Atomi, T. Allers, Model organisms for genetics in the domain
609 Archaea: methanogens, halophiles, *Thermococcales* and *Sulfolobales*. *FEMS Microbiol Rev*
610 **35**, 577–608 (2011).

611 5. C. W. Beukes, *et al.*, Genome Data Provides High Support for Generic Boundaries in
612 Burkholderia Sensu Lato. *Front. Microbiol.* **8**, 1154 (2017).

613 6. T. Van Rossum, P. Ferretti, O. M. Maistrenko, P. Bork, Diversity within species:
614 interpreting strains in microbiomes. *Nat Rev Microbiol* **18**, 491–506 (2020).

615 7. K. C. Crosby, *et al.*, Genomic delineation and description of species and within-species
616 lineages in the genus Pantoea. *Front. Microbiol.* **14**, 1254999 (2023).

617 8. M. Baker-Goering, K. Roy, C. Edens, S. Collier, Economic Burden of Legionnaires' Disease,
618 United States, 2014. *Emerg. Infect. Dis.* **27**, 255–257 (2021).

619 9. M. A. Moffa, *et al.*, Legionellosis on the rise: A scoping review of sporadic, community-
620 acquired incidence in the United States. *Epidemiol. Infect.* **151**, e133 (2023).

621 10. LPSN, Family Legionellaceae. Available at: https://lpsn.dsmz.de/family/legionellaceae
622 [Accessed 8 December 2024].

623 11. V. L. Yu, *et al.*, Distribution of *Legionella* Species and Serogroups Isolated by Culture in
624 Patients with Sporadic Community-Acquired Legionellosis: An International Collaborative
625 Survey. *J INFECT DIS* **186**, 127–128 (2002).

626 12. P. W. J. J. van der Wielen, W. Wierenga, F. I. H. M. Oesterholt, A. Oostdijk, A. van der
627 Werff, "Rightly to an effective Legionella prevention (in Dutch)" (2021).

628 13. National Academies of Sciences, Engineering, and Medicine, *Management of Legionella*
629 *in Water Systems* (National Academies Press, 2020).

630 14. E. Van Kenhove, K. Dinne, A. Janssens, J. Laverge, Overview and comparison of Legionella
631 regulations worldwide. *American Journal of Infection Control* **47**, 968–978 (2019).

632 15. S. J. Joseph, *et al.*, Dynamics of genome change among Legionella species. *Sci Rep* **6**,
633 33442 (2016).

634 16. L. Gomez-Valero, *et al.*, More than 18,000 effectors in the *Legionella* genus genome
635 provide multiple, independent combinations for replication in human cells. *Proc. Natl.*
636 *Acad. Sci. U.S.A.* **116**, 2265–2273 (2019).

637 17. N. J. Tobias, *et al.*, *Legionella* shows a diverse secondary metabolism dependent on a
638 broad spectrum Sfp-type phosphopantetheinyl transferase. *PeerJ* **4**, e2720 (2016).

639 18. T. Graells, H. Ishak, M. Larsson, L. Guy, The all-intracellular order Legionellales is
640 unexpectedly diverse, globally distributed and lowly abundant. *FEMS Microbiology*
641 *Ecology* **94** (2018).

642 19. S. R. Vartoukian, R. M. Palmer, W. G. Wade, Strategies for culture of 'unculturable'
643 bacteria: Culturing the unculturable. *FEMS Microbiology Letters* **309**, 1–7 (2010).

644 20. L. M. Rodriguez-R, *et al.*, An ANI gap within bacterial species that advances the definitions
645 of intra-species units. *mBio* **15**, e02696-23 (2024).

646 21. A. Oren, G. Garrity, Notification of changes in taxonomic opinion previously published
647 outside the IJSEM. List of changes in taxonomic opinion no. 35. *International Journal of*
648 *Systematic and Evolutionary Microbiology* **72** (2022).

649 22. N. Saini, R. S. Gupta, A robust phylogenetic framework for members of the order
650 Legionellales and its main genera (Legionella, Aquicella, Coxiella and Rickettsiella) based

651 on phylogenomic analyses and identification of molecular markers demarcating different
652 clades. *Antonie van Leeuwenhoek* **114**, 957–982 (2021).

653 23. N. A. Kozak-Muiznieks, *et al.*, Comparative genome analysis reveals a complex population
654 structure of Legionella pneumophila subspecies. *Infection, Genetics and Evolution* **59**,
655 172–185 (2018).

656 24. L. Philippot, *et al.*, The ecological coherence of high bacterial taxonomic ranks. *Nat Rev*
657 *Microbiol* **8**, 523–529 (2010).

658 25. K. T. Konstantinidis, J. M. Tiedje, Prokaryotic taxonomy and phylogeny in the genomic
659 era: advancements and challenges ahead. *Current Opinion in Microbiology* **10**, 504–509
660 (2007).

661 26. D. H. Parks, *et al.*, A standardized bacterial taxonomy based on genome phylogeny
662 substantially revises the tree of life. *Nat Biotechnol* **36**, 996–1004 (2018).

663 27. S. T. Chambers, S. Slow, A. Scott-Thomas, D. R. Murdoch, Legionellosis Caused by Non-
664 Legionella pneumophila Species, with a Focus on Legionella longbeachae.
665 *Microorganisms* **9**, 291 (2021).

666 28. R. R. Muder, V. L. Yu, Infection Due to *Legionella* Species Other Than *L. pneumophila*.
667 *Clinical Infectious Diseases* **35**, 990–998 (2002).

668 29. C. Wang, X. Chuai, M. Liang, Legionella feeleii: pneumonia or Pontiac fever? Bacterial
669 virulence traits and host immune response. *Med Microbiol Immunol* **208**, 25–32 (2019).

670 30. G. M. Garrity, A. Brown, R. M. Vickers, Tatlockia and Fluoribacter: Two New Genera of
671 Organisms Resembling Legionella pneumophila. *International Journal of Systematic*
672 *Bacteriology* **30**, 609–614 (1980).

673 31. N. K. Fry, S. Warwick, N. A. Saunders, T. M. Embley, The use of 16S ribosomal RNA
674 analyses to investigate the phylogeny of the family Legionellaceae. *Journal of General*
675 *Microbiology* **137**, 1215–1222 (1991).

676 32. H. B. Hassler, *et al.*, Phylogenies of the 16S rRNA gene and its hypervariable regions lack
677 concordance with core genome phylogenies. *Microbiome* **10**, 104 (2022).

678 33. S. Köstlbacher, *et al.*, Pangenomics reveals alternative environmental lifestyles among
679 chlamydiae. *Nat Commun* **12**, 4021 (2021).

680 34. A. E. Dewar, C. Hao, L. J. Belcher, M. Ghoul, S. A. West, Bacterial lifestyle shapes
681 pangenomes. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2320170121 (2024).

682 35. H. Tettelin, D. Riley, C. Cattuto, D. Medini, Comparative genomics: the bacterial pan-
683 genome. *Current Opinion in Microbiology* **11**, 472–477 (2008).

684 36. V. Poirier, Y. Av-Gay, Intracellular Growth of Bacterial Pathogens: The Role of Secreted
685 Effector Proteins in the Control of Phagocytosed Microorganisms. *Microbiol Spectr* **3**,
686 3.6.12 (2015).

687 37. K. Wallden, A. Rivera-Calzada, G. Waksman, Microreview: Type IV secretion systems:
688 versatility and diversity in function: Diversity and versatility of the type IV secretion
689 system. *Cellular Microbiology* **12**, 1203–1212 (2010).

690 38. R. C. White, N. P. Cianciotto, Assessing the impact, genomics and evolution of type II
691 secretion across a large, medically important genus: the Legionella type II secretion
692 paradigm. *Microbial Genomics* **5** (2019).

693 39. T. Qin, H. Zhou, H. Ren, W. Liu, Distribution of Secretion Systems in the Genus Legionella
694 and Its Correlation with Pathogenicity. *Front. Microbiol.* **08** (2017).

695 40. M. A. Söderberg, J. Dao, S. R. Starkenburg, N. P. Cianciotto, Importance of Type II
696 Secretion for Survival of *Legionella pneumophila* in Tap Water and in Amoebae at Low
697 Temperatures. *Appl Environ Microbiol* **74**, 5583–5588 (2008).

698 41. V. J. Carrión, *et al.*, Pathogen-induced activation of disease-suppressive functions in the
699 endophytic root microbiome. *Science* **366**, 606–612 (2019).

700 42. L. Wang, *et al.*, Mining of novel secondary metabolite biosynthetic gene clusters from
701 acid mine drainage. *Sci Data* **9**, 760 (2022).

702 43. E. De Buck, J. Anné, E. Lammertyn, The role of protein secretion systems in the virulence
703 of the intracellular pathogen Legionella pneumophila. *Microbiology* **153**, 3948–3953
704 (2007).

705 44. L. Gomez-Valero, *et al.*, Comparative analyses of Legionella species identifies genetic
706 features of strains causing Legionnaires disease. *Genome Biology* **15** (2014).

707 45. F. Fuche, A. Vianney, C. Andrea, P. Doublet, C. Gilbert, Functional Type 1 Secretion
708 System Involved in Legionella pneumophila Virulence. *J. Bacteriol.* **197**, 563–571 (2015).

709 46. I. Meuskens, A. Saragliadis, J. C. Leo, D. Linke, Type V Secretion Systems: An Overview of
710 Passenger Domain Functions. *Front. Microbiol.* **10**, 1163 (2019).

711 47. R. Gallegos-Monterrosa, S. J. Coulthurst, The ecological impact of a bacterial weapon:
712 microbial interactions and the Type VI secretion system. *FEMS Microbiology Reviews* **45**,
713 fuab033 (2021).

714 48. A. K. M. Stubbusch, *et al.*, Antagonism as a foraging strategy in microbial communities.
715 [Preprint] (2024). Available at: http://biorxiv.org/lookup/doi/10.1101/2024.11.04.621785
716 [Accessed 12 December 2024].

717 49. C. L. Brown, *et al.*, Whole genome sequence analysis reveals the broad distribution of the
718 RtxA type 1 secretion system and four novel putative type 1 secretion systems
719 throughout the Legionella genus. *PLoS ONE* **15**, e0223033 (2020).

720 50. A. W. Brown, P. S. Simone, J. C. York, G. L. Emmert, A device for fully automated on-site
721 process monitoring and control of trihalomethane concentrations in drinking water.
722 *Analytica Chimica Acta* **853**, 351–359 (2015).

723 51. T. Qin, H. Zhou, H. Ren, W. Liu, Distribution of Secretion Systems in the Genus Legionella
724 and Its Correlation with Pathogenicity. *Front. Microbiol.* **08** (2017).

725 52. A. W. Ensminger, Legionella pneumophila, armed to the hilt: justifying the largest arsenal
726 of effectors in the bacterial world. *Current Opinion in Microbiology* **29**, 74–80 (2016).

727 53. L. Gomez-Valero, *et al.*, More than 18,000 effectors in the *Legionella* genus genome
728 provide multiple, independent combinations for replication in human cells. *Proc. Natl.*
729 *Acad. Sci. U.S.A.* **116**, 2265–2273 (2019).

730 54. J.-B. Raina, V. Fernandez, B. Lambert, R. Stocker, J. R. Seymour, The role of microbial
731 motility and chemotaxis in symbiosis. *Nat Rev Microbiol* **17**, 284–294 (2019).

732 55. M. Shaheen, N. J. Ashbolt, Differential Bacterial Predation by Free-Living Amoebae May
733 Result in Blooms of Legionella in Drinking Water Systems. *Microorganisms* **9**, 174 (2021).

734 56. J. Hoppe, *et al.*, PilY1 Promotes Legionella pneumophila Infection of Human Lung Tissue
735 Explants and Contributes to Bacterial Adhesion, Host Cell Invasion, and Twitching
736 Motility. *Front. Cell. Infect. Microbiol.* **7** (2017).

737 57. C. Cazalet, *et al.*, Analysis of the Legionella longbeachae Genome and Transcriptome
738 Uncovers Unique Strategies to Cause Legionnaires' Disease. *PLoS Genet* **6**, e1000851
739 (2010).

740 58. C. Duncan, *et al.*, Lcl of Legionella pneumophila Is an Immunogenic GAG Binding Adhesin
741 That Promotes Interactions with Lung Epithelial Cells and Plays a Crucial Role in Biofilm
742 Formation. *Infect Immun* **79**, 2168–2181 (2011).

743 59. M. A. Matilla, T. Krell, The effect of bacterial chemotaxis on host infection and
744 pathogenicity. *FEMS Microbiology Reviews* **42** (2018).

745 60. C. Dietrich, K. Heuner, B. C. Brand, J. Hacker, M. Steinert, Flagellum of *Legionella*
746 *pneumophila* Positively Affects the Early Phase of Infection of Eukaryotic Host Cells. *Infect*
747 *Immun* **69**, 2116–2122 (2001).

748 61. K. Wuichet, I. B. Zhulin, Origins and Diversification of a Complex Signal Transduction
749 System in Prokaryotes. *Sci. Signal.* **3** (2010).

750 62. B. Chaban, H. V. Hughes, M. Beeby, The flagellum in bacterial pathogens: For motility and
751 a whole lot more. *Seminars in Cell & Developmental Biology* **46**, 91–103 (2015).

752 63. D. A. Coil, J. Anné, Twitching motility in *Legionella pneumophila*. *FEMS Microbiology*
753 *Letters* **293**, 271–277 (2009).

754 64. O. Rendueles, M. Garcia-Garcerà, B. Néron, M. Touchon, E. P. C. Rocha, Abundance and
755 co-occurrence of extracellular capsules increase environmental breadth: Implications for
756 the emergence of pathogens. *PLoS Pathog* **13**, e1006525 (2017).

757 65. S. Schmidt, *et al.*, The unique Legionella longbeachae capsule favors intracellular
758 replication and immune evasion. *PLoS Pathog* **20**, e1012534 (2024).

759 66. S. A. Hsieh, D. L. Donermeyer, S. C. Horvath, P. M. Allen, Phase-variable bacteria
760 simultaneously express multiple capsules: Read the story behind the paper on the
761 Microbe Post here. *Microbiology* **167** (2021).

762 67. G. Oliva, T. Sahr, C. Buchrieser, The Life Cycle of L. pneumophila: Cellular Differentiation
763 Is Linked to Virulence and Metabolism. *Front. Cell. Infect. Microbiol.* **8**, 3 (2018).

764 68. E. Eylert, *et al.*, Isotopologue Profiling of Legionella pneumophila. *Journal of Biological*
765 *Chemistry* **285**, 22232–22243 (2010).

766 69. W. Eisenreich, K. Heuner, The life stage-specific pathometabolism of *Legionella*
767 *pneumophila*. *FEBS Letters* **590**, 3868–3886 (2016).

768 70. I. Häuslein, C. Manske, W. Goebel, W. Eisenreich, H. Hilbi, Pathway analysis using ^{13}C-
769 glycerol and other carbon tracers reveals a bipartite metabolism of *Legionella*
770 *pneumophila*. *Molecular Microbiology* **100**, 229–246 (2016).

771 71. R. Eisenhofer, I. Odriozola, A. Alberdi, Impact of microbial genome completeness on
772 metagenomic functional inference. *ISME Communications* **3**, 12 (2023).

773 72. B. L. Turner, M. J. Papházy, P. M. Haygarth, I. D. Mckelvie, Inositol phosphates in the
774 environment. *Phil. Trans. R. Soc. Lond. B* **357**, 449–469 (2002).

775 73. N. Gillmaier, *et al.*, Growth-related Metabolism of the Carbon Storage Poly-3-
776 hydroxybutyrate in Legionella pneumophila. *Journal of Biological Chemistry* **291**, 6471–
777 6482 (2016).

778 74. C. Manske, U. Schell, H. Hilbi, Metabolism of *myo* -Inositol by Legionella pneumophila
779 Promotes Infection of Amoebae and Macrophages. *Appl Environ Microbiol* **82**, 5000–5014
780 (2016).

781 75. C. M. Lopez-Vazquez, *et al.*, Modeling the PAO–GAO competition: Effects of carbon
782 source, pH and temperature. *Water Research* **43**, 450–462 (2009).

783 76. T. T. H. Nguyen, D. D. Myrold, R. S. Mueller, Distributions of Extracellular Peptidases
784 Across Prokaryotic Genomes Reflect Phylogeny and Habitat. *Front. Microbiol.* **10**, 413
785 (2019).

786 77. B. W. James, W. S. Mauchline, P. J. Dennis, C. W. Keevil, R. Wait, Poly-3-Hydroxybutyrate
787 in *Legionella pneumophila* , an Energy Source for Survival in Low-Nutrient Environments.
788 *Appl Environ Microbiol* **65**, 822–827 (1999).

789 78. J. B. McKinlay, Are Bacteria Leaky? Mechanisms of Metabolite Externalization in Bacterial
790 Cross-Feeding. *Annu. Rev. Microbiol.* **77**, 277–297 (2023).

791 79. J. Ramoneda, T. B. N. Jensen, M. N. Price, E. O. Casamayor, N. Fierer, Taxonomic and
792 environmental distribution of bacterial amino acid auxotrophies. *Nat Commun* **14**, 7608
793 (2023).

794 80. G. D'Souza, *et al.*, Less is more: selective advantages can explain the prevalent loss of
795 biosynthetic genes in bacteria: adaptive loss of biosynthetic genes in bacteria. *Evolution*
796 **68**, 2559–2570 (2014).

797 81. M. N. Price, A. M. Deutschbauer, A. P. Arkin, GapMind: Automated Annotation of Amino
798 Acid Biosynthesis. *mSystems* **5**, 10.1128/msystems.00291-20 (2020).

799 82. J. Zimmermann, C. Kaleta, S. Waschina, gapseq: informed prediction of bacterial
800 metabolic pathways and reconstruction of accurate metabolic models. *Genome Biol* **22**,
801 81 (2021).

802 83. M. H. Stipanuk, SULFUR AMINO ACID METABOLISM: Pathways for Production and
803 Removal of Homocysteine and Cysteine. *Annu. Rev. Nutr.* **24**, 539–577 (2004).

804 84. N. M. Kredich, Biosynthesis of Cysteine. *EcoSal Plus* **3**, 10.1128/ecosalplus.3.6.1.11
805 (2008).

806 85. T. C. Lee, R. M. Vickers, V. L. Yu, M. M. Wagener, Growth of 28 Legionella species on
807 selective culture media: a comparative study. *J Clin Microbiol* **31**, 2764–2768 (1993).

808 86. J. F. Matias Rodrigues, T. S. B. Schmidt, J. Tackmann, C. Von Mering, MAPseq: highly
809 efficient k-mer search with confidence estimates, for rRNA sequence analysis.
810 *Bioinformatics* **33**, 3808–3810 (2017).

811 87. D. Emuron, V. Senadhi, Gastrointestinal Legionella: An Underecognized Entity and Cause
812 of Secretory Diarrhea: A Case and Review of the Literature: 889. *Official journal of the*
813 *American College of Gastroenterology | ACG* **106** (2011).

814 88. M. T. Collins, S. N. Cho, J. S. Reif, Prevalence of antibodies to Legionella pneumophila in
815 animal populations. *J Clin Microbiol* **15**, 130–136 (1982).

816 89. A. Casadevall, Evolution of Intracellular Pathogens. *Annu. Rev. Microbiol.* **62**, 19–33
817 (2008).

818 90. N. Segata, On the Road to Strain-Resolved Comparative Metagenomics. *mSystems* **3**,
819 e00190-17 (2018).

820 91. J. S. Johnson, *et al.*, Evaluation of 16S rRNA gene sequencing for species and strain-level
821 microbiome analysis. *Nat Commun* **10**, 5029 (2019).

822 92. R. P. A. Pereira, J. Peplies, I. Brettar, M. G. Höfle, Development of a genus-specific next
823 generation sequencing approach for sensitive and quantitative determination of the
824 Legionella microbiome in freshwater systems. *BMC Microbiol* **17**, 79 (2017).

825 93. J. Ramoneda, M. Hoffert, E. Stallard-Olivera, E. O. Casamayor, N. Fierer, Leveraging
826 genomic information to predict environmental preferences of bacteria. *The ISME Journal*
827 wrae195 (2024). https://doi.org/10.1093/ismejo/wrae195.

828 94. A. Omsland, T. Hackstadt, R. A. Heinzen, Bringing Culture to the Uncultured: Coxiella
829 burnetii and Lessons for Obligate Intracellular Bacterial Pathogens. *PLoS Pathog* **9**,
830 e1003540 (2013).

831 95. D. K. Boamah, G. Zhou, A. W. Ensminger, T. J. O'Connor, From Many Hosts, One
832 Accidental Pathogen: The Diverse Protozoan Hosts of Legionella. *Front. Cell. Infect.*
833 *Microbiol.* **7**, 477 (2017).

834 96. I.-M. A. Chen, *et al.*, IMG/M v.5.0: an integrated data management and comparative
835 analysis system for microbial genomes and microbiomes. *Nucleic Acids Research* **47**,
836 D666–D677 (2019).

837 97. K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, E. W. Sayers, GenBank. *Nucleic Acids*
838 *Res* **44**, D67–D72 (2016).

839 98. R. Leinonen, H. Sugawara, M. Shumway, on behalf of the International Nucleotide
840 Sequence Database Collaboration, The Sequence Read Archive. *Nucleic Acids Research*
841 **39**, D19–D21 (2011).

842 99. B. A. Wee, *et al.*, Population analysis of Legionella pneumophila reveals a basis for
843 resistance to complement-mediated killing. *Nat Commun* **12**, 7165 (2021).

844 100. M. Buck, *et al.*, Comprehensive dataset of shotgun metagenomes from oxygen stratified
845 freshwater lakes and ponds. *Sci Data* **8**, 131 (2021).

846 101. T. A. Gurbich, *et al.*, MGnify Genomes: A Resource for Biome-specific Microbial Genome
847 Catalogues. *Journal of Molecular Biology* 168016 (2023).
848 https://doi.org/10.1016/j.jmb.2023.168016.

849 102. S. Nayfach, *et al.*, A genomic catalog of Earth's microbiomes. *Nat Biotechnol* **39**, 499–509
850 (2021).

851 103. Y. Nishimura, S. Yoshizawa, The OceanDNA MAG catalog contains over 50,000
852 prokaryotic genomes originated from various marine environments. *Sci Data* **9**, 305
853 (2022).

854 104. A. S. Sudarshan, *et al.*, New Drinking Water Genome Catalog Identifies a Globally
855 Distributed Bacterial Genus Adapted to Disinfected Drinking Water Systems. *Environ. Sci.*
856 *Technol.* **58**, 16475–16487 (2024).

857 105. A. Prjibelski, D. Antipov, D. Meleshko, A. Lapidus, A. Korobeynikov, Using SPAdes De Novo
858 Assembler. *CP in Bioinformatics* **70**, e102 (2020).

859 106. S. Chen, Y. Zhou, Y. Chen, J. Gu, fastp: an ultra-fast all-in-one FASTQ preprocessor.
860 *Bioinformatics* **34**, i884–i890 (2018).

861 107. Md. Vasimuddin, S. Misra, H. Li, S. Aluru, Efficient Architecture-Aware Acceleration of
862 BWA-MEM for Multicore Systems in *2019 IEEE International Parallel and Distributed*
863 *Processing Symposium (IPDPS)*, (IEEE, 2019), pp. 314–324.

864 108. H. Li, *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**,
865 2078–2079 (2009).

866 109. A. Chklovski, D. H. Parks, B. J. Woodcroft, G. W. Tyson, CheckM2: a rapid, scalable and
867 accurate tool for assessing microbial genome quality using machine learning. *Nat*
868 *Methods* **20**, 1203–1212 (2023).

869 110. A. Orakov, *et al.*, GUNC: detection of chimerism and contamination in prokaryotic
870 genomes. *Genome Biol* **22**, 178 (2021).

871 111. P.-A. Chaumeil, A. J. Mussig, P. Hugenholtz, D. H. Parks, GTDB-Tk v2: memory friendly
872 classification with the genome taxonomy database. *Bioinformatics* **38**, 5315–5316 (2022).

873 112. A. M. Eren, *et al.*, Community-led, integrated, reproducible multi-omics with anvi'o. *Nat*
874 *Microbiol* **6**, 3–6 (2020).

875 113. B. D. Ondov, *et al.*, Mash: fast genome and metagenome distance estimation using
876 MinHash. *Genome Biol* **17**, 132 (2016).

877 114. J. Shaw, Y. W. Yu, Fast and robust metagenomic sequence comparison through sparse
878 chaining with skani. *Nat Methods* **20**, 1661–1665 (2023).

879 115. M. R. Olm, C. T. Brown, B. Brooks, J. F. Banfield, dRep: a tool for fast and accurate
880 genomic comparisons that enables improved genome recovery from metagenomes
881 through de-replication. *ISME J* **11**, 2864–2868 (2017).

882 116. LeCo Project, LegioSpecies - an updated and referenced list of Legionellaceae species.
883 (2024). https://doi.org/10.5281/ZENODO.11072744. Deposited 2024.

884 117. O. Schwengers, *et al.*, Bakta: rapid and standardized annotation of bacterial genomes via
885 alignment-free sequence identification: Find out more about Bakta, the motivation,
886 challenges and applications, here. *Microbial Genomics* **7** (2021).

887 118. C. P. Cantalapiedra, A. Hernández-Plaza, I. Letunic, P. Bork, J. Huerta-Cepas, eggNOG-
888 mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the
889 Metagenomic Scale. *Molecular Biology and Evolution* **38**, 5825–5829 (2021).

890 119. K. Konstantinidis, *et al.*, FastAAI: Efficient Estimation of Genome Average Amino Acid
891 Identity and Phylum-level relationships using Tetramers of Universal Proteins. [Preprint]
892 (2022). Available at: https://www.researchsquare.com/article/rs-1459378/v1 [Accessed
893 11 October 2024].

894 120. J. Cury, S. S. Abby, O. Doppelt-Azeroual, B. Néron, E. P. C. Rocha, "Identifying Conjugative
895 Plasmids and Integrative Conjugative Elements with CONJscan" in *Horizontal Gene*
896 *Transfer*, Methods in Molecular Biology., F. De La Cruz, Ed. (Springer US, 2020), pp. 265–
897 283.

898 121. S. S. Abby, *et al.*, Identification of protein secretion systems in bacterial genomes. *Sci Rep*
899 **6**, 23080 (2016).

900 122. O. Schwengers, *et al.*, Platon: identification and characterization of bacterial plasmid
901 contigs in short-read draft assemblies exploiting protein sequence-based replicon
902 distribution scores. *Microbial Genomics* **6** (2020).

903 123. V. M. Gumerov, L. E. Ulrich, I. B. Zhulin, MiST 4.0: a new release of the microbial signal
904 transduction database, now with a metagenomic component. *Nucleic Acids Research* **52**,
905 D647–D653 (2024).

906 124. S. R. Eddy, Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**, e1002195 (2011).

907 125. M. Shaffer, *et al.*, DRAM for distilling microbial metabolism to automate the curation of
908 microbiome function. *Nucleic Acids Research* **48**, 8883–8900 (2020).

909 126. J. Zheng, *et al.*, dbCAN3: automated carbohydrate-active enzyme and substrate
910 annotation. *Nucleic Acids Research* **51**, W115–W121 (2023).

911 127. N. D. Rawlings, *et al.*, The MEROPS database of proteolytic enzymes, their substrates and
912 inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic*
913 *Acids Research* **46**, D624–D632 (2018).

914 128. M. Steinegger, J. Söding, MMseqs2 enables sensitive protein sequence searching for the
915 analysis of massive data sets. *Nat Biotechnol* **35**, 1026–1028 (2017).

916 129. F. Teufel, *et al.*, SignalP 6.0 predicts all five types of signal peptides using protein
917 language models. *Nat Biotechnol* **40**, 1023–1025 (2022).

918 130. C. Camacho, *et al.*, BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421
919 (2009).

920 131. G. Gautreau, *et al.*, PPanGGOLiN: Depicting microbial diversity via a partitioned
921 pangenome graph. *PLoS Comput Biol* **16**, e1007732 (2020).

922 132. M. Vences, *et al.*, iTaxoTools 0.1: Kickstarting a specimen-based software toolkit for
923 taxonomists. *MT* **6** (2021).

924 133. K. Katoh, D. M. Standley, MAFFT Multiple Sequence Alignment Software Version 7:
925 Improvements in Performance and Usability. *Molecular Biology and Evolution* **30**, 772–
926 780 (2013).

927 134. B. Q. Minh, *et al.*, IQ-TREE 2: New Models and Efficient Methods for Phylogenetic
928 Inference in the Genomic Era. *Molecular Biology and Evolution* **37**, 1530–1534 (2020).

929 135. S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. Von Haeseler, L. S. Jermiin,
930 ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* **14**,
931 587–589 (2017).

932 136. D. T. Hoang, O. Chernomor, A. Von Haeseler, B. Q. Minh, L. S. Vinh, UFBoot2: Improving
933 the Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution* **35**, 518–522
934 (2018).

935 137. I. Letunic, P. Bork, Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic
936 tree display and annotation tool. *Nucleic Acids Research* **52**, W78–W82 (2024).

937 138. R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for
938 Statistical Computing, 2023).

939 139. H. Wickham, *et al.*, Welcome to the Tidyverse. *JOSS* **4**, 1686 (2019).

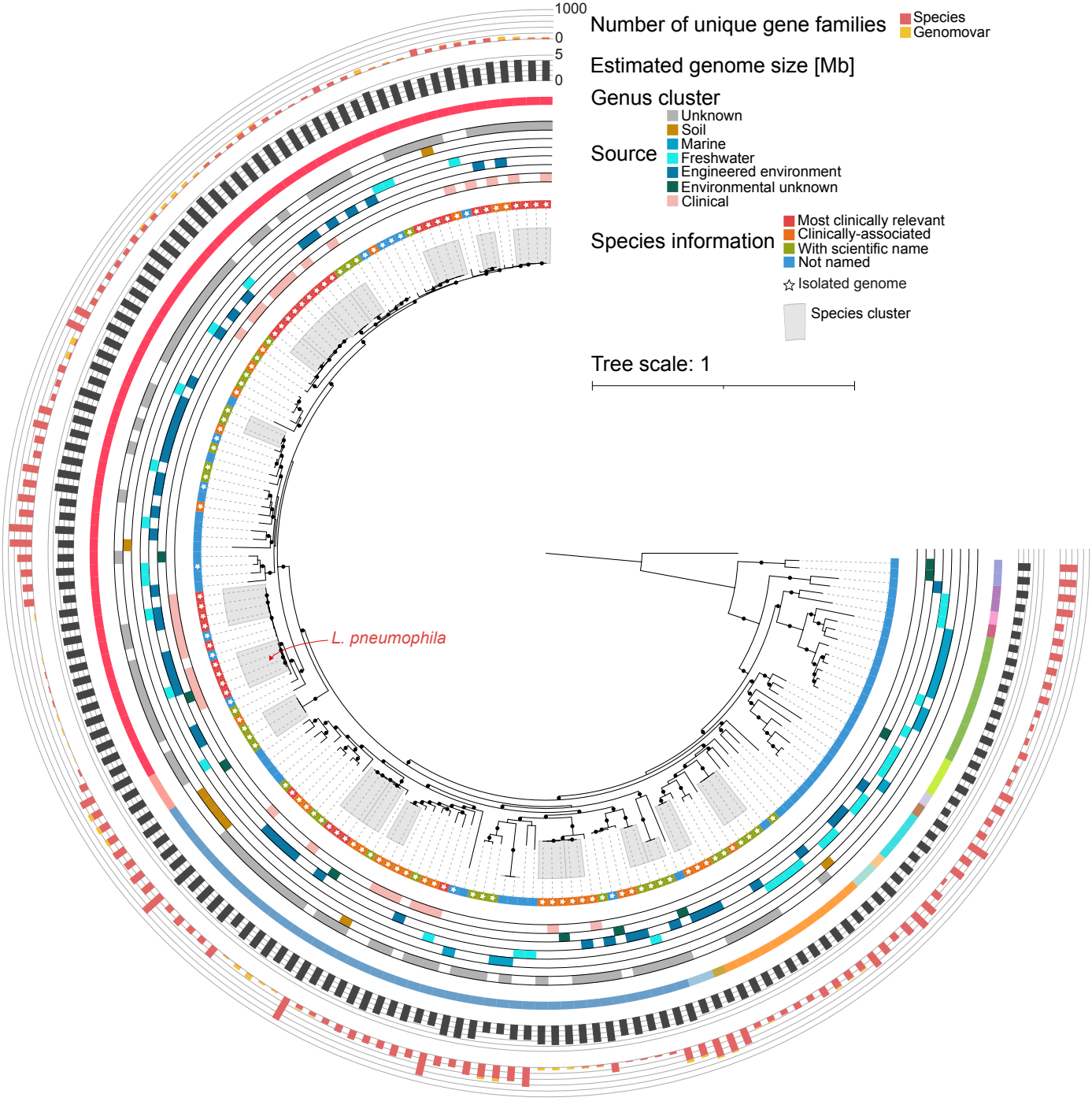
140. E. Paradis, J. Claude, K. Strimmer, APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289–290 (2004).

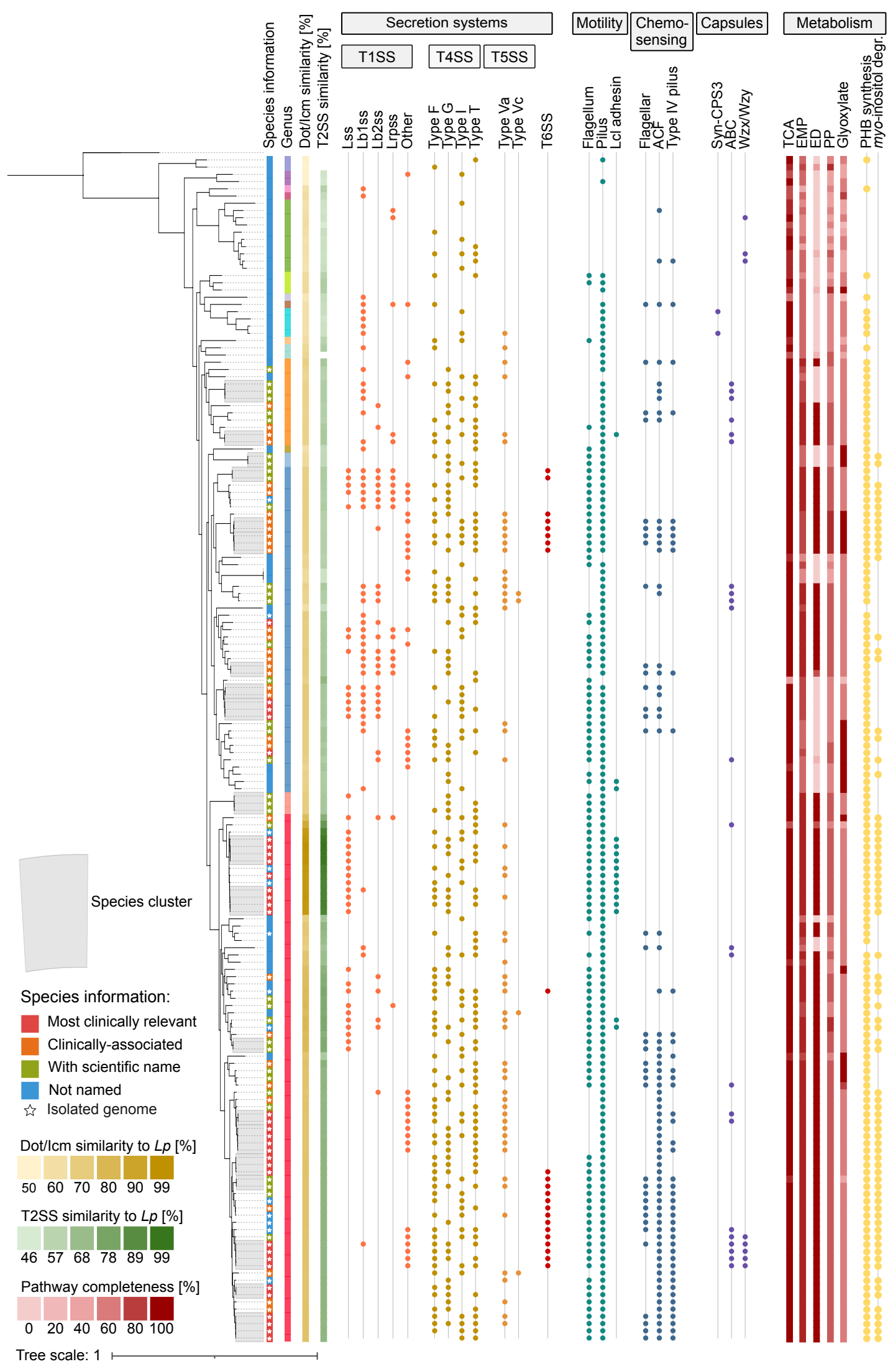
141. M. Hahsler, M. Piekenbrock, D. Doran, dbscan : Fast Density-Based Clustering with *R*. *J. Stat. Soft.* **91** (2019).

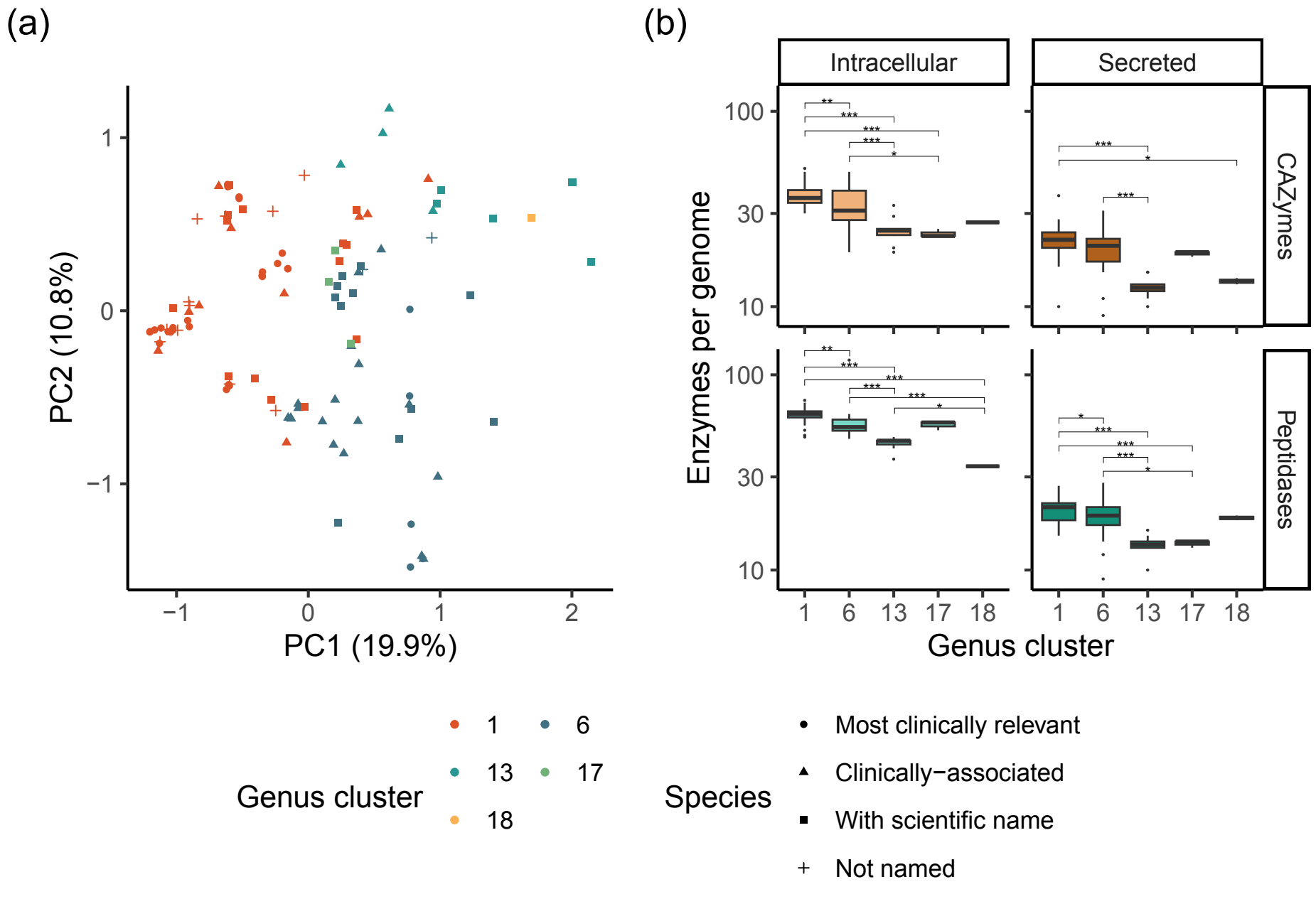
142. J. Oksanen, *et al.*, *vegan: Community Ecology Package* (2022).

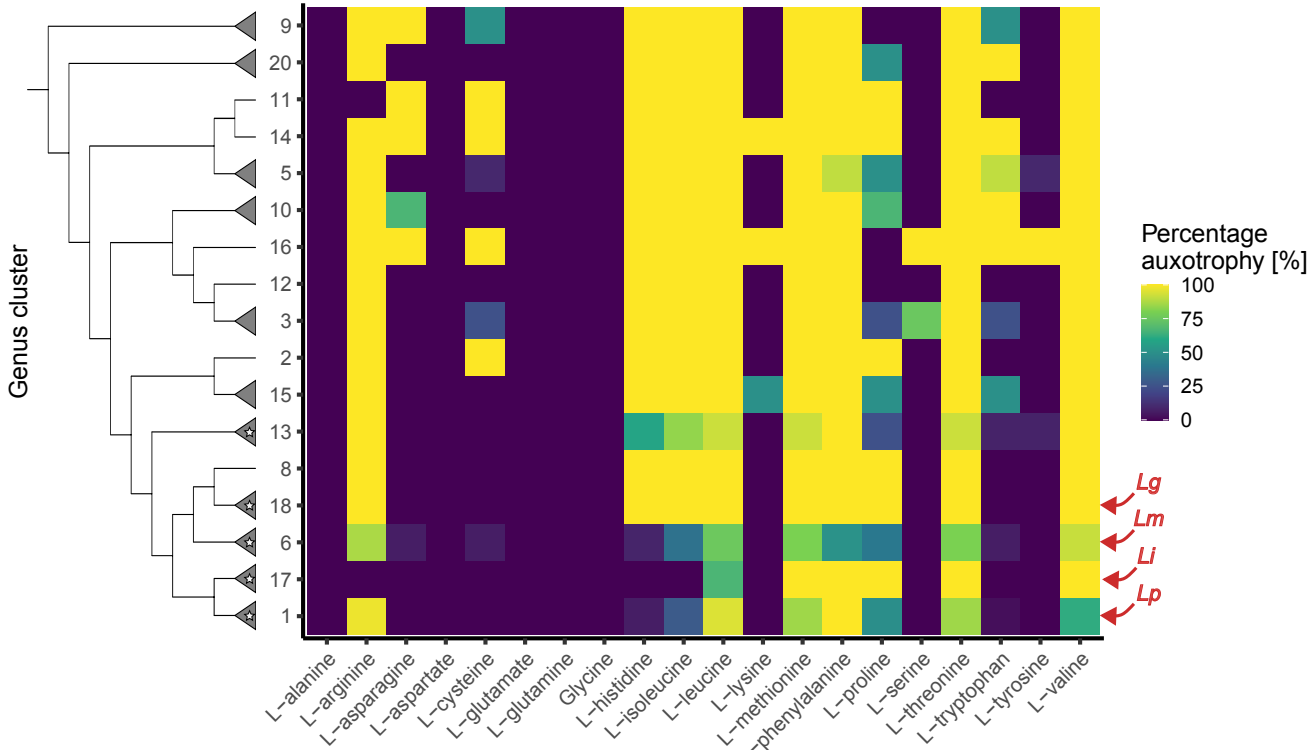
143. J. R. George, L. Pine, M. W. Reeves, W. K. Harrell, Amino acid requirements of Legionella pneumophila. *J Clin Microbiol* **11**, 286–291 (1980).

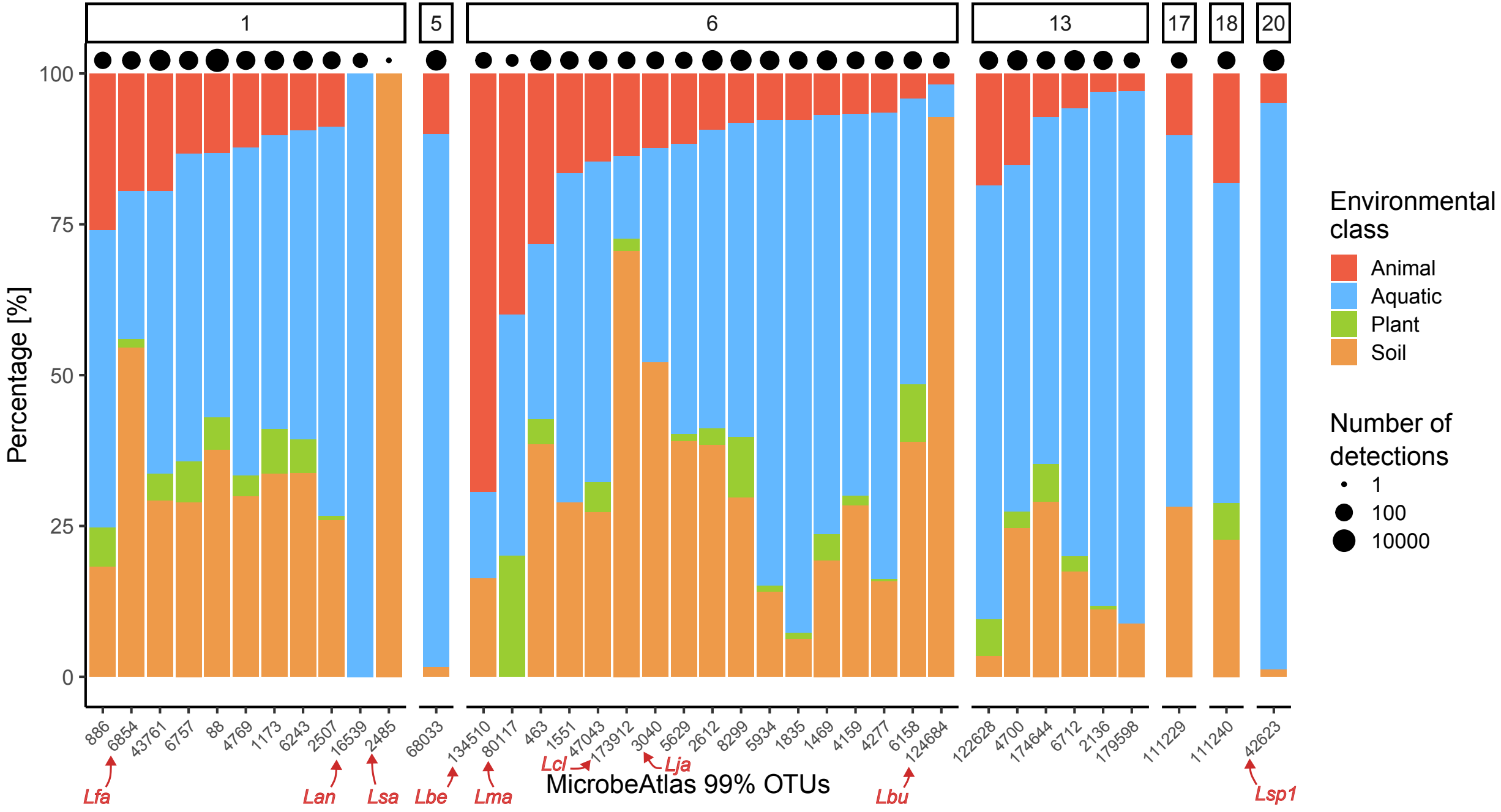
144. M. J. Tesh, R. D. Miller, Amino acid requirements for Legionella pneumophila growth. *J Clin Microbiol* **13**, 865–869 (1981).


1000
0
5
0
Number of unique gene families
Species
Genomovar
Estimated genome size [Mb]
Genus cluster
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
Li
Lp


1
5
6
13
17
18
20
Percentage [%]
100
75
50
25
0
886
6854
43761
6757
88
4769
1173
6243
2507
16539
2485
68033
134510
80117
463
1551
47043
173912
3040
5629
2612
8299
5934
1835
1469
4159
4277
6158
124684
122628
4700
174644
6712
2136
179598
111229
111240
42623
Lfa
Lan
Lsa
Lbe
Lma
Lcl
Lja
Lbu
Lsp1
MicrobeAtlas 99% OTUs
Environmental class
Animal
Aquatic
Plant
Soil
Number of detections
1
100
10000