Global climate gradients structure soil *Legionella* diversity, relative abundance and pathogen distributions

Hans W. Singh[1], Mengting Maggie Yuan[1]

[1]Pacific Biosciences Research Center, University of Hawai'i at Mānoa, Honolulu, HI, 96822, USA

Keywords: *Legionella*, Legionellosis, soil-borne pathogen, biogeography

Abstract

Legionnaires' disease has increased sixfold since 2000 in the United States, yet environmental reservoirs such as soils remain understudied sources of infection. It remains unclear how the diversity and abundance of environmental pathogens, such as *Legionella*, will shift as the climate continues to change. Here, we utilize 4,287 publicly available global soil 16S rRNA gene amplicon datasets to analyze how the distribution of the genus *Legionella*, including pathogenic lineages, varies across continental-scale biogeographic regions and along climatic and geochemical gradients. *Legionella* relative abundance increased across the joint precipitation-air temperature gradient, and mixed-effects modeling indicated that the chance of detecting *Legionella* in soils increased substantially when annual precipitation reached ~500 mm. *Legionella* diversity was characterized by high spatial turnover, with the majority of abundant ASVs restricted to individual regions despite broad phylogenetic representation across the genus. Further, fewer than 2% of our 16S-derived *Legionella* sequences matched ASVs corresponding to characterized species, and the clinically dominant species *L. pneumophila* was rarely detected. In contrast, ASVs matching several other known pathogens (*L. longbeachae, L. anisa, L. bozemanae, L. cincinnatiensis*) were orders of magnitude more common than *L. pneumophila* in soils. Moreover, pathogenic *Legionella* ASVs were detected significantly more frequently at warmer and wetter conditions. Collectively, our findings suggest that *Legionella* relative abundance and diversity in soils, especially pathogenic lineages, are shaped by both dispersal and climatic filtering. As temperature and precipitation regimes shift in the future, our results imply that soils should be an important pathogen source to monitor and that *Legionella* species beyond *L. pneumophila* warrant increased ecological and public health attention.

Introduction

The gammaproteobacterial genus *Legionella* is the causative agent of Legionnaires' disease, a severe, pneumonia-like pulmonary infection that has increased sixfold since 2000 in the United States (1). Currently, the rate of Legionnaires' disease is over 2.4 cases per 100,000 people in both the U.S. and Europe (2-3), and it carries a high economic burden ($800 million annually) (4), hospitalization rate (95%) and case fatality rate (10%) (2). *Legionella* can also cause the milder, flu-like Pontiac fever, and together these two illnesses are collectively referred

to as legionellosis (1). *Legionella* has tight evolutionary relationships with free-living amoebae and other protozoan hosts, which function as analogous systems to human alveolar macrophages. Mechanistically, *Legionella* utilize type I, II, and IV secretion systems to deliver a large suite of effector proteins (~300) that hijack host cellular functions and enable intracellular survival (1,5-8). While *Legionella* spp. have evolved the capacity to replicate intracellularly, they can also exist as free-living or within biofilms, and are commonly detected in freshwater systems (1,9). Despite the widespread environmental occurrence of *Legionella* (10), the climatic, geochemical, and ecological factors that shape their global distribution and diversity across natural reservoirs, such as soils, remain poorly understood.

Following its first identification associated with cooling towers (11), *Legionella* has been intensively studied as a human pathogen in built water systems (12), but the ecology of the genus in soil and freshwater remains far less clear. Typical *Legionella* infections follow inhalation of aerosolized water from contaminated built systems like cooling towers and showers (12), and usually involve *L. pneumophila* (80% of cases in the U.S. and Europe) (1-3). Direct exposure to *Legionella* from environmental reservoirs such as soils remains understudied as a source of legionellosis, although soil-associated *Legionella* have been shown to reach humans through aerosolization or hydrological transport (13). It is important to note that 95% of legionellosis cases in the U.S. are diagnosed using urinary antigen tests, which, while rapid and convenient, are limited to detecting *L. pneumophila* serogroup 1 (14-15). Notably, 23 of the 73 characterized *Legionella* species have been implicated in human infections, highlighting a significant gap in diagnostic sensitivity and species-level resolution (16-38). Legionellosis incidence over the past few decades is strongly correlated with precipitation, temperature and humidity (39-40). These observations are supported by increased infection rates in summer and fall months (2-3) when precipitation and temperature peak in most climate types. Sporadic, community-acquired cases without clear sources are also rising, with 66% of recent cases not linked to healthcare, travel or senior living (2). As cases grow despite improved disinfection and surveillance, environmental reservoirs beyond built systems should be considered as transmission sources. In Australasia, over half of reported cases are caused by *L. longbeachae*, a species traced to potting soils (1,41), implying a potentially significant contribution of infections not associated with built water systems.

L. pneumophila dominates genome sequencing efforts (79.7% of complete *Legionella* genomes in the NCBI database) (42), yet their well-studied physiology and pathogenesis do not necessarily apply to other species. For example, *L. longbeachae* uses a different intracellular mechanism and only shares 30% of its effector proteins with *L. pneumophila* (43-44). Effector diversity is high and largely species-specific, with eukaryotic-like domains commonly seen across the genus, reflecting cases of gene acquisition from their hosts (45). While *Legionella* effectors constitute a taxonomically diverse virulence factor set, implying varied host interactions, most remain poorly understood (46). As non-*pneumophila* infections are likely underrecognized, their risks under global change are highly uncertain. It remains unclear how the diversity and abundance of environmental pathogens, including *Legionella*, will shift as the global climate continues to change. Environmental reservoirs such as soils may be significant yet underrecognized sources of infection. Soils in particular remain poorly represented in culture-based and genomic surveys of *Legionella* diversity, limiting our understanding of legionellosis risk (46-47). Understanding the global diversity and relative abundance of *Legionella* species is an essential first step to assess their potential links to the rising incidence of legionellosis.

Over the past two decades, the acceleration in genomic, metagenomic, and high-throughput amplicon sequencing capabilities have revolutionized our ability to perform large-scale assessments of microbial diversity (48-50). In conjunction, the development of large data repositories such as NCBI and JGI has enabled the accumulation of microbiological datasets that are often linked to accompanying metadata (42,51). From a sequencing perspective, *Legionella* is a rare taxon in natural environmental systems, making it difficult to recover in metagenomic bins. For example, of the nearly 350,000 high-quality and medium-quality environmental metagenomic bins within the JGI database, only 109 (0.03%) belong to the genus *Legionella*, highlighting the challenges of using metagenomic sequencing to assess the global diversity of rare taxa (51). Complete genomes across a breadth of strains can also be challenging to compare due to biases in target samples (towards infected patients) and cultivation techniques (towards *L. pneumophila*-specific media) (15,42,47). High-throughput targeted amplicon sequencing of published *Legionella*-specific primers is also insufficient for a global comparison of *Legionella* diversity as they are rarely utilized over large geographic distances, primers are biased towards clinical species, and infrequently targeted in soil samples (52). The most widely used 16S rRNA gene primer set to date (515F-806R) optimizes broad prokaryotic taxonomic coverage and is utilized in large-scale analysis of microbial diversity across every continent (12,48). These attributes make it optimal for a global analysis, with large numbers of samples and high sequencing depth compensating for the rarity of *Legionella* reads.

Nonetheless, taxonomic classification of 16S reads to rare members of the microbiome remains a challenge (53-54). Utilization of both k-mer based and identity-based classifiers depend on existing reference sequences, but the sparse and *L. pneumophila*-skewed representation of *Legionella* in 16S databases such as SILVA, Greengenes and GDTB limits accurate detection (42,55-56). When a query sequence lacks a close match in the database, classifiers default to higher taxonomic levels (e.g., class or order), thereby masking genus- or species-level diversity (54-55). A further limitation within the sequencing pipeline is that rare taxa like *Legionella* often resolve into singleton amplicon sequence variants (ASVs), which are often removed during quality filtering (57-58). Phylogenetic insertion offers the most accurate approach for assigning taxonomy from short 16S amplicons, particularly for rare taxa, as they place sequences on an evolutionary backbone rather than relying on k-mer frequencies or pairwise identity (59-61). These methods are computationally intensive and impractical for analysis of complete 16S datasets (59), but become more tractable when applied to targeted analyses of a single genus, such as *Legionella*.

Here, we analyze the global soil ecology of *Legionella* using 4,287 publicly available 16S rRNA amplicon datasets spanning diverse biogeographic regions and environmental gradients. Specifically, we identify environmental predictors of *Legionella* abundance, characterize patterns of diversity and biogeographic turnover, and assess the prevalence of ASVs corresponding to known pathogenic *Legionella* species.

Results

Two-step taxonomic approach accurately classifies *Legionella* sequences

To maximize the recovery, speed and accuracy of *Legionella* detection, we implemented a two-step taxonomic classification pipeline in which 16S amplicon sequences (the first 150 bp following the 515F primer) were first screened against a custom BLAST search, after which

candidate *Legionella* hits were inserted into a backbone reference phylogeny built from representative and complete genomes. Analysis using all Gammaproteobacteria sequences within the Silva database highlights the accuracy of our approach - every *Legionella* sequence (n=1586) in SILVA was correctly detected and placed within the *Legionella* clade, and no non-*Legionella* Gammaproteobacteria sequences (n=16,926) were phylogenetically verified as *Legionella* sequences (i.e., no false positives). As an additional validation to our pipeline, we sequenced a mock community consisting of DNA from *L. pneumophila*, *L. longbeachae* and *L. micdadei*, and as expected found these to be the only three ASVs recovered.

Relative abundance of *Legionella* sequences across climatic and biogeochemical gradients

In total, 4287 globally distributed soil 16S rRNA amplicon sequencing samples were analyzed, representing eleven distinct BioProjects (58-73) across eight distinct geographic locations: North America, Europe, Africa, Australia, Hawaii, Iceland, China, and Chile (Fig. 1a-b). The bioprojects we selected broadly fell into one of two categories, with the first comprising microbial surveys carried out across large biogeographical ranges that span substantial climatic and biogeochemical gradients. Our second category corresponded to bioprojects from fixed geographic locations that imposed controlled environmental manipulations: temperature in Iceland, precipitation in Iowa and both temperature and precipitation in Oklahoma. Across our entire dataset, the average sequencing depth was 56,960 sequences with 22.76 *Legionella* sequences recovered (0.039%). Despite *Legionella* being a rare taxon, the overwhelming majority of our samples (93.0%) contained at least one *Legionella* read, highlighting that *Legionella* are a consistent member of soil microbial communities.

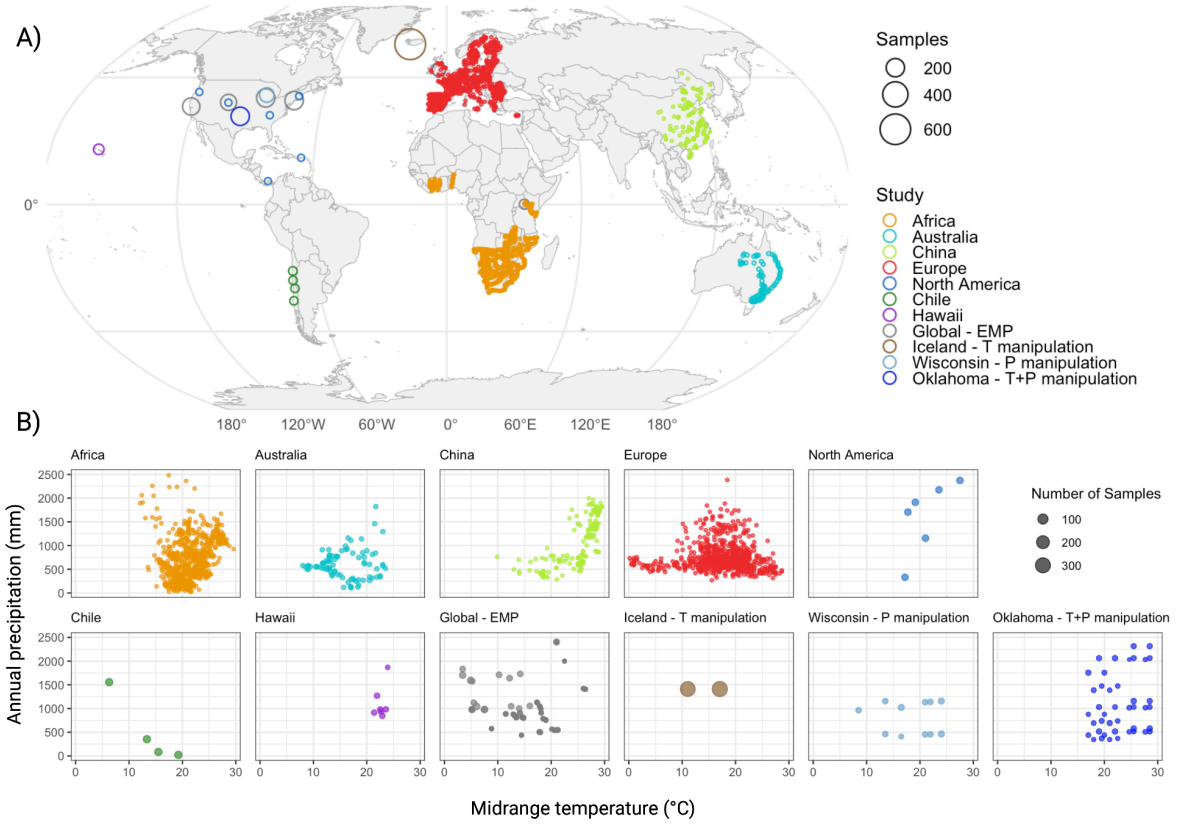

Fig. 1 - Samples selected for global analysis. A) Geographic distribution of soil samples used in this study. Colors indicate sequencing projects and point size corresponds to the number of samples at each location. B) Climatic ranges for each sequencing project. Points show individual sampling locations plotted by midrange temperature (°C) during the sampling month (x-axis) and annual precipitation (mm) (y-axis). Panels correspond to sequencing projects, and point size indicates number of samples.

From each sample, the following metadata were collected as indicators of the climatic and biogeochemical conditions that *Legionella* was recovered from: midrange monthly temperature, annual precipitation, soil pH, total organic carbon, and total organic nitrogen. When we analyzed how *Legionella* relative abundance varied across climatic and biogeochemical variables, Spearman correlations revealed weak but significant monotonic relationships (Fig. S1) with all environmental variables ($p<10^{-8}$). The strongest associations were negative with elevation ($\rho = -0.24$) and positive with soil carbon content ($\rho = 0.24$). Annual precipitation ($\rho = 0.2$) and soil nitrogen content ($\rho = 0.18$) displayed moderate positive correlations, while pH ($\rho = -0.13$) and temperature ($\rho = 0.09$) exhibited weaker correlations. Although statistically robust due to large sample sizes, all correlations were small in magnitude, indicating that *Legionella* abundance is only weakly structured by any single climatic or biogeochemical gradient.

Since Spearman correlations only quantify monotonic trends and do not capture non-linear or threshold responses, we next evaluated differences in *Legionella* relative abundance across discrete environmental categories (Fig. 2a, Fig. S2). Each environmental variable was partitioned into five buckets representing the full range of metadata, and Kruskal-Wallis tests revealed highly significant differences among buckets for all six variables (all $p < 10^{-15}$). Dunn's post hoc comparisons show clear, structured shifts in abundance across these buckets. For temperature, *Legionella* abundance was significantly higher in the 25-30°C range compared to all other temperature ranges (Fig. 2a). Precipitation showed a strong threshold response: the driest sites (0-400 mm) had significantly lower abundance than the other four wetter categories, which formed a single statistically homogeneous group (Fig. 2a). Soil carbon exhibited a graded increase in abundance with each successive bucket (0-1% < 1-2%, 2-3% < 3-4% < 4+%), producing four distinct groups (a-d) (Fig. S2). Nitrogen content followed a similar positive pattern with low-N soils (0-0.1%), moderate-N (0.1-0.2%, 0.2-0.3%, 0.3-0.4%) and high-N (0.4+%) soils falling into three distinct groups (a-c) (Fig. S2). For pH, abundance was highest under strongly acidic conditions (<5), decreased across intermediate pH levels (5-8) and was lowest in alkaline soils (8+), forming three groups (a-c) (Fig. S2). Lastly, elevation strongly structured *Legionella* abundance, which displayed a clear low elevation optimum (0-200 m.), declined at middle elevations (200-400 m.) and displayed minimum abundance at higher elevations (400-600 m., 600-800 m., 800+ m.) (Fig. S2). Collectively, these categorical patterns reveal a mix of gradual monotonic and threshold responses, suggesting that depending on the environmental driver, *Legionella* abundance can be shaped by both sharp environmental transitions and smooth continuous gradients.

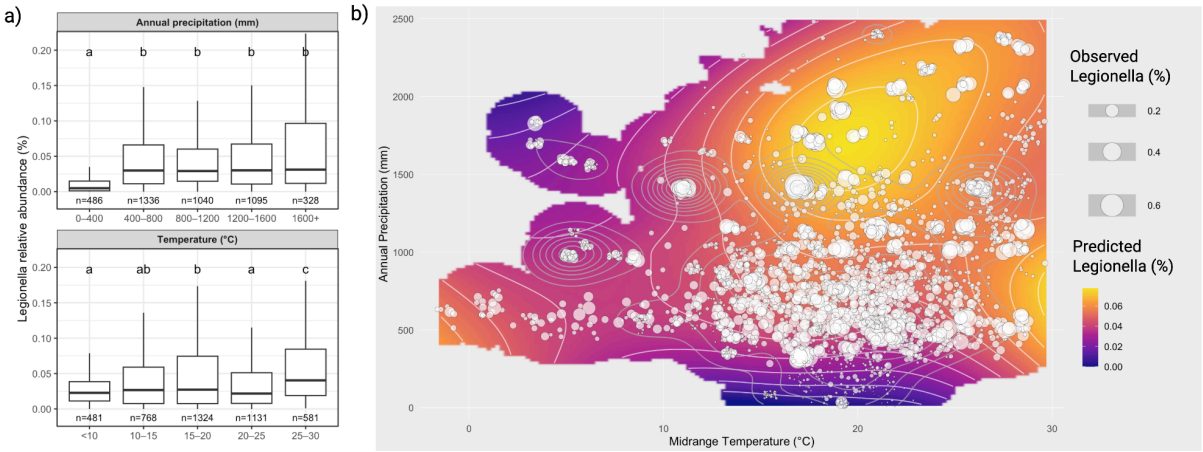

Fig. 2 - Climatic controls on *Legionella* relative abundance (a) Relative abundance of *Legionella* sequences across climatic ranges. Samples are grouped into five bins for annual precipitation (upper panel) and midrange temperature during the sampling month (lower panel). Boxplots show the distribution of *Legionella* relative abundance within each bin. Letters indicate statistically significant differences among groups (Kruskal-Wallis with Dunn's post hoc test). (b) Relative abundance of *Legionella* sequences across the joint temperature-precipitation gradient. Circles indicate observed *Legionella* relative abundance. The colored surface represents LOESS-predicted *Legionella* relative abundance, with contour lines indicating predicted abundance levels.

As climatic data are the easiest to monitor at scale, we next queried the interactive effects of climate on *Legionella* relative abundance by analyzing the two-dimensional environmental space of temperature and precipitation. A bivariate temperature-precipitation response surface revealed clear interactive structure, with the highest predicted *Legionella* relative abundances occurring under both warm conditions and high precipitation (Fig. 2b). A two-dimensional GAM (generalized additive model) revealed a highly significant joint temperature-precipitation effect (EDF = 21.7, $p<10^{-15}$) and significantly improved model fit over a model with independent smooth functions for temperature and precipitation (ANOVA: F=2.82, p=0.012), confirming that *Legionella* relative abundance is shaped by nonlinear interactive effects of climatic conditions.

To evaluate the joint effects of all climatic and biogeochemical variables while accounting for site-to-site variation, we used a hurdle mixed-effects modeling framework, consisting of a binomial mixed model for presence-absence and a beta mixed-effects regression for relative abundance conditional on presence, with study included as a random intercept. Detection probability of *Legionella* sequences increased significantly with only annual precipitation (Fig. S3), whereas no other variables were associated with changes in presence of *Legionella*. In contrast, conditional relative abundance increased significantly with temperature and soil carbon content, and declined with increasing pH and soil nitrogen content (Fig. S3). Notably, annual precipitation showed no association with relative abundance. Together, these results indicate that moisture availability primarily governs *Legionella* occurrence, while temperature and local biogeochemical variables govern *Legionella* relative abundance. Further, holding all other climatic and biogeochemical variables constant and accounting for study-to-study variation altered the inferred effects of certain predictors, such as nitrogen content (positive in isolation, negative in joint model) and precipitation (positive in isolation, neutral in joint model).

Legionella relative abundance at experimentally manipulated sites

Experimental field manipulations are powerful tools for identifying predictors of *Legionella* abundance, as they selectively alter target variables while keeping all other variables constant. Analysis of 16S data from a geothermally-warmed site in Iceland (+6°C) showed that *Legionella* relative abundance was significantly higher than in neighboring ambient sites ($p<0.01$) (Fig. S4). Similarly, analysis of rainfall-manipulation sites in Wisconsin revealed that *Legionella* increased with volumetric water content ($r = 0.39$, $p < 0.01$) (Fig. S5). Finally, at a site combining precipitation and temperature manipulations in Oklahoma, *Legionella* relative abundance was positively correlated with annual moisture ($r = 0.32$, $p < 0.001$) and negatively correlated with annual temperature ($r = -0.34$, $p < 0.001$) (Fig. S6). These results highlight the role of environmental selection in *Legionella* distribution.

Legionella diversity across biogeographic, climatic and biogeochemical gradients

Non-metric multidimensional scaling (NMDS) based on Bray-Curtis dissimilarities revealed weak but significant regional structuring of *Legionella* community composition (Fig. S7). Although global samples largely overlapped in ordination space, Chilean samples exhibited a pronounced shift along NMDS1 and greater dispersion than other regions, indicating elevated within-region heterogeneity. Permutation-based multivariate analysis of variance (PERMANOVA) confirmed a significant effect of region on *Legionella* community composition ($r^2 = 0.02$, $p < 0.001$). Permutation-based fitting of environmental variables onto the NMDS ordination identified pH as the only factor significantly correlated with Legionella community composition at the global scale (envfit, $r^2 = 0.0043$, $p = 0.043$) (Fig. S8). While other environmental factors were not significant predictors, binned ordinations suggest gradual, directional turnover in community composition across the midrange temperature, soil carbon and soil nitrogen gradients (Fig. S8). For these variables, lower-value bins transition directionally across ordination space toward higher-value bins rather than forming discrete clusters, which is consistent with continuous, non-threshold responses to environmental variation.

To further resolve regional structuring within the dominant members of the *Legionella* community, we examined the 300 most abundant *Legionella* ASVs within each of our 8 regions. The majority of dominant *Legionella* sequences (70.0%) were only detected in one region, with just 11.5% observed in 4+ regions and none detected in all 8 regions (Fig. 3a). The proportion of region-restricted dominant taxa varied among regions, with isolated regions such as Chile (84.6%), Hawaii (81.3%) and Iceland (74.0%) displaying high levels of regional specificity, whereas Europe (49.3%) and North America (59.0%) showed less (Fig. 3a). Pairwise comparisons showed limited overlap between Chile and any other region (1-8%), and limited overlap was also observed for paired regions with vastly different climates: Iceland/Africa (4%), Iceland/Hawaii (4%) and Iceland/Australia (6%) (Fig. S9). Regions with less geographic distance or similar climates contained the highest overlap - North America/Europe (30%), Iceland/Europe (20%) and China/Europe (18%) (Fig. S9). Indeed, mapping climatic differences onto these pairwise patterns revealed that overlap in dominant *Legionella* ASVs declined with increasing environmental dissimilarity (Fig. 3b), with the percentage of shared ASVs negatively correlated with increasing differences in both annual precipitation (Spearman $\rho = -0.45$, $p=0.017$) and temperature ($\rho = -0.37$, $p=0.05$) between regions. Collectively, these results suggest that both geographic separation and climatic similarity are associated with the structure of dominant *Legionella* ASVs across regions.

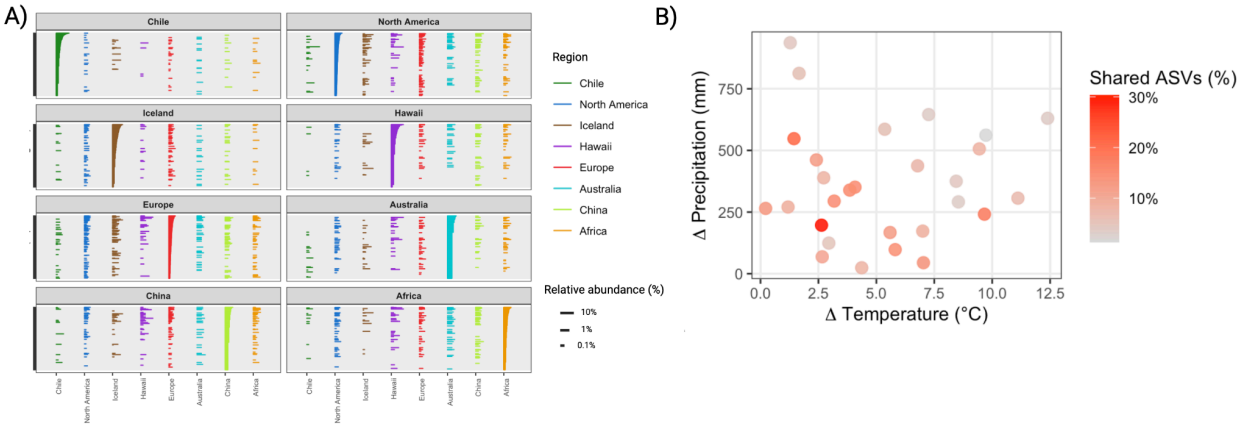

Fig. 3 - Regional distribution and climatic overlap of dominant *Legionella* ASVs (a) Distribution of the 300 most abundant *Legionella* ASVs for each region. Each panel shows the dominant ASVs from one region, with the length of the bar indicating relative abundance, and colors indicating the presence of those ASVs across all 8 regions. (b) Climatic similarity and overlap of dominant *Legionella* ASVs among regions. Each point represents a pairwise comparison between regions, with axes showing differences in midrange temperature (°C) and annual precipitation (mm). Point color indicates the percentage of shared dominant ASVs between regions.

Finally, to determine whether the climatic and biogeographic-associated turnover in dominant *Legionella* ASVs corresponded to deep phylogenetic divergence, we constructed a maximum-likelihood phylogeny using IQ-TREE based on the 300 most abundant ASVs from each of the eight regions. Environmental ASVs spanned multiple deeply branching lineages and were broadly interspersed with both cultured and pathogenic *Legionella* reference sequences, indicating that abundant environmental *Legionella* represent a wide range of evolutionary lineages rather than a restricted subset (Fig. 4). Pathogenic *Legionella* references were distributed across multiple phylogenetic lineages rather than forming a single monophyletic group, consistent with phylogenetically widespread pathogenic potential within the genus. Although ASVs from different biogeographic regions did not segregate into discrete, region-specific clades and instead co-occurred throughout the phylogeny, null-model analysis revealed significant phylogenetic structure within several regions. Net Relatedness Index (NRI) values indicated significant deep phylogenetic clustering in North America, Iceland, Europe, China, and Africa, whereas Chile alone (lowest average midrange temperature and annual precipitation) exhibited strong deep overdispersion (Fig. S10). Nearest Taxon Index (NTI) values further revealed significant tip-level clustering in most regions, including Europe, Iceland, Africa, North America, Australia, and Chile (Fig. S10). These results indicate that while dominant environmental *Legionella* lineages are broadly distributed across the genus, climatic and biogeographic filtering operate within regions to structure communities at both deep (NRI) and fine (NTI) phylogenetic scales, with substantial heterogeneity among regions.

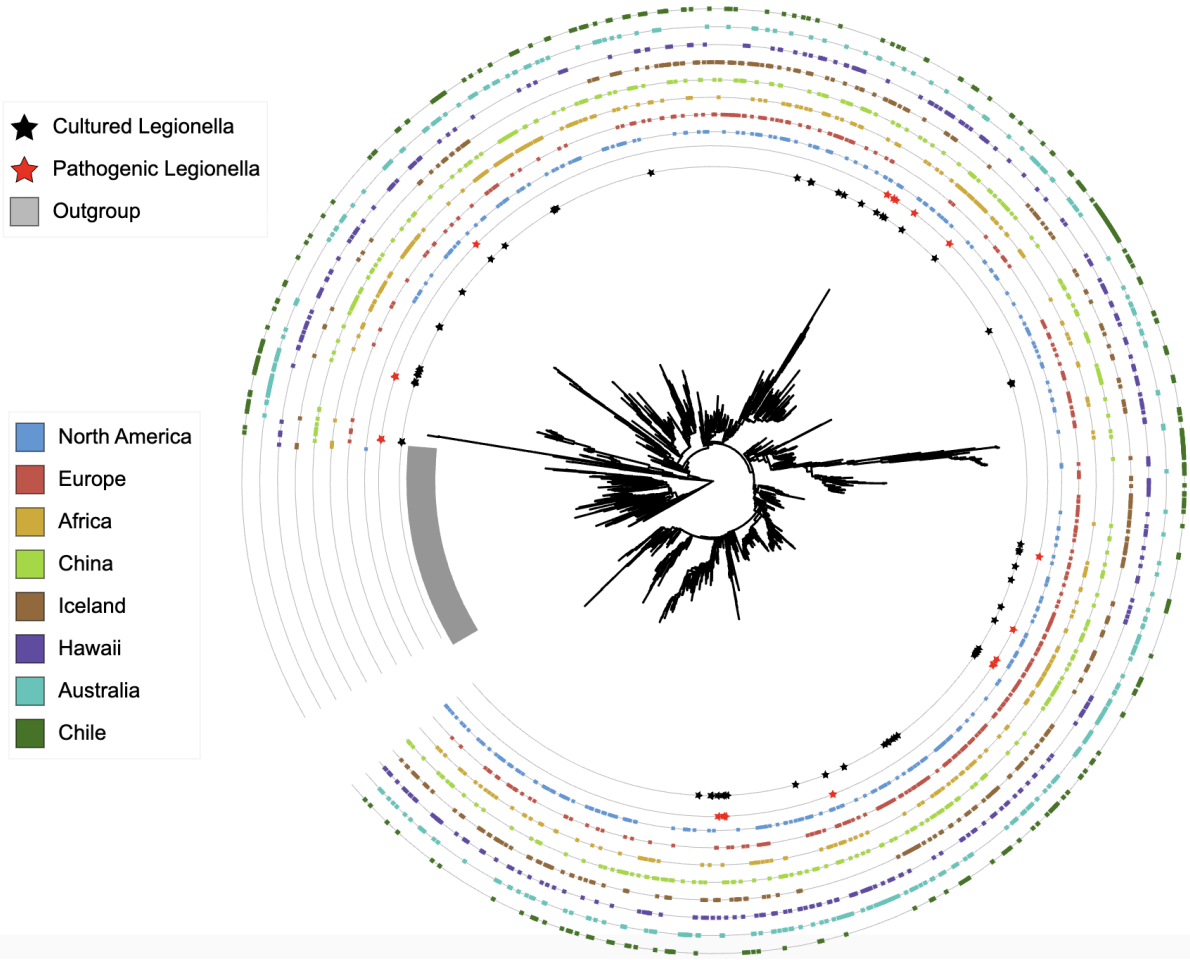

Fig. 4 - Phylogeny of dominant *Legionella* ASVs. The 300 most abundant *Legionella* ASVs from each region were placed into a maximum-likelihood phylogeny. Stars on the innermost rings indicate ASVs matching representative or complete genomes of cultured *Legionella*, with red stars marking human pathogens. Colored outer rings indicate the presence of each ASV among the dominant *Legionella* ASVs in each region.

Spatial structuring of *Legionella* diversity

To evaluate spatial structuring of *Legionella* community composition at a finer scale, we quantified pairwise beta diversity among samples using Bray-Curtis dissimilarity calculated from *Legionella*-only ASV profiles and examined its relationship with geographic distance. Community similarity declined systematically with increasing geographic distance, exhibiting a significant distance-decay relationship (Fig. S11). Fitting a power-law model to our data demonstrated a strong negative scaling between similarity and distance ($\beta = -0.134$, $p < 0.001$), indicating that *Legionella* communities become progressively more dissimilar as spatial separation increases. Multiple regression on distance matrices (MRM) indicated that geographic distance remained a significant predictor of community dissimilarity after accounting for climatic and biogeochemical differences among sites ($p < 0.001$).

Community context of *Legionella*-rich microbiomes

To identify microbial community features associated with *Legionella*-rich samples, we performed a differential abundance analysis using 97% OTUs. Samples were stratified into

upper (>0.062%) and lower (<0.009%) quartiles of *Legionella* relative abundance, and in total 114 OTUs were enriched in *Legionella*-rich samples and 290 OTUs were enriched in *Legionella*-poor samples ($|\log_2$ fold change$| \geq 2.5$, FDR < 0.01) (Fig. S12). OTUs enriched in *Legionella*-rich samples were most frequently classified as Bacteroidota (35.1%) and Acidobacteria (27.2%) (Fig. S12). In contrast, the sharpest compositional shift was observed for Actinomycetota, which comprised 30.7% of OTUs enriched in *Legionella*-poor samples but only 3.5% of OTUs enriched in *Legionella*-rich samples (Fig. S12). Notably, many Actinomycetota OTUs enriched in *Legionella*-poor samples were classified to the family Rubrobacteria (33.7%), a lineage commonly associated with arid soil environments. Differentially abundant OTUs were broadly prevalent within their respective groups, with median prevalence of 40.7% among OTUs enriched in *Legionella*-poor samples and 38.4% among OTUs enriched in *Legionella*-rich samples, indicating community-wide shifts rather than effects driven by rare taxa.

Relative abundance and diversity of *Legionella* ASVs from known pathogens

In total, 257 *Legionella* genomes in the NCBI database are annotated as either reference-quality or complete, representing 74 distinct species. When we clustered the first 150 bp following the 515F forward primer, hereafter referred to as the target 16S region, we found that this region resolves 51 species into unique (non-shared) ASVs (Fig. 5a). Only six ASVs were shared among multiple species, with one ASV representing twelve different species - *L. anisa*, *L. cherii*, *L. dumoffii*, *L. gormanii*, *L. longbeachae*, *L. parisiensis*, *L. qingyii*, *L. resiliens*, *L. santicrucis*, *L. sheltonii*, *L. sp 227* and *L. sp PC1000* (Fig. 5a). Of the 23 recognized pathogenic *Legionella* species, 12 were represented by unique ASVs (known pathogen ASVs), whereas the remaining 11 clustered into ASVs shared with at least one species not confirmed to be pathogenic (potentially pathogenic ASVs) (Fig. 5a). It is important to note that *Legionella* species are designated as pathogenic based on culture-confirmed human infections, a criterion that almost certainly underestimates pathogenic potential across the genus.

When we searched our dataset for pathogenic or potentially pathogenic *Legionella* ASVs (pASVs), the most abundant pASV matched two pathogenic species (*L. bozemanae* and *L. cincinnatiensis*), whereas the next most abundant pASV matched six pathogenic species (*L. anisa*, *L. cherrii*, *L. dumoffii*, *L. gormanii*, *L. longbeachae* and *L. parisiensis*), accounting for 0.48% and 0.43% of total *Legionella* sequences, respectively (Fig. 5b). Across all *Legionella* ASVs from our global analysis, these two pASVs ranked as the second and fourth most abundant sequences overall (Fig. S13). Consistent with their high global abundance, both pASVs were consistently detected across six of our eight geographic regions (>0.1% of *Legionella* sequences), with Chile and Iceland as exceptions. The only other pathogenic or potentially pathogenic *Legionella* ASVs recovered were those matching to *L. tucsonensis* (0.11%), *L. sainthelensi* (0.03%), *L. micdadei* (0.03%), *L. feeleii* (0.02%), *L. oakridgensis* (0.01%), *L. wadsworthii* (0.01%) and *L. pneumophila* (0.01%) (Fig. S5b).

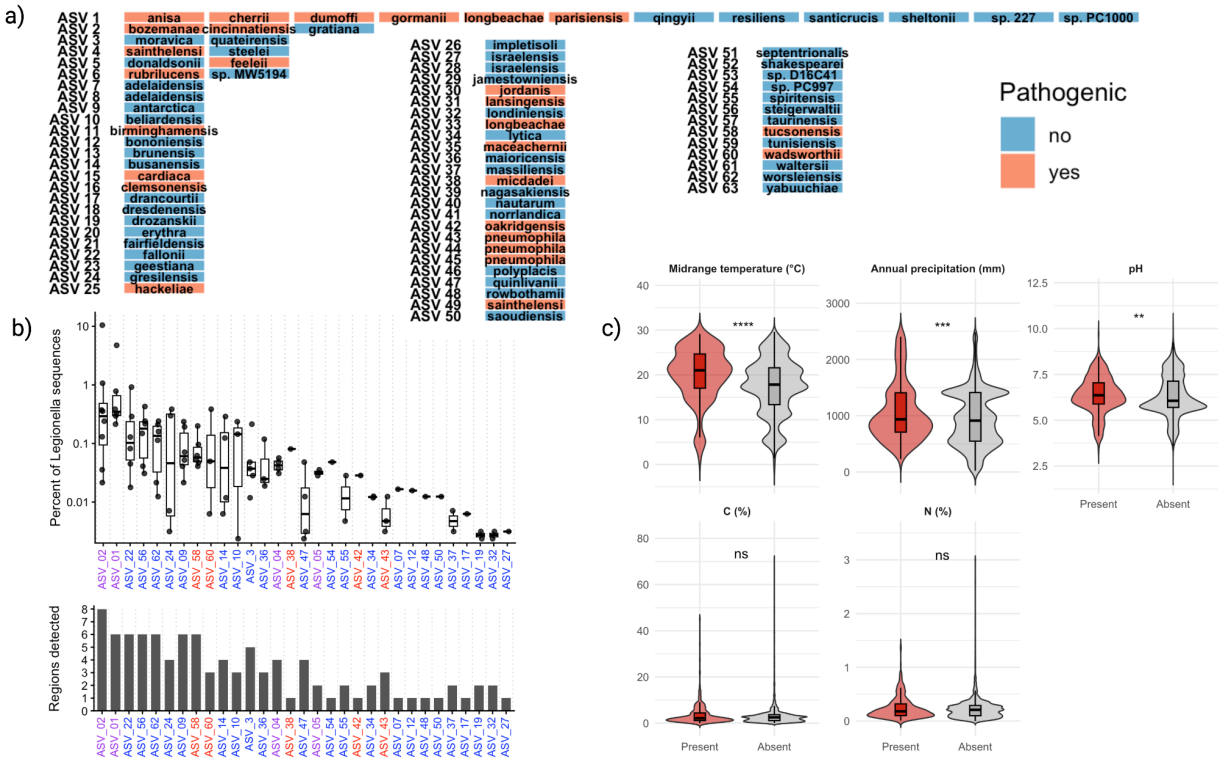

Fig. 5 - Distribution and environmental associations of characterized *Legionella* ASVs. (a) Clustering of characterized *Legionella* species based on the first 150 bp following the 515F primer of the 16S rRNA gene. Species sharing identical sequences are grouped into the same ASV and colored according to whether the species is a known human pathogen. (b) Relative abundance and geographic distribution of characterized *Legionella* ASVs. The upper panel shows the percentage of total *Legionella* sequences in each region that mapped to a given characterized ASV. Points represent individual regions and boxplots summarize the distribution across regions where the ASV was detected. X-axis labels are colored by pathogenic status (red = pathogenic, purple = potentially pathogenic, blue = non-pathogenic). ASV numbers correspond to those shown in panel (a). The lower section of panel (b) shows the number of regions in which each ASV was detected. (c) Environmental distribution of pathogenic or potentially pathogenic *Legionella* ASVs. Violin plots show the distribution of samples in which pathogenic ASVs were present (red) or absent (grey) across climatic and biogeochemical variables.

To evaluate whether the presence of pathogenic *Legionella* was associated with environmental gradients while accounting for detection bias, we compared samples in which *Legionella* pASVs were present or absent across our environmental variables and fitted a logistic regression model including sequencing depth as a covariate. Our results illustrate that pASVs tended to occur significantly more often at higher midrange temperatures and under wetter climate conditions (Fig. 5c). Consistent with these patterns, logistic regression revealed that after controlling for sequencing depth, both midrange temperature and annual precipitation remained significant predictors of *Legionella* pASVs ($p < 10^{-5}$). Model coefficients indicate that a 10°C increase in midrange temperature was associated with a 55% increase in the odds of detecting pathogenic *Legionella* ASVs, and a 500 mm increase in precipitation was associated with an approximately 45% increase in detection odds. In contrast, pH, soil carbon, and nitrogen content were not significant predictors of pathogenic *Legionella* ASVs after accounting for sequencing depth.

Discussion

By leveraging 4,287 publicly available global soil 16S rRNA datasets coupled with an efficient rara-taxa detection method based on targeted phylogenetic insertion, we found that the rare (only 0.039% of the overall soil microbiome) but widespread (present in 93% of soil samples) gammaproteobacterial genus *Legionella* exhibits predictable environmental and biogeographic patterns across global soils. Warmer and wetter environments were consistently associated with higher relative abundances of *Legionella* and increased probabilities of detecting pathogenic *Legionella* ASVs, highlighting the role of climate in structuring environmental pathogen reservoirs. Consistent with these patterns, *Legionella* community composition also exhibited continental-scale structuring, with dominant ASVs showing strong geographic specificity and decreasing overlap among regions with more distinct climates. Notably, the most clinically significant species, *L. pneumophila* (1), was rare in soils, whereas the soil-associated *L. longbeachae* was expectedly among the most abundant *Legionella* ASVs detected, providing empirical support for the long-standing view that different pathogenic *Legionella* species occupy distinct environmental reservoirs (41). Together, these findings suggest that *Legionella* persistence in soils is governed by a set of ecological filters: geographic distance and climate influences which lineages dominate regionally within the global species pool; soil water availability determines establishment, and temperature and local biogeochemical conditions regulate population growth.

Climatic influence on *Legionella* presence and relative abundance in global soils

Mapping the ecological niches of environmental pathogens is increasingly important as climate change alters environmental conditions. The combination of extremely low relative abundance yet near-ubiquitous detection across soils suggests that *Legionella* behave as a member of the microbial “rare biosphere” (74), persisting at low abundance across diverse environments while responding opportunistically to favorable ecological niches. Such regulation of microbial growth and activity often reflects environmental controls, particularly by temperature and precipitation, which are strong drivers of the distribution and transmission of many human pathogens–from dengue viruses to *Leptospira* bacteria (75-76). Temperature governs biochemical process rates and often exerts selective pressure on human pathogens; precipitation influences air humidity, soil hydrology, solute movement, and nutrient deposition, which are all critical controls of microbial activity in the air, soil, and aquatic systems (77). In addition, surface air temperature and precipitation are regularly monitored globally with a high spatial resolution (78), and many of these datasets are publicly available, providing a pragmatic framework for disease risk assessment at scale.

Previous work showed that Legionnaires’ disease incidence is positively associated with warmer and wetter climate (79); our finding that *Legionella* relative abundance increased globally in soils along the joint temperature-precipitation gradient implies that increasing environmental load under these conditions may contribute to exposure and infection. Warmer and wetter soils may favor *Legionella* persistence through several mechanisms. Elevated soil moisture promotes biofilm formation and supports higher densities of free-living amoebae and other protist hosts that facilitate intracellular replication (9,72), while warmer temperatures increase microbial activity and host-pathogen interactions (80-82). Consistent with this biofilm- and host-mediated ecology, our mixed-effects modeling identified precipitation as the primary determinant of *Legionella* presence. In contrast, temperature and local biogeochemical variables were more strongly associated with changes in *Legionella* relative abundance, suggesting that different environmental factors govern establishment and population growth. This pattern is consistent with general principles of soil microbial ecology, in which moisture availability constrains habitat suitability while temperature and nutrient availability regulate

metabolic rates and microbial growth (83-84). Notably, a previous global analysis showed that relative abundance of the order *Legionellales*, including multiple genera within the families *Legionellaceae* and *Coxiellaceae*, decreases with soil temperature (10). The discrepancy between this previous study and ours likely reflects differences in taxonomic resolution (order versus genus) as well as precipitation effects that were not considered in the previous analysis. As sporadic, community-acquired cases of legionellosis continue to rise (2), our findings collectively reinforce the possibility that environmental reservoirs such as soils may represent significant yet underrecognized sources of *Legionella* exposure.

Patterns in the broader soil microbiome further support the ecological context of *Legionella* persistence. *Legionella*-rich samples were enriched in Bacteroidota and Acidobacteria, groups commonly associated with soils containing higher organic matter availability and lower pH, respectively (85-86). This pattern aligns with our observation that *Legionella* relative abundance was higher in carbon-rich and more acidic soils. In contrast, Actinomycetota were strongly enriched in *Legionella*-poor samples; these taxa are commonly associated with drier soils and are well known for producing antimicrobial secondary metabolites (87-88), which may inhibit the presence of other microbial taxa, including *Legionella*. Under culture conditions, Pseudomonas has been shown to suppress *Legionella* growth through antimicrobial metabolites and quorum-sensing molecules (89), supporting the idea that microbial antagonism may also affect *Legionella* persistence in natural environments. However, bacterial-bacterial interactions involving *Legionella* in soil systems remain poorly characterized. Understanding interactions between *Legionella* and surrounding microbial communities may therefore be critical for determining the ecological conditions that promote or suppress its persistence, including potential competitive suppression, facilitation within biofilms, or indirect effects mediated through protist hosts.

The rapid expansion of sequencing technology and deposition of microbiome datasets in public repositories have enabled large-scale global analyses that were previously infeasible to conduct (90-91). However, as with many data reuse studies, this approach introduces potential biases, including uneven geographic sampling, variable sequencing depth and study-to-study methodological variation. Nonetheless, the large number of samples analyzed, consistency of these patterns with experimental field manipulations, and the use of hurdle mixed-effects models accounting for sequencing depth and study-level variation suggest that the broad climatic patterns observed here are robust. Although our analysis is based on relative abundance, previous studies have shown that warmer and wetter soils generally support higher microbial activity (81,83), suggesting that increases in *Legionella* relative abundance observed here may also correspond to elevated absolute abundance. However, confirming this will require future work using *Legionella*-specific primer sets and quantitative techniques such as qPCR or ddPCR to directly measure the absolute abundance of *Legionella* in soils.

Biogeography and climate structure global soil *Legionella* diversity

Multiple lines of evidence indicate that uncultured *Legionella* account for the vast majority of within-genus diversity (1,10,47), a conclusion reinforced by our observation that fewer than 2% of *Legionella* sequences matched ASVs corresponding to characterized species. This gap likely reflects limitations of cultivation approaches, as *Legionella* typically require specific growth conditions or host associations that are rarely reproduced *in vitro* (9,92-93). In addition, most isolation efforts and culture media have historically targeted clinical isolates or water-associated sources, leaving soil-associated species comparatively underexplored (15).

At a global scale, *Legionella* community similarity declined gradually across environmental gradients and geographic distance. This pattern is consistent with models of microbial biogeography in which environmental filtering and spatial processes jointly structure community composition (94). Furthermore, dominant lineages exhibited substantial regional differentiation, with the highest regional specificity occurring in geographically isolated areas such as Chile and Iceland. This pattern is consistent with studies of island and geographically isolated ecosystems, which suggest that dispersal limitation and historical contingency can influence which taxa become abundant (95-96). In addition, climatic conditions also shape which *Legionella* ASVs become regionally abundant: taxa were more frequently shared among regions with similar temperature and precipitation regimes, consistent with patterns resulting from environmental filtering (97). Dominant *Legionella* ASVs also showed phylogenetic clustering, indicated by high NRI and NTI values within several regions, further supporting homogeneous selection on phylogenetically conserved traits and resulting in assemblages dominated by closely related taxa (98).

A complementary explanation for the observed phylogenetic clustering is biotic filtering mediated by protist hosts. As *Legionella* replicate intracellularly within amoebae (9), host compatibility can be a major determinant of environmental persistence. If regional differences in amoebal community composition structure which *Legionella* lineages are able to proliferate, then local assemblages may reflect host-mediated selection acting within a broadly distributed evolutionary pool. Such dynamics could generate tip-level clustering without producing regionally endemic deep clades. Because global 18S rRNA gene datasets remain far less abundant than 16S datasets (90), and paired bacterial-protist datasets are still relatively scarce (99), the role of protist community structure in shaping *Legionella* biogeography remains underexplored and represents an important avenue for future research.

Distribution of potentially pathogenic *Legionella* in soils

Pathogenic or potentially pathogenic ASVs were consistently detected globally, and their enrichment under warmer and wetter conditions aligns with epidemiological evidence linking legionellosis incidence to increased temperature and precipitation. As such, environmental reservoirs such as soils may represent an underappreciated exposure pathway, particularly under conditions where soil disturbance can generate aerosols. Notably, the most clinically significant species, *L. pneumophila* (1), was rare in soils, whereas the ASV corresponding to *L. longbeachae* (47), a species commonly associated with soil environments, ranked among the more abundant *Legionella* lineages detected. Indeed, the prominence of *L. longbeachae* in soils is consistent with epidemiological patterns in Australasia, where potting soils are a major source of infection (47). The enrichment of pathogenic lineages under warmer and wetter conditions is also consistent with the expectation that species with higher infectivity may be better adapted to elevated temperatures and may maintain closer evolutionary relationships with amoebal hosts, which serve as natural reservoirs and training grounds for intracellular pathogens (9,100). This underscores that infection risk is shaped not solely by environmental abundance but also by transmission pathways and genomic traits that are selectively favored in human-associated environments such as engineered water systems. Accordingly, understanding pathogenic mechanisms of environmentally-enriched *Legionella* taxa may be important under future climatic conditions if natural reservoirs become a significant source of transmission.

Comparisons of 16S rRNA genes from complete and reference-quality *Legionella* genomes demonstrates that the commonly sequenced 16S rRNA gene region provides moderate taxonomic resolution within the genus, resolving many species into unique ASVs while collapsing others - including several clinically relevant taxa - into shared sequence

variants. While in this study we only utilized the first 150 bp of the 515F-806R region, additional analysis indicates that these trends would persist even when using the entire region (Fig. S14). This partial resolution underscores a key limitation of amplicon-based approaches for inferring pathogenic potential: the presence of an ASV matching a known pathogen does not necessarily uniquely identify a single pathogenic species. Additionally, the absence of such matches also does not preclude pathogenic capacity, given that pathogenicity in *Legionella* is narrowly defined by culture-confirmed human infections (101). Further studies are needed to evaluate *Legionella*-specific primer sets for taxonomic resolution, coverage and specificity. Additionally, since our analysis is based on amplicon sequencing of environmental DNA, detection does not necessarily indicate metabolically active populations. Future work incorporating RNA-based approaches, environmental aerosol measurements, or targeted cultivation may help distinguish active or potentially transmissible populations from dead or dormant cells.

Concluding remarks

In summary, our global analysis parses large-scale microbiome datasets to show that *Legionella* are a widespread but low-abundance member of soil microbiomes whose distribution is strongly structured by climate and biogeography. The enrichment of *Legionella* and pathogenic *Legionella* ASVs under warmer and wetter conditions suggests that climatic factors play a critical role in shaping environmental reservoirs of these bacteria. Future studies using *Legionella*-specific primer sets and quantitative approaches such as qPCR or ddPCR will help resolve absolute abundance patterns and clarify the role of soils as potential reservoirs of exposure.

Methods

Global soil 16S rRNA gene amplicon datasets

Publicly available 16S rRNA gene amplicon datasets were retrieved from the NCBI Sequence Read Archive (SRA) (102). Studies were identified through targeted searches for soil microbiome datasets and either (i) spanned broad continental-scale geographic ranges or (ii) consisted of individual field sites that were manipulated to test specific environmental variables. To enable global-scale comparisons, only studies including at least the first 150 bp downstream of the bacterial 16S rRNA gene 515F primer were included (58-73). Studies were further screened to ensure the availability of geographic coordinates, sampling date, and key edaphic metadata, including soil pH, total organic carbon, and total organic nitrogen. Climatic variables, including midrange monthly temperature at the time of sampling and annual precipitation, were extracted from the WorldClim database using sample coordinates (78).

Sequence processing

Processed 16S rRNA gene amplicon sequences in FASTA format were downloaded from the NCBI Sequence Read Archive (SRA) and subjected to a standardized downstream processing pipeline. Sequences were trimmed to a uniform length of 150 bp using seqkit (103), and reads shorter than 150 bp were removed. Quality-controlled sequences were denoised using Deblur (104), resulting in error-corrected amplicon sequence variants. Resulting sequences were dereplicated using VSEARCH (105) to obtain sequence abundances across samples. To reduce potential artifacts and sequencing noise, sequences were clustered at 99% identity and global singleton clusters were removed to reduce potential artifacts. After arriving at this final dataset, all downstream analyses were performed using amplicon sequence variants (ASVs) rather than clustered OTUs to get fine-scale resolution.

Taxonomic identification of *Legionella*

To maximize sensitivity and specificity of *Legionella* detection, we implemented a two-step taxonomic classification approach. First, all ASVs were screened using BLAST (106) against a custom *Legionella* reference database. This database comprised 16S sequences from all reference-quality and complete *Legionella* genomes available in the NCBI database, together with all sequences from neighboring family Coxiellaceae (also within the order *Legionellales*), as well as a representative selection of 16S sequences spanning the broader Gammaproteobacteria and bacterial domain. Candidate sequences were then placed into a backbone phylogenetic tree constructed from this custom reference database using PhyML (107) to confirm phylogenetic affiliation. Only ASVs that clustered within the *Legionella* clade were retained. The accuracy and specificity of this approach was verified using all Gammaproteobacteria reference sequences from the SILVA database, as well as a mock community composed of DNA from *L. pneumophila*, *L. longbeachae* and *L. micdadei*.

Association of *Legionella* relative abundance with environmental metadata

Legionella relative abundance was calculated as the proportion of *Legionella* reads relative to the total number of bacterial reads per sample. Samples lacking detectable *Legionella* sequences were retained for presence-absence and modeling analyses. Associations between *Legionella* relative abundance and environmental variables were assessed using Spearman rank correlations. For field experiments involving controlled environmental manipulations, *Legionella* relative abundance was compared between treatment

and control plots using non-parametric tests or correlation analyses, as appropriate for each experimental design. To examine non-monotonic and threshold responses, each environmental variable was partitioned into five bins spanning the full observed range, and differences in *Legionella* relative abundance among bins were tested using Kruskal-Wallis tests followed by Dunn's post hoc comparisons with multiple-testing correction. To evaluate interactive climatic effects, *Legionella* relative abundance was modeled across the joint temperature-precipitation space using a two-dimensional generalized additive model (GAM) (108). Model performance was compared to additive models with independent smooth terms for temperature and precipitation using analysis of variance (ANOVA) to assess improvement in fit.

Mixed-effects modeling of *Legionella* relative abundance

To evaluate the joint effects of climatic and biogeochemical variables while accounting for study-level variation, a hurdle modeling framework was applied (109). Presence-absence of *Legionella* was modeled using a binomial mixed-effects model, and relative abundance conditional on presence was modeled using a beta mixed-effects regression. Study identity was included as a random intercept in both models to account for non-independence among samples (110). Predictor variables included temperature, precipitation, soil pH, elevation, total organic carbon, and total organic nitrogen.

***Legionella* community composition and diversity**

Legionella community composition was assessed using Bray-Curtis dissimilarities (111) calculated from *Legionella*-only ASV profiles. Ordination was performed using non-metric multidimensional scaling (NMDS) (112). Differences in community composition among regions were tested using permutation-based multivariate analysis of variance (PERMANOVA) (113). Environmental variables were fitted onto ordinations using permutation-based vector fitting (envfit) (114). Spatial structuring of *Legionella* communities was assessed by examining the relationship between Bray-Curtis similarity and geographic distance, with geographic distance calculated from sample coordinates. Distance-decay relationships were modeled using power-law regression (115). Multiple regression on distance matrices (MRM) was used to test whether geographic distance remained a significant predictor of community dissimilarity after accounting for climatic and biogeochemical differences among sites (116).

To examine dominant *Legionella* ASVs and their regional structuring, the 300 most abundant *Legionella* ASVs within each geographic region were selected. Overlap among regions were quantified using pairwise comparisons, and relationship between ASV overlap and climatic dissimilarity among regions were assessed using Spearman correlations. To evaluate whether biogeographic and climatic regions displayed deep phylogenetic divergence, a maximum-likelihood phylogeny was constructed using the 300 most abundant *Legionella* ASVs from each region together with reference sequences from complete and reference-quality *Legionella* genomes. Sequences were aligned using MAFFT (117), and phylogenetic inference was performed using IQ-Tree (118) under an appropriate nucleotide substitution model. Phylogenetic trees were visualized and annotated using iTOL (119). To quantify within-region phylogenetic structure, clustering and dispersion were assessed using the Net Relatedness Index (NRI) and Nearest Taxon Index (NTI) (98), calculated using the picante R package (120) as standardized effect sizes of mean pairwise phylogenetic distance (MPD) and mean nearest taxon distance (MNTD) relative to a null model with 999 randomizations. Significance was determined using rank-based two-tailed p-values ($\alpha=0.01$), where $p < 0.01$ indicated significant clustering and $p > 0.99$ indicated significant overdispersion.

Community context of *Legionella*-rich samples

To identify microbial taxa associated with *Legionella*-rich communities, differential abundance analysis was performed using 97% OTUs, which were clustered using VSEARCH (105). Samples were stratified into upper and lower quartiles of *Legionella* relative abundance. Differentially abundant OTUs were identified using $\log_2$fold-change thresholds and false discovery rate (FDR) correction (121). Taxonomic composition and prevalence of enriched OTUs were plotted to assess taxa that were abundant in both *Legionella*-rich and *Legionella*-poor communities.

Identification of pathogenic *Legionella* ASVs

Reference-quality and complete *Legionella* renomes were retrieved from NCBI (122), trimmed to the first 150 bp following the 515F primer and clustered to determine species-level resolution. ASVs matching reference sequences from known pathogenic species were designated as pathogenic ASVs, while ASVs shared with both pathogenic and non-pathogenic species were designated as potentially pathogenic. Associations between pathogenic or potentially pathogenic *Legionella* ASVs and environmental variables were assessed using logistic regression models. Presence-absence of pathogenic ASVs was modeled as a function of climatic and biogeochemical variables, with sequencing depth included as a covariate to account for detection bias.

Acknowledgements
Research reported in this publication was supported by the National Institute Of General Medical Sciences of the National Institutes of Health under Award Number P20GM125508. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We thank the investigators and research teams who generated and made publicly available the sequencing datasets used in this study. We also thank the Center for Microbial Analysis, Innovation, and Knowledge Integration (CMAIKI) at the University of Hawai'i for computational resources and support.

Data availability
All sequencing data used in this study are publicly available through the NCBI BioProject repositories listed in Supplementary Table 1. All analyses were conducted using open-access software.

References

1. Mondino S, Schmidt S, Rolando M, Escoll P, Gomez-Valero L, Buchrieser C. 2020. Legionnaires' Disease: State of the Art Knowledge of Pathogenesis Mechanisms of Legionella. Annu Rev Pathol Mech Dis 15:439–466.
2. Centers for Disease Control and Prevention. 2023. Legionnaires' disease surveillance summary report, United States, 2020–2021. U.S. Department of Health and Human Services, Atlanta, GA.
3. European Centre for Disease Prevention and Control. 2024. Legionnaires' disease annual epidemiological report for 2022. ECDC, Stockholm, Sweden.
4. Baker-Goering M, Roy K, Edens C, Collier S. 2021. Economic Burden of Legionnaires' Disease, United States, 2014. Emerg Infect Dis 27:255–257.
5. Al-Quadan T, Price CT, Abu Kwaik Y. 2012. Exploitation of evolutionarily conserved amoeba and mammalian processes by Legionella. Trends in Microbiology 20:299–306.
6. Durie CL, Sheedlo MJ, Chung JM, Byrne BG, Su M, Knight T, Swanson M, Lacy DB, Ohi MD. 2020. Structural analysis of the Legionella pneumophila Dot/Icm type IV secretion system core complex. eLife 9:e59530.
7. Fuche F, Vianney A, Andrea C, Doublet P, Gilbert C. 2015. Functional Type 1 Secretion System Involved in Legionella pneumophila Virulence. J Bacteriol 197:563–571.
8. Cianciotto NP. 2013. Type II Secretion and Legionella Virulence, p. 81–102. In Hilbi, H (ed.), Molecular Mechanisms in Legionella Pathogenesis. Springer Berlin Heidelberg, Berlin, Heidelberg.
9. Molmeret M, Horn M, Wagner M, Santic M, Abu Kwaik Y. 2005. Amoebae as Training Grounds for Intracellular Bacterial Pathogens. Appl Environ Microbiol 71:20–28.
10. Graells T, Ishak H, Larsson M, Guy L. 2018. The all-intracellular order Legionellales is unexpectedly diverse, globally distributed and lowly abundant. FEMS Microbiology Ecology 94.
11. Fraser DW, Tsai TR, Orenstein W, Parkin WE, Beecham HJ, Sharrar RG, Harris J, Mallison GF, Martin SM, McDade JE, Shepard CC, Brachman PS, the Field Investigation Team*. 1977. Legionnaires' Disease: Description of an Epidemic of Pneumonia. N Engl J Med 297:1189–1197.
12. 2007. Legionella and the prevention of legionellosis. World Health Organization, Geneva.
13. Van Heijnsbergen E, Schalk JAC, Euser SM, Brandsema PS, Den Boer JW, De Roda Husman AM. 2015. Confirmed and Potential Sources of Legionella Reviewed. Environ Sci Technol 49:4797–4815.
14. Pierre DM, Baron J, Yu VL, Stout JE. 2017. Diagnostic testing for Legionnaires' disease. Ann Clin Microbiol Antimicrob 16:59.
15. Mercante JW, Winchell JM. 2015. Current and Emerging Legionella Diagnostics for Laboratory and Outbreak Investigations. Clin Microbiol Rev 28:95–133.
16. Bornstein N, Mercatello A, Marmet D, Surgot M, Deveaux Y, Fleurette J. 1989. Pleural infection caused by Legionella anisa. J Clin Microbiol 27:2100–2101.
17. Wilkinson HW, Thacker WL, Benson RF, Polt SS, Brookings E, Mayberry WR, Brenner DJ, Gilley RG, Kirklin JK. 1987. Legionella birminghamensis sp. nov. isolated from a cardiac transplant recipient. J Clin Microbiol 25:2120–2122.
18. Mitchell RG, Pasvol G, Newnham RS. 1984. Pneumonia due to Legionella bozemanii: First report of a case in Europe. Journal of Infection 8:251–255.
19. Pearce MM, Theodoropoulos N, Mandel MJ, Brown E, Reed KD, Cianciotto NP. 2012. Legionella cardiaca sp. nov., isolated from a case of native valve endocarditis in a

human heart. International Journal of Systematic and Evolutionary Microbiology 62:2946–2954.
20. Fang G-D, Yu VL, Vickers RM. 1989. Disease Due to the Legionellaceae (Other than Legionella pneumophila): Historical, Microbiological, Clinical, and Epidemiological Review. Medicine 68:116.
21. Thacker WL, Benson RF, Staneck JL, Vincent SR, Mayberry WR, Brenner DJ, Wilkinson HW. 1988. Legionella cincinnatiensis sp. nov. isolated from a patient with pneumonia. J Clin Microbiol 26:418–420.
22. Palmer A, Painter J, Hassler H, Richards VP, Bruce T, Morrison S, Brown E, Kozak-Muiznieks NA, Lucas C, McNealy TL. 2016. Legionella clemsonensis sp. nov.: a green fluorescing Legionella strain from a patient with pneumonia. Microbiology and Immunology 60:694–701.
23. Joly JR, Déry P, Gauvreau L, Coté L, Trépanier C. 1986. Legionnaires' disease caused by Legionella dumoffii in distilled water. CMAJ. 135(11):1274-1277.
24. Palutke WA, Crane LR, Wentworth BB, Geiger JG, Cardozo L, Singhakowinta A, Bartley J, Robinson BE. 1986. Legionella feeleii– Associated Pneumonia in Humans. American Journal of Clinical Pathology 86:348–351.
25. Griffith ME, Lindquist DS, Benson RF, Thacker WL, Brenner DJ, Wilkinson HW. 1988. First isolation of Legionella gormanii from human disease. J Clin Microbiol 26:380–381.
26. Wilkinson HW, Thacker WL, Steigerwalt AG, Brenner DJ, Ampel NM, Wing EJ. 1985. Second serogroup of Legionella hackeliae isolated from a patient with pneumonia. J Clin Microbiol 22:488–489.
27. Littrup P, Madsen JK, Lind K. 1987. Aortic valve endocarditis associated with Legionella infection after Mycoplasma pneumonia. Heart 58:293–295.
28. Thacker WL, Dyke JW, Benson RF, Havlichek DH, Robinson-Dunn B, Stiefel H, Schneider W, Moss CW, Mayberry WR, Brenner DJ. 1992. Legionella lansingensis sp. nov. isolated from a patient with pneumonia and underlying chronic lymphocytic leukemia. J Clin Microbiol 30:2398–2401.
29. McKinney RM, Porschen RK, Edelstein PH, Bissett ML, Harris PP, Bondell SP, Steigerwalt AG, Weaver RE, Ein ME, Lindquist DS, Kops RS, Brenner DJ. 1981. Legionella longbeachae Species Nova, Another Etiologic Agent of Human Pneumonia. Ann Intern Med 94:739–743.
30. Wilkinson HW, Thacker WL, Brenner DJ, Ryan KJ. 1985. Fatal Legionella maceachernii pneumonia. J Clin Microbiol 22:1055.
31. Myerowitz RL, Pasculle AW, Dowling JN, Pazin GJ, Puerzer M, Yee RB, Rinaldo CR, Hakala TR. 1979. Opportunistic Lung Infection Due to “Pittsburgh Pneumonia Agent.” N Engl J Med 301:953–958.
32. Tang PW, Toma S, MacMillan LG. 1985. Legionella oakridgensis: laboratory diagnosis of a human infection. J Clin Microbiol 21:462–463.
33. Lo Presti F, Riffard S, Vandenesch F, Reyrolle M, Ronco E, Ichai P, Etienne J. 1997. The first clinical isolate of Legionella parisiensis, from a liver transplant patient with pneumonia. J Clin Microbiol 35:1706–1709.
34. Brenner DJ, Steigerwalt AG, McDADE JE. 1979. Classification of the Legionnaires’ Disease Bacterium: Legionella pneumophila , genus novum, species nova, of the Family Legionellaceae, familia nova. Ann Intern Med 90:656–658.
35. Matsui M, Fujii S, Shiroiwa R, Amemura-Maekawa J, Chang B, Kura F, Yamauchi K. 2010. Isolation of Legionella rubrilucens from a pneumonia patient co-infected with Legionella pneumophila. Journal of Medical Microbiology 59:1242–1246.
36. Yuan X, Meng F, Yu X, Bai C, Jia R, Meng F, Wang F, Fu J. 2022. Legionella sainthelensi Pneumonia with Abdominal Symptoms as Prominent Manifestations: A Case Report. Infect Dis Immun 2:285–288.

37. Thacker WL, Benson RF, Schifman RB, Pugh E, Steigerwalt AG, Mayberry WR, Brenner DJ, Wilkinson HW. 1989. Legionella tucsonensis sp. nov. isolated from a renal transplant recipient. J Clin Microbiol 27:1831–1834.
38. Edelstein PH, Brenner DJ, Moss CW, Steigerwalt AG, Francis EM, George WL. 1982. Legionella wadsworthii Species Nova: A Cause of Human Pneumonia. Ann Intern Med 97:809–813.
39. Moffa MA, Rock C, Galiatsatos P, Gamage SD, Schwab KJ, Exum NG. 2023. Legionellosis on the rise: A scoping review of sporadic, community-acquired incidence in the United States. Epidemiol Infect 151:e133.
40. Han XY. 2021. Effects of climate changes and road exposure on the rapidly rising legionellosis incidence rates in the United States. PLoS ONE 16:e0250364.
41. Bacigalupe R, Lindsay D, Edwards G, Fitzgerald JR. 2017. Population Genomics of Legionella longbeachae and Hidden Complexities of Infection Source Attribution. Emerg Infect Dis 23:750–757.
42. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Geer LY, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Ostell J, Miller V, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L, Yaschenko E. 2007. Database resources of the National Center for Biotechnology Information. Nucleic Acids Research 35:D5–D12.
43. Cazalet C, Gomez-Valero L, Rusniok C, Lomma M, Dervins-Ravault D, Newton HJ, Sansom FM, Jarraud S, Zidane N, Ma L, Bouchier C, Etienne J, Hartland EL, Buchrieser C. 2010. Analysis of the Legionella longbeachae Genome and Transcriptome Uncovers Unique Strategies to Cause Legionnaires' Disease. PLoS Genet 6:e1000851.
44. Gomez-Valero L, Rusniok C, Cazalet C, Buchrieser C. 2011. Comparative and Functional Genomics of Legionella Identified Eukaryotic Like Proteins as Key Players in Host-Pathogen Interactions. Front Microbio 2:208
45. Burstein D, Amaro F, Zusman T, Lifshitz Z, Cohen O, Gilbert JA, Pupko T, Shuman HA, Segal G. 2016. Genomic analysis of 38 Legionella species identifies large and diverse effector repertoires. Nat Genet 48:167–175.
46. Gomez-Valero L, Rusniok C, Carson D, Mondino S, Pérez-Cobas AE, Rolando M, Pasricha S, Reuter S, Demirtas J, Crumbach J, Descorps-Declere S, Hartland EL, Jarraud S, Dougan G, Schroeder GN, Frankel G, Buchrieser C. 2019. More than 18,000 effectors in the Legionella genus genome provide multiple, independent combinations for replication in human cells. Proc Natl Acad Sci USA 116:2265–2273.
47. Whiley H, Bentham R. 2011. Legionella longbeachae and Legionellosis. Emerg Infect Dis 17:579–583.
48. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, Owens SM, Betley J, Fraser L, Bauer M, Gormley N, Gilbert JA, Smith G, Knight R. 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. The ISME Journal 6:1621–1624.
49. Hugenholtz P, Tyson GW. 2008. Metagenomics. Nature 455:481–483.
50. Shokralla S, Spall JL, Gibson JF, Hajibabaei M. 2012. Next-generation sequencing technologies for environmental DNA research. Molecular Ecology 21:1794–1805.
51. Chen I-MA, Chu K, Palaniappan K, Ratner A, Huang J, Huntemann M, Hajek P, Ritter SJ, Webb C, Wu D, Varghese NJ, Reddy TBK, Mukherjee S, Ovchinnikova G, Nolan M, Seshadri R, Roux S, Visel A, Woyke T, Eloe-Fadrosh EA, Kyrpides NC, Ivanova NN. 2023. The IMG/M data management and analysis system v.7: content updates and new features. Nucleic Acids Research 51:D723–D732.

52. Collins S, Jorgensen F, Willis C, Walker J. 2015. Real-time PCR to supplement gold-standard culture-based detection of Legionella in environmental samples. J Appl Microbiol 119:1158–1169.
53. Janda JM, Abbott SL. 2007. 16S rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls. J Clin Microbiol 45:2761–2764.
54. Edgar RC. 2018. Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences. PeerJ 6:e4652.
55. Bokulich NA, Kaehler BD, Rideout JR, Dillon M, Bolyen E, Knight R, Huttley GA, Gregory Caporaso J. 2018. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. Microbiome 6:90.
56. Balvočiūtė M, Huson DH. 2017. SILVA, RDP, Greengenes, NCBI and OTT — how do these taxonomies compare? BMC Genomics 18:114.
57. Edgar RC. 2013. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat Methods 10:996–998.
58. Labouyrie M, Ballabio C, Romero F, Panagos P, Jones A, Schmid MW, Mikryukov V, Dulya O, Tedersoo L, Bahram M, Lugato E, Van Der Heijden MGA, Orgiazzi A. 2023. Patterns in soil microbial diversity across Europe. Nat Commun 14:3311.
59. Cowan D, Lebre P, Amon C, Becker R, Boga H, Boulangé A, Chiyaka T, Coetzee T, De Jager P, Dikinya O, Eckardt F, Greve M, Harris M, Hopkins D, Houngnandan H, Houngnandan P, Jordaan K, Kaimoyo E, Kambura A, Kamgan-Nkuekam G, Makhalanyane T, Maggs-Kölling G, Marais E, Mondlane H, Nghalipo E, Olivier B, Ortiz M, Pertierra L, Ramond J-B, Seely M, Sithole-Niang I, Valverde A, Varliero G, Vikram S, Wall D, Zeze A. 2022. Biogeographical survey of soil microbiomes across sub-Saharan Africa: structure, drivers, and predicted climate-driven changes. Microbiome 10:131.
60. Zhou J, Deng Y, Shen L, Wen C, Yan Q, Ning D, Qin Y, Xue K, Wu L, He Z, Voordeckers JW, Nostrand JDV, Buzzard V, Michaletz ST, Enquist BJ, Weiser MD, Kaspari M, Waide R, Yang Y, Brown JH. 2016. Temperature mediates continental-scale diversity of microbes in forest soils. Nat Commun 7:12083.
61. Rodriguez V, Moskwa L-M, Oses R, Kühn P, Riveras-Muñoz N, Seguel O, Scholten T, Wagner D. 2022. Impact of Climate and Slope Aspects on the Composition of Soil Bacterial Communities Involved in Pedogenetic Processes along the Chilean Coastal Cordillera. Microorganisms 10:847.
62. Amend AS, Swift SOI, Darcy JL, Belcaid M, Nelson CE, Buchanan J, Cetraro N, Fraiola KMS, Frank K, Kajihara K, McDermot TG, McFall-Ngai M, Medeiros M, Mora C, Nakayama KK, Nguyen NH, Rollins RL, Sadowski P, Sparagon W, Téfit MA, Yew JY, Yogi D, Hynson NA. 2022. A ridge-to-reef ecosystem microbial census reveals environmental reservoirs for animal and plant microbiomes. Proc Natl Acad Sci USA 119:e2204146119.
63. Dai Z, Guo X, Lin J, Wang X, He D, Zeng R, Meng J, Luo J, Delgado-Baquerizo M, Moreno-Jiménez E, Brookes PC, Xu J. 2023. Metallic micronutrients are associated with the structure and function of the soil microbiome. Nat Commun 14:8456.
64. Ramirez KS, Leff JW, Barberán A, Bates ST, Betley J, Crowther TW, Kelly EF, Oldfield EE, Shaw EA, Steenbock C, Bradford MA, Wall DH, Fierer N. 2014. Biogeographic patterns in below-ground diversity in New York City's Central Park are similar to those observed globally. Proc R Soc B 281:20141988.
65. O'Brien SL, Gibbons SM, Owens SM, Hampton-Marcell J, Johnston ER, Jastrow JD, Gilbert JA, Meyer F, Antonopoulos DA. 2016. Spatial scale drives patterns in soil bacterial diversity. Environmental Microbiology 18:2039–2051.
66. Potter C, Freeman C, Golyshin PN, Ackermann G, Fenner N, McDonald JE, Ehbair A, Jones TG, Murphy LM, Creer S. 2017. Subtle shifts in microbial communities occur

alongside the release of carbon induced by drought and rewetting in contrasting peatland ecosystems. Sci Rep 7:11314.
67. Burns KN, Bokulich NA, Cantu D, Greenhut RF, Kluepfel DA, O'Geen AT, Strauss SL, Steenwerth KL. 2016. Vineyard soil bacterial diversity and composition revealed by 16S rRNA genes: Differentiation by vineyard management. Soil Biology and Biochemistry 103:337–348.
68. Wood SA, Bradford MA, Gilbert JA, McGuire KL, Palm CA, Tully KL, Zhou J, Naeem S. 2015. Agricultural intensification and the functional capacity of soil microbes on smallholder African farms. Journal of Applied Ecology 52:744–752.
69. Kaufmann C, Cassin-Sackett L. 2021. Fine-Scale Spatial Structure of Soil Microbial Communities in Burrows of a Keystone Rodent Following Mass Mortality. Front Ecol Evol 9:758348.
70. Nguyen TBA, Chen Q, Yan Z, Li C, He J, Hu H. 2023. Trophic interrelationships of bacteria are important for shaping soil protist communities. Environ Microbiol Rep 15:298–307.
71. Metze D, Schnecker J, De Carlan CLN, Bhattarai B, Verbruggen E, Ostonen I, Janssens IA, Sigurdsson BD, Hausmann B, Kaiser C, Richter A. 2024. Soil warming increases the number of growing bacterial taxa but not their growth rates. Sci Adv 10:eadk6295.
72. Wu L, Zhang Y, Guo X, Ning D, Zhou X, Feng J, Yuan MM, Liu S, Guo J, Gao Z, Ma J, Kuang J, Jian S, Han S, Yang Z, Ouyang Y, Fu Y, Xiao N, Liu X, Wu L, Zhou A, Yang Y, Tiedje JM, Zhou J. 2022. Reduction of microbial diversity in grassland soil is driven by long-term climate warming. Nat Microbiol 7:1054–1062.
73. Lemoine NP, Budny ML, Rose E, Lucas J, Marshall CW. 2024. Seasonal soil moisture thresholds inhibit bacterial activity and decomposition during drought in a tallgrass prairie. Oikos 2024:e10201.
74. Lynch MDJ, Neufeld JD. 2015. Ecology and exploration of the rare biosphere. Nat Rev Microbiol 13:217–229.
75. Brady OJ, Gething PW, Bhatt S, Messina JP, Brownstein JS, Hoen AG, Moyes CL, Farlow AW, Scott TW, Hay SI. 2012. Refining the Global Spatial Limits of Dengue Virus Transmission by Evidence-Based Consensus. PLoS Negl Trop Dis 6:e1760.
76. Vinetz JM. 2001. Leptospirosis: Current Opinion in Infectious Diseases 14:527–538.
77. Fierer N. 2017. Embracing the unknown: disentangling the complexities of the soil microbiome. Nat Rev Microbiol 15:579–590.
78. Fick SE, Hijmans RJ. 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. Intl Journal of Climatology 37:4302–4315.
79. Phin N, Parry-Ford F, Harrison T, Stagg HR, Zhang N, Kumar K, Lortholary O, Zumla A, Abubakar I. 2014. Epidemiology and clinical management of Legionnaires' disease. The Lancet Infectious Diseases 14:1011–1021.
80. Declerck P. 2010. Biofilms: the environmental playground of Legionella pneumophila. Environmental Microbiology 12:557–566.
81. Davidson EA, Janssens IA. 2006. Temperature sensitivity of soil carbon decomposition and feedbacks to climate change. Nature 440:165–173.
82. Harvell CD, Mitchell CE, Ward JR, Altizer S, Dobson AP, Ostfeld RS, Samuel MD. 2002. Climate Warming and Disease Risks for Terrestrial and Marine Biota. Science 296:2158–2162.
83. Schimel J, Balser TC, Wallenstein M. 2007. Microbial stress-response physiology and its implications for ecosystem function. Ecology 88:1386–1394.
84. Allison SD, Martiny JBH. 2008. Resistance, resilience, and redundancy in microbial communities. Proc Natl Acad Sci USA 105:11512–11519.
85. Kielak AM, Barreto CC, Kowalchuk GA, Van Veen JA, Kuramae EE. 2016. The Ecology of Acidobacteria: Moving beyond Genes and Genomes. Front Microbiol 7.

86. Fierer N, Bradford MA, Jackson RB. 2007. Toward an ecological classification of soil bacteria. Ecology 88:1354–1364.
87. Barka EA, Vatsa P, Sanchez L, Gaveau-Vaillant N, Jacquard C, Klenk H-P, Clément C, Ouhdouch Y, Van Wezel GP. 2016. Taxonomy, Physiology, and Natural Products of Actinobacteria. Microbiol Mol Biol Rev 80:1–43.
88. Delgado-Baquerizo M, Oliverio AM, Brewer TE, Benavent-González A, Eldridge DJ, Bardgett RD, Maestre FT, Singh BK, Fierer N. 2018. A global atlas of the dominant bacteria found in soil. Science 359:320–325.
89. Guerrieri E, Bondi M, Sabia C, De Niederhäusern S, Borella P, Messi P. 2008. Effect of Bacterial Interference on Biofilm Development by Legionella pneumophila. Curr Microbiol 57:532–536.
90. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, Tripathi A, Gibbons SM, Ackermann G, Navas-Molina JA, Janssen S, Kopylova E, Vázquez-Baeza Y, González A, Morton JT, Mirarab S, Zech Xu Z, Jiang L, Haroon MF, Kanbar J, Zhu Q, Jin Song S, Kosciolek T, Bokulich NA, Lefler J, Brislawn CJ, Humphrey G, Owens SM, Hampton-Marcell J, Berg-Lyons D, McKenzie V, Fierer N, Fuhrman JA, Clauset A, Stevens RL, Shade A, Pollard KS, Goodwin KD, Jansson JK, Gilbert JA, Knight R, The Earth Microbiome Project Consortium, Rivera JLA, Al-Moosawi L, Alverdy J, Amato KR, Andras J, Angenent LT, Antonopoulos DA, Apprill A, Armitage D, Ballantine K, Bárta J, Baum JK, Berry A, Bhatnagar A, Bhatnagar M, Biddle JF, Bittner L, Boldgiv B, Bottos E, Boyer DM, Braun J, Brazelton W, Brearley FQ, Campbell AH, Caporaso JG, Cardona C, Carroll J, Cary SC, Casper BB, Charles TC, Chu H, Claar DC, Clark RG, Clayton JB, Clemente JC, Cochran A, Coleman ML, Collins G, Colwell RR, Contreras M, Crary BB, Creer S, Cristol DA, Crump BC, Cui D, Daly SE, Davalos L, Dawson RD, Defazio J, Delsuc F, Dionisi HM, Dominguez-Bello MG, Dowell R, Dubinsky EA, Dunn PO, Ercolini D, Espinoza RE, Ezenwa V, Fenner N, Findlay HS, Fleming ID, Fogliano V, Forsman A, Freeman C, Friedman ES, Galindo G, Garcia L, Garcia-Amado MA, Garshelis D, Gasser RB, Gerdts G, Gibson MK, Gifford I, Gill RT, Giray T, Gittel A, Golyshin P, Gong D, Grossart H-P, Guyton K, Haig S-J, Hale V, Hall RS, Hallam SJ, Handley KM, Hasan NA, Haydon SR, Hickman JE, Hidalgo G, Hofmockel KS, Hooker J, Hulth S, Hultman J, Hyde E, Ibáñez-Álamo JD, Jastrow JD, Jex AR, Johnson LS, Johnston ER, Joseph S, Jurburg SD, Jurelevicius D, Karlsson A, Karlsson R, Kauppinen S, Kellogg CTE, Kennedy SJ, Kerkhof LJ, King GM, Kling GW, Koehler AV, Krezalek M, Kueneman J, Lamendella R, Landon EM, Lane-deGraaf K, LaRoche J, Larsen P, Laverock B, Lax S, Lentino M, Levin II, Liancourt P, Liang W, Linz AM, Lipson DA, Liu Y, Lladser ME, Lozada M, Spirito CM, MacCormack WP, MacRae-Crerar A, Magris M, Martín-Platero AM, Martín-Vivaldi M, Martínez LM, Martínez-Bueno M, Marzinelli EM, Mason OU, Mayer GD, McDevitt-Irwin JM, McDonald JE, McGuire KL, McMahon KD, McMinds R, Medina M, Mendelson JR, Metcalf JL, Meyer F, Michelangeli F, Miller K, Mills DA, Minich J, Mocali S, Moitinho-Silva L, Moore A, Morgan-Kiss RM, Munroe P, Myrold D, Neufeld JD, Ni Y, Nicol GW, Nielsen S, Nissimov JI, Niu K, Nolan MJ, Noyce K, O'Brien SL, Okamoto N, Orlando L, Castellano YO, Osuolale O, Oswald W, Parnell J, Peralta-Sánchez JM, Petraitis P, Pfister C, Pilon-Smits E, Piombino P, Pointing SB, Pollock FJ, Potter C, Prithiviraj B, Quince C, Rani A, Ranjan R, Rao S, Rees AP, Richardson M, Riebesell U, Robinson C, Rockne KJ, Rodriguezl SM, Rohwer F, Roundstone W, Safran RJ, Sangwan N, Sanz V, Schrenk M, Schrenzel MD, Scott NM, Seger RL, Seguin-Orlando A, Seldin L, Seyler LM, Shakhsheer B, Sheets GM, Shen C, Shi Y, Shin H, Shogan BD, Shutler D, Siegel J, Simmons S, Sjöling S, Smith DP, Soler JJ, Sperling M, Steinberg PD, Stephens B, Stevens MA, Taghavi S, Tai V, Tait K, Tan CL, Tas¸ N, Taylor DL, Thomas T, Timling I, Turner BL, Urich T, Ursell LK, Van Der Lelie D, Van Treuren W, Van Zwieten L, Vargas-Robles D, Thurber RV, Vitaglione P, Walker DA,

Walters WA, Wang S, Wang T, Weaver T, Webster NS, Wehrle B, Weisenhorn P, Weiss S, Werner JJ, West K, Whitehead A, Whitehead SR, Whittingham LA, Willerslev E, Williams AE, Wood SA, Woodhams DC, Yang Y, Zaneveld J, Zarraonaindia I, Zhang Q, Zhao H. 2017. A communal catalogue reveals Earth's multiscale microbial diversity. Nature 551:457–463.
91. Bahram M, Hildebrand F, Forslund SK, Anderson JL, Soudzilovskaia NA, Bodegom PM, Bengtsson-Palme J, Anslan S, Coelho LP, Harend H, Huerta-Cepas J, Medema MH, Maltz MR, Mundra S, Olsson PA, Pent M, Põlme S, Sunagawa S, Ryberg M, Tedersoo L, Bork P. 2018. Structure and function of the global topsoil microbiome. Nature 560:233–237.
92. Fields BS, Benson RF, Besser RE. 2002. Legionella and Legionnaires' Disease: 25 Years of Investigation. Clin Microbiol Rev 15:506–526.
93. Feeley JC, Gibson RJ, Gorman GW, Langford NC, Rasheed JK, Mackel DC, Baine WB. 1979. Charcoal-yeast extract agar: primary isolation medium for Legionella pneumophila. J Clin Microbiol 10:437–441.
94. Hanson CA, Fuhrman JA, Horner-Devine MC, Martiny JBH. 2012. Beyond biogeographic patterns: processes shaping the microbial landscape. Nat Rev Microbiol 10:497–506.
95. Green JL, Bohannan BJM, Whitaker RJ. 2008. Microbial Biogeography: From Taxonomy to Traits. Science 320:1039–1043.
96. Papke RT, Ramsing NB, Bateson MM, Ward DM. 2003. Geographical isolation in hot spring cyanobacteria. Environmental Microbiology 5:650–659.
97. Fierer N, Jackson RB. 2006. The diversity and biogeography of soil bacterial communities. Proc Natl Acad Sci USA 103:626–631.
98. Webb CO, Ackerly DD, McPeek MA, Donoghue MJ. 2002. Phylogenies and Community Ecology. Annu Rev Ecol Syst 33:475–505.
99. Geisen S, Mitchell EAD, Adl S, Bonkowski M, Dunthorn M, Ekelund F, Fernández LD, Jousset A, Krashevska V, Singer D, Spiegel FW, Walochnik J, Lara E. 2018. Soil protists: a fertile frontier in soil biology research. FEMS Microbiology Reviews 42:293–323.
100. Gomez-Valero L, Buchrieser C. 2019. Intracellular parasitism, the driving force of evolution of Legionella pneumophila and the genus Legionella. Microbes and Infection 21:230–236.
101. Newton HJ, Ang DKY, Van Driel IR, Hartland EL. 2010. Molecular Pathogenesis of Infections Caused by Legionella pneumophila. Clin Microbiol Rev 23:274–298.
102. Leinonen R, Sugawara H, Shumway M, on behalf of the International Nucleotide Sequence Database Collaboration. 2011. The Sequence Read Archive. Nucleic Acids Research 39:D19–D21.
103. Shen W, Le S, Li Y, Hu F. 2016. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. PLoS ONE 11:e0163962.
104. Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Zech Xu Z, Kightley EP, Thompson LR, Hyde ER, Gonzalez A, Knight R. 2017. Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. mSystems 2:e00191-16.
105. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016. VSEARCH: a versatile open source tool for metagenomics. PeerJ 4:e2584.
106. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.
107. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Systematic Biology 59:307–321.
108. Wood SN. 2017. Generalized Additive Models: An Introduction with R, 2nd ed. Chapman and Hall/CRC. https://www.taylorfrancis.com/books/9781498728348.

109. Zuur AF, Ieno EN, Walker N, Saveliev AA, Smith GM. 2009. Mixed effects models and extensions in ecology with R. Springer New York, New York, NY. https://link.springer.com/10.1007/978-0-387-87458-6.
110. Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, White J-SS. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. Trends in Ecology & Evolution 24:127–135.
111. Bray JR, Curtis JT. 1957. An Ordination of the Upland Forest Communities of Southern Wisconsin. Ecological Monographs 27:325–349.
112. Kruskal JB. 1964. Nonmetric Multidimensional Scaling: A Numerical Method. Psychometrika 29:115–129.
113. Anderson MJ. 2001. A new method for non-parametric multivariate analysis of variance. Austral Ecology 26:32–46.
114. Oksanen FJ. 2017. Vegan: Community Ecology Package. R package Version 2.4-3. https://CRAN.R-project.org/package=vegan
115. Nekola JC, White PS. 1999. The distance decay of similarity in biogeography and ecology. Journal of Biogeography 26:867–878.
116. Lichstein JW. 2007. Multiple regression on distance matrices: a multivariate spatial analysis tool. Plant Ecol 188:117–131.
117. Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution 30:772–780.
118. Nguyen L-T, Schmidt HA, Von Haeseler A, Minh BQ. 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Molecular Biology and Evolution 32:268–274.
119. Letunic I, Bork P. 2021. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Research 49:W293–W296.
120. Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, Blomberg SP, Webb CO. 2010. Picante: R tools for integrating phylogenies and ecology. Bioinformatics 26:1463–1464.
121. Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B: Statistical Methodology 57:289–300.
122. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, Farrell CM, Feldgarden M, Fine AM, Funk K, Hatcher E, Kannan S, Kelly C, Kim S, Klimke W, Landrum MJ, Lathrop S, Lu Z, Madden TL, Malheiro A, Marchler-Bauer A, Murphy TD, Phan L, Pujar S, Rangwala SH, Schneider VA, Tse T, Wang J, Ye J, Trawick BW, Pruitt KD, Sherry ST. 2023. Database resources of the National Center for Biotechnology Information in 2023. Nucleic Acids Research 51:D29–D38.

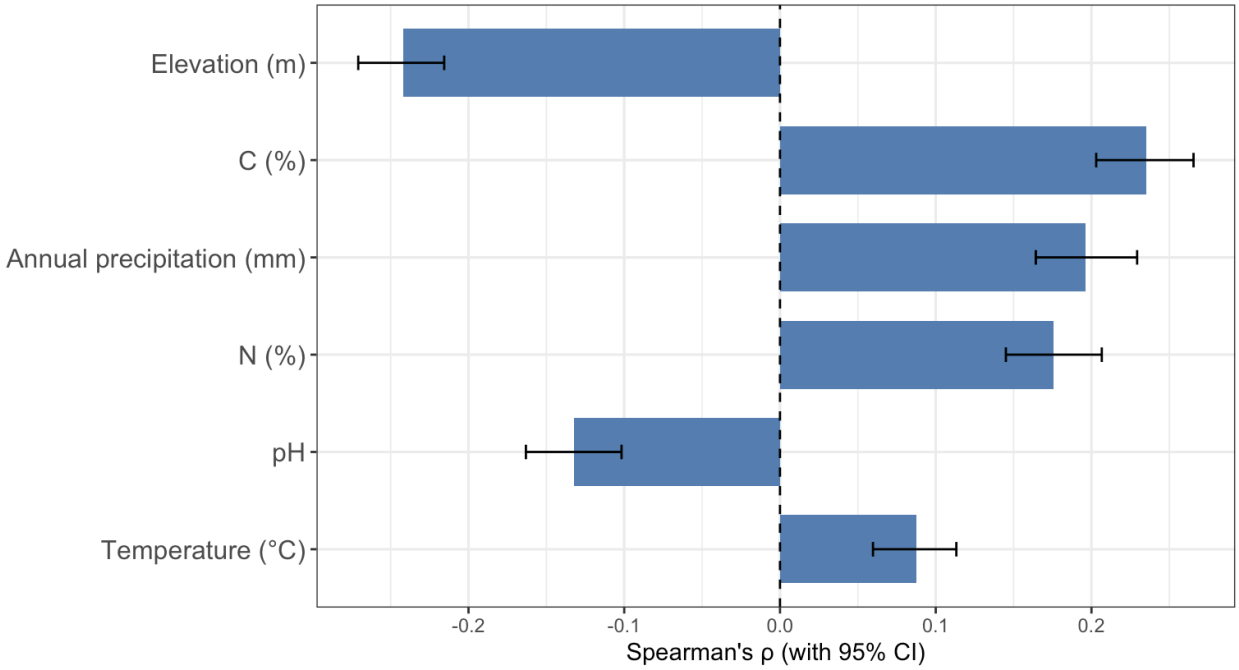

Fig. S1 - Spearman correlations between *Legionella* relative abundance and climatic and biogeochemical variables. Bars represent Spearman's rank correlation coefficients (ρ) between *Legionella* relative abundance and each environmental variable across all samples. Error bars indicate 95% confidence intervals.

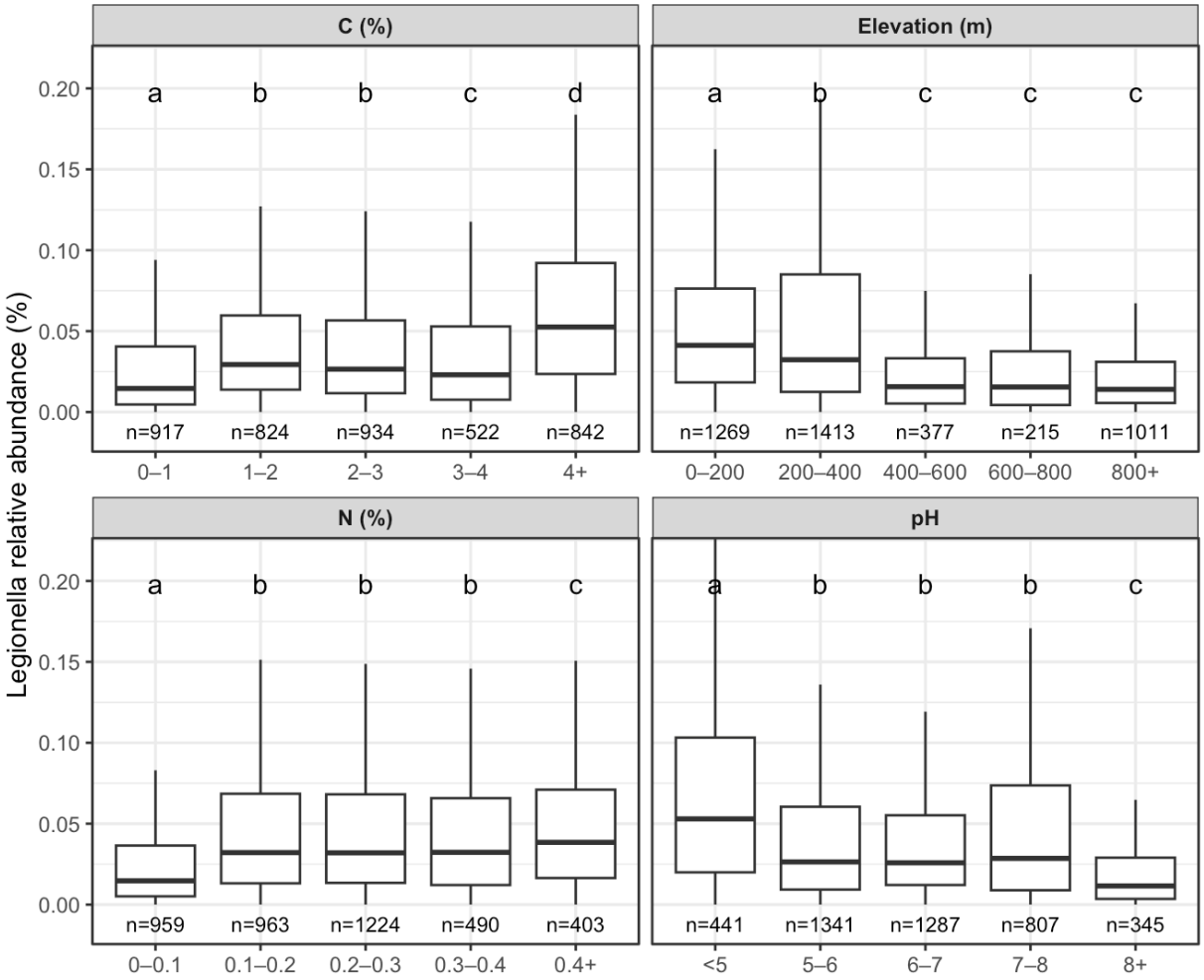

Fig. S2 - Variation in *Legionella* relative abundance across biogeochemical and abiotic gradients. Boxplots show *Legionella* relative abundance across five bins spanning the observed range of soil carbon (C), elevation, nitrogen (N) and pH values. Boxes represent the interquartile range (IQR), center lines indicate medians and whiskers extend to 1.5x the IQR. Sample sizes for each bin are indicated below the x-axis. Letters above boxes denote statistically distinct groups identified using Dunn's post hoc comparisons following Kruskal-Wallis tests ($p<0.05$).

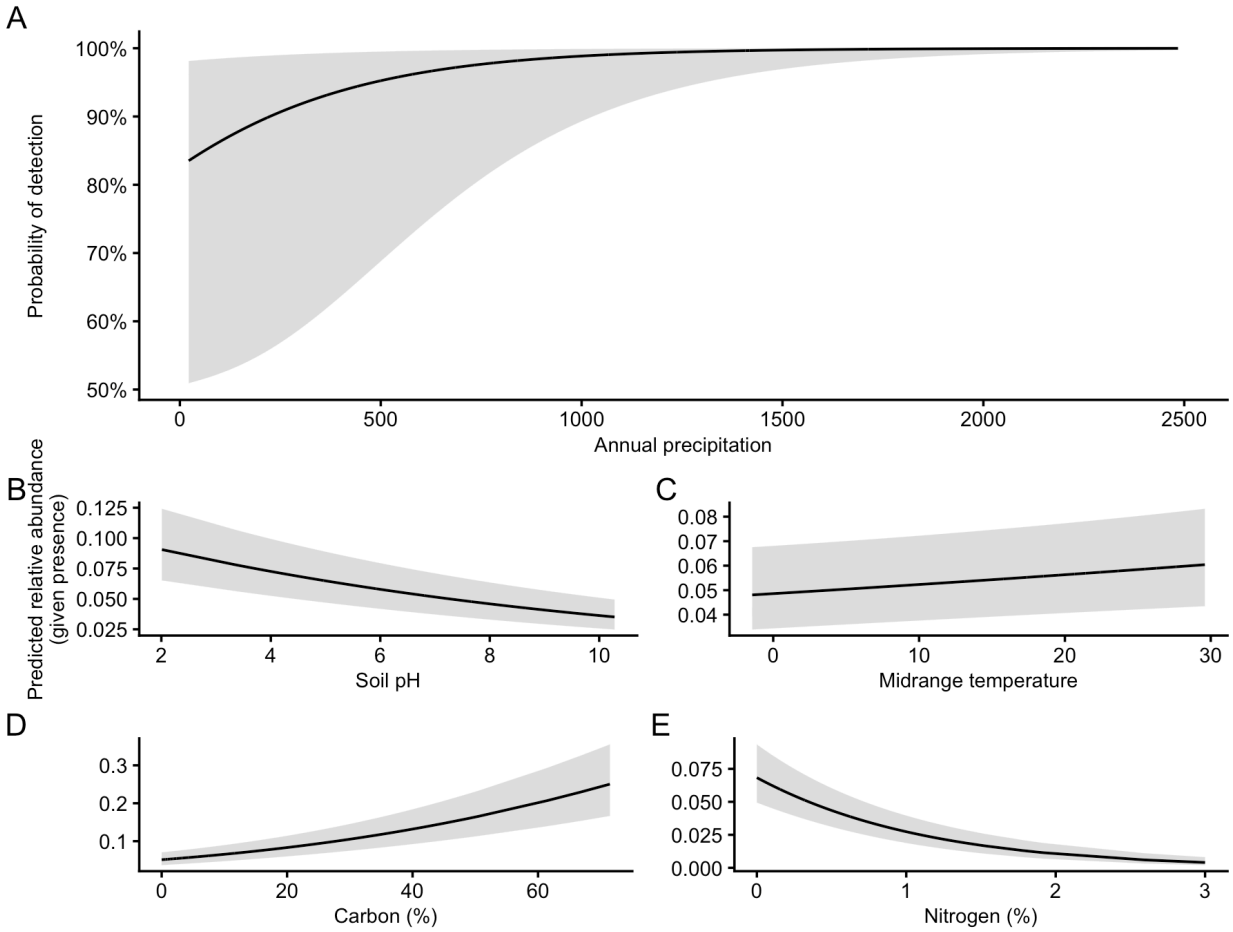

Fig. S3 - Hurdle mixed-effects model results for predictors of *Legionella* occurrence and abundance. (A) Predicted probability of detecting *Legionella* as a function of annual precipitation (mm) from a binomial mixed-effects model. (B-E) Predicted relative abundance of *Legionella* conditional on presence from a beta mixed-effects regression model across gradients of soil pH, midrange temperature, soil carbon (C), and soil nitrogen (N). Lines represent model predictions and shaded regions indicate 95% confidence intervals. Study was included as a random intercept in all models.

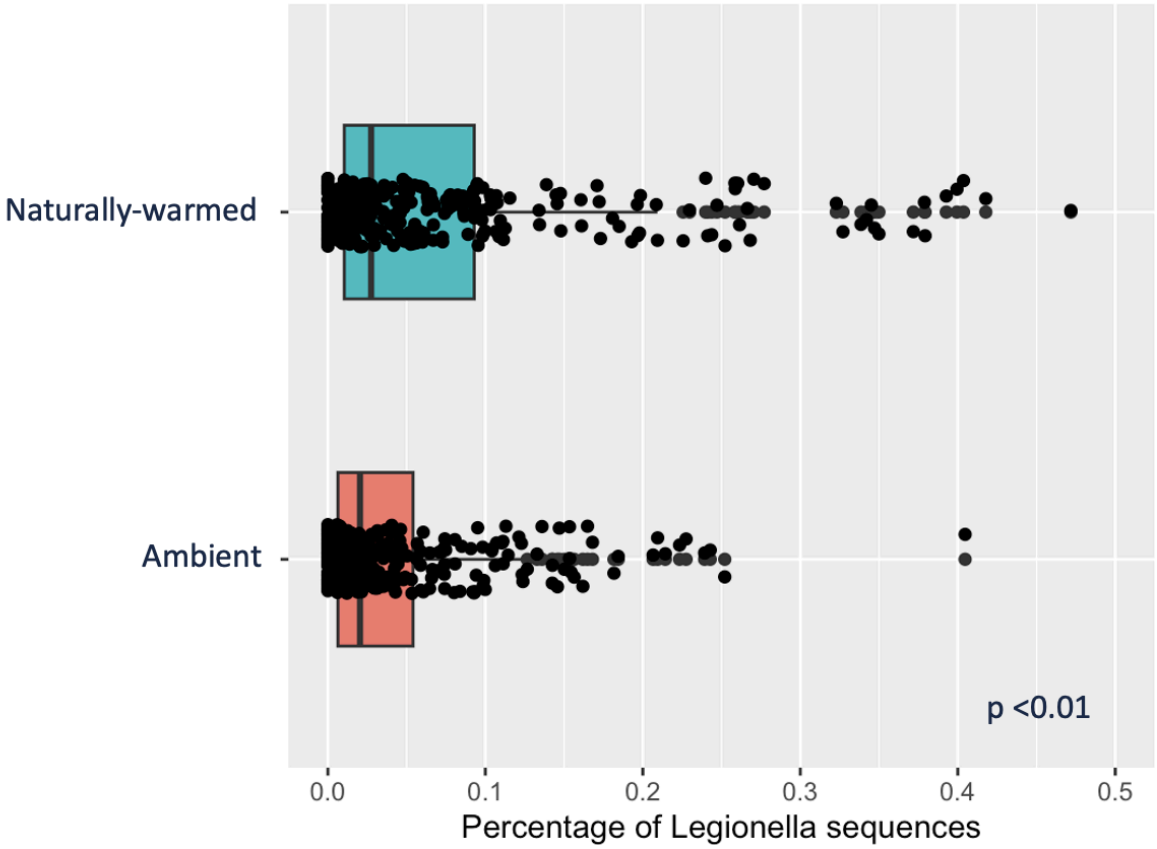

Fig. S4 - *Legionella* relative abundance at geothermally warmed and ambient sites in Iceland. Comparison of *Legionella* relative abundance between naturally warmed soil (+6°C) and nearby ambient sites. Points represent individual samples and boxplots summarize the distribution (median and interquartile range).

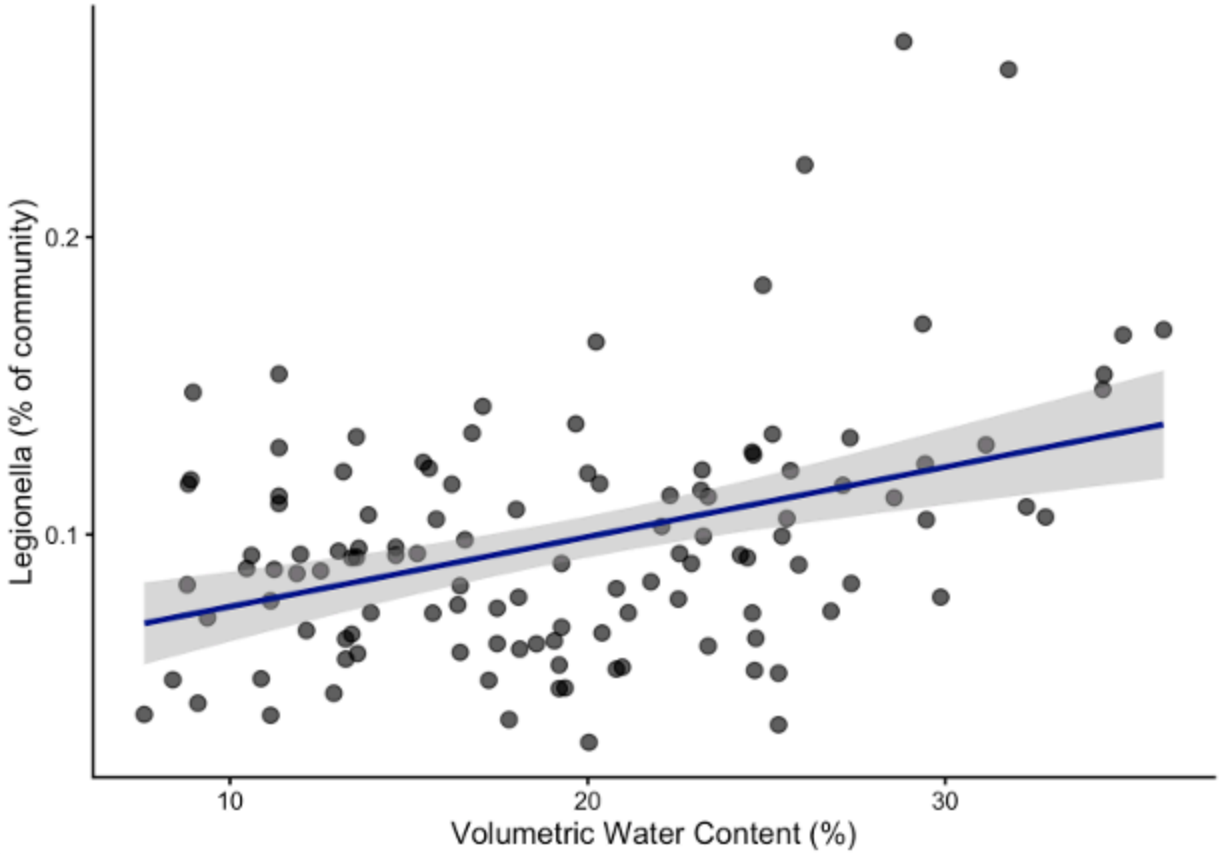

Fig. S5 - Relationship between soil moisture and *Legionella* relative abundance at rainfall-manipulation sites in Wisconsin. Relative abundance of *Legionella* sequences plotted against volumetric water content (%). Points represent individual samples. The line indicates the fitted linear regression and the shaded region shows the 95% confidence interval.

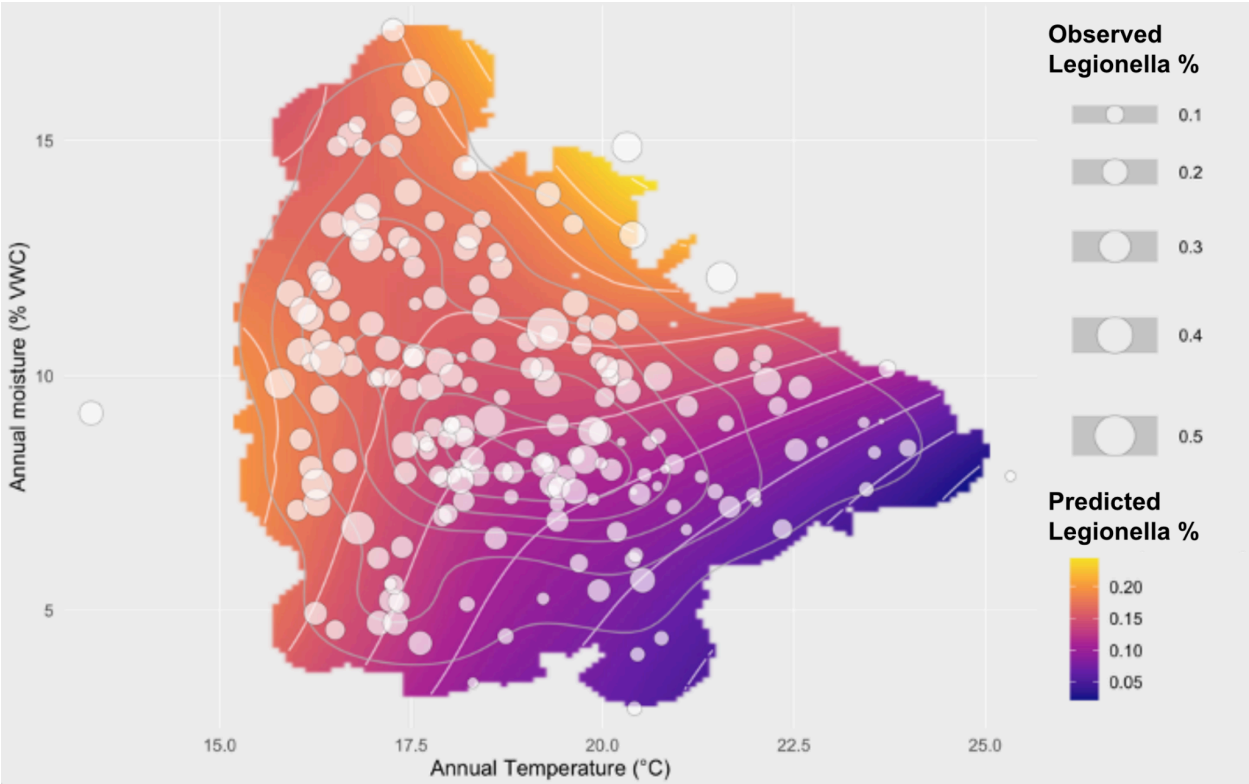

Fig. S6 - *Legionella* relative abundance across temperature and soil moisture gradients at a combined climate-manipulation experiment in Oklahoma. Observed *Legionella* relative abundance is shown as points, with point size proportional to the percentage of *Legionella* sequences detected in each sample. The colored surface represents LOESS-predicted *Legionella* relative abundance, with contour lines indicating predicted abundance levels.

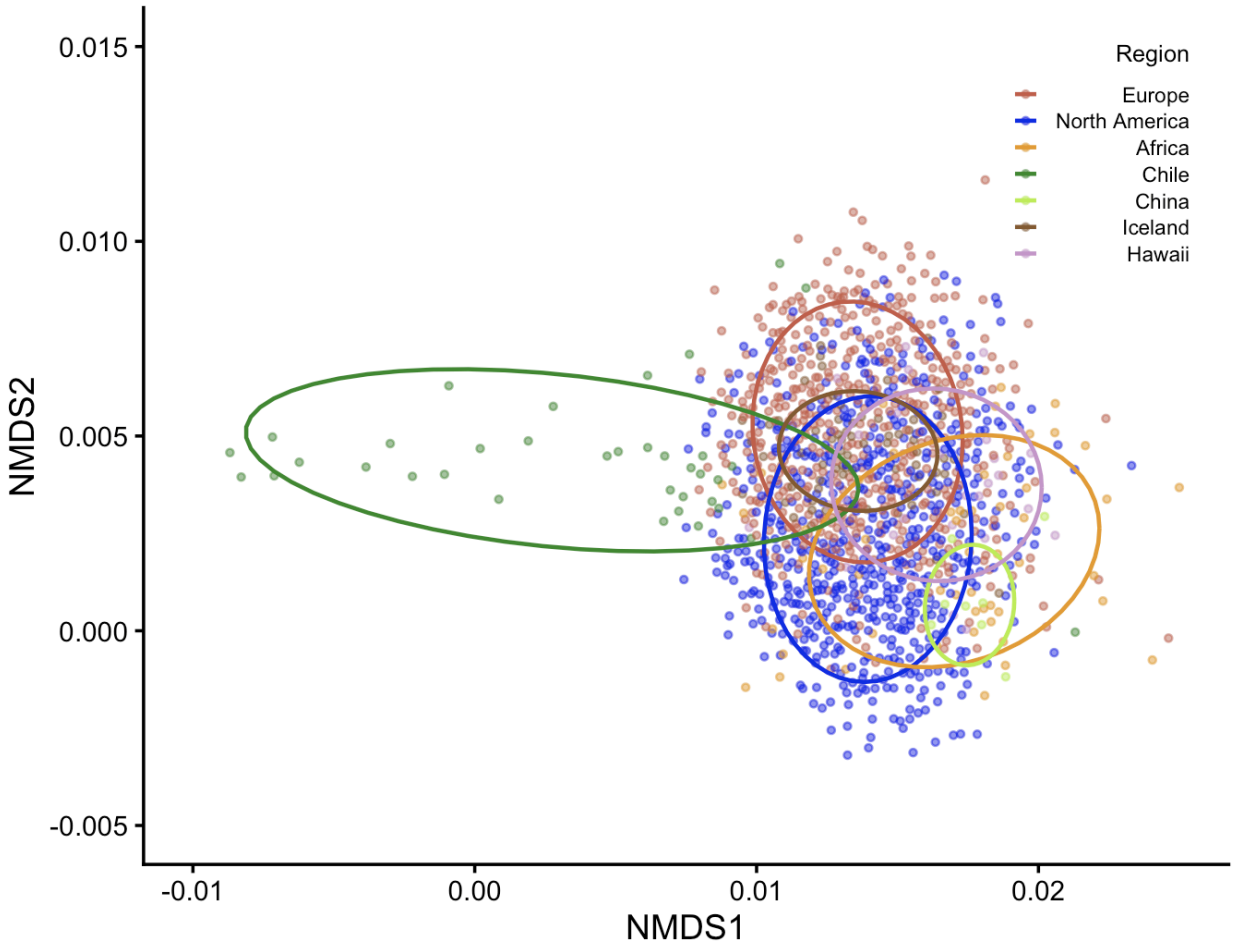

Fig. S7 - Regional structuring of *Legionella* community composition. Non-metric dimensional scaling (NMDS) ordination of *Legionella* community composition based on Bray-Curtis dissimilarities calculated from ASV relative abundances. Each point represents an individual sample and colors indicate geographic region. Ellipses represent 95% confidence intervals around the centroid of each region.

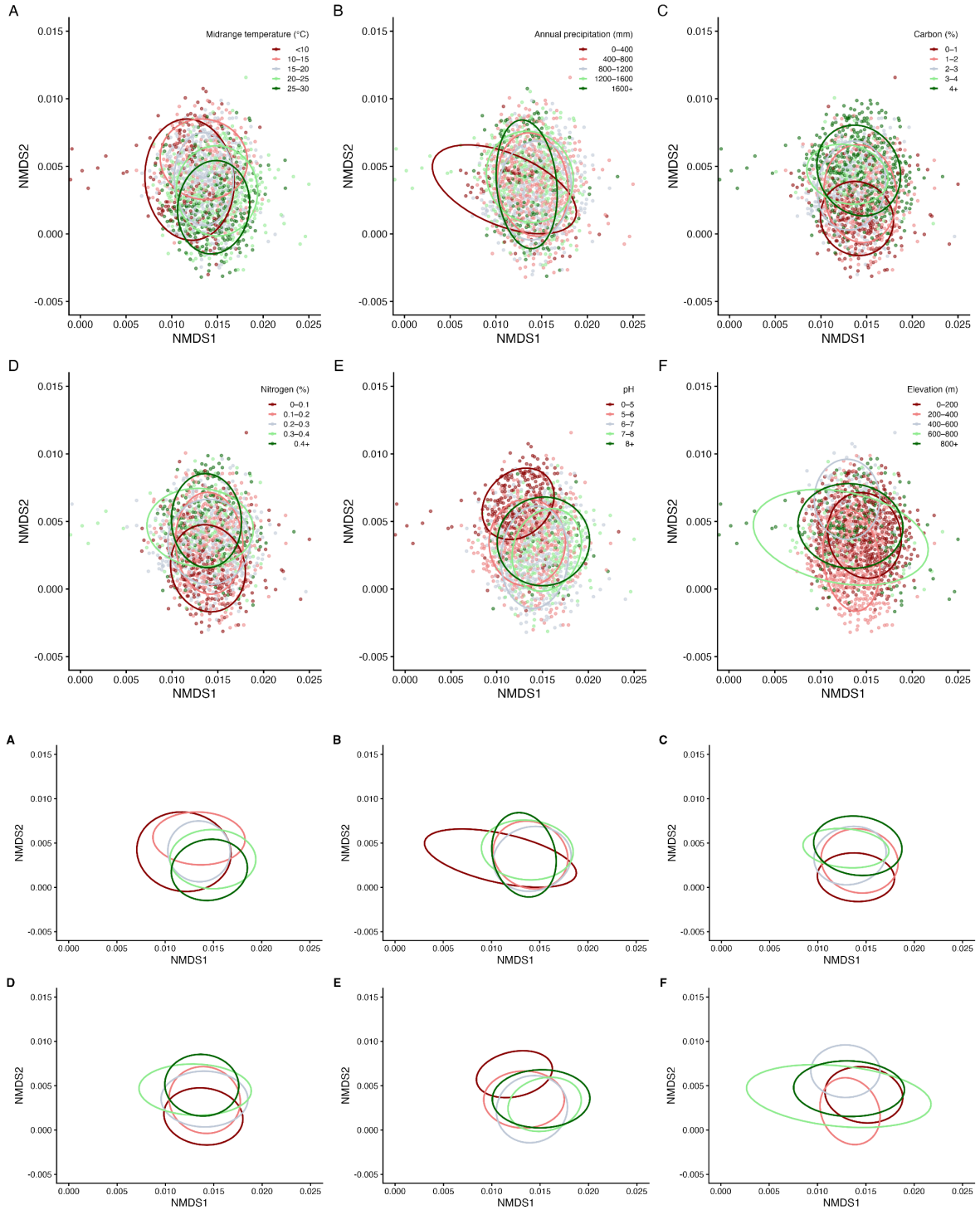

Fig. S8 - Environmental structuring of *Legionella* community composition. NMDS ordination of *Legionella* community composition based on Bray-Curtis dissimilarities calculated from *Legionella* ASV relative abundances. Samples are colored according to environmental bins for midrange temperature, annual precipitation, soil carbon (C), soil nitrogen (N), soil pH and elevation. Ellipses represent 95% confidence intervals around the centroid of samples within each environmental bin.

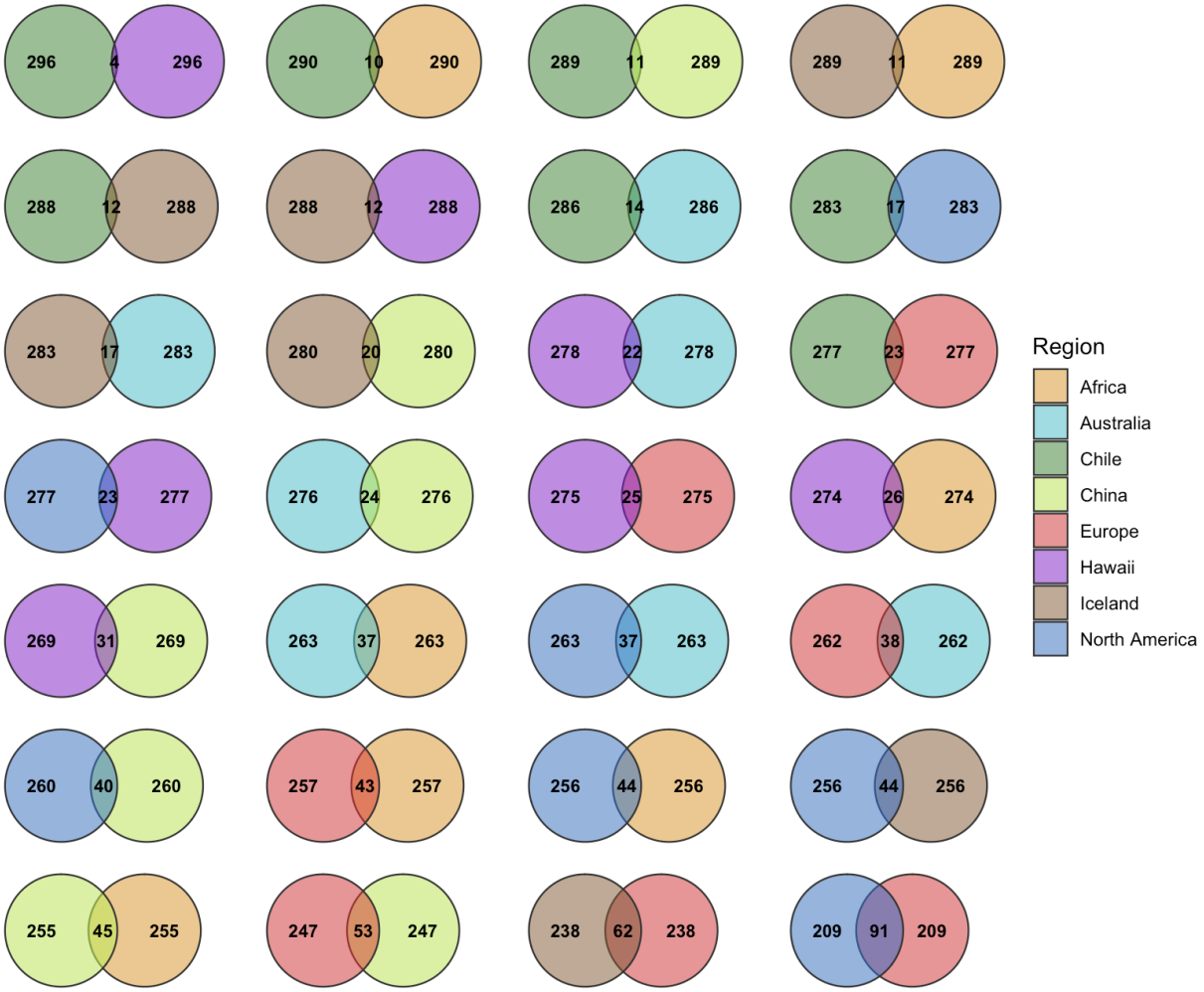

Fig. S9 - Pairwise overlap of dominant *Legionella* ASVs across regions. Venn diagrams illustrate overlap among the 300 most abundant *Legionella* ASVs for each pairwise comparison of the eight regions. Colors indicate regions, and numbers represent the number of dominant ASVs unique to each region (outer sections) or shared between regions (center). Comparisons are arranged from lowest overlap (upper left) to highest overlap (bottom right).

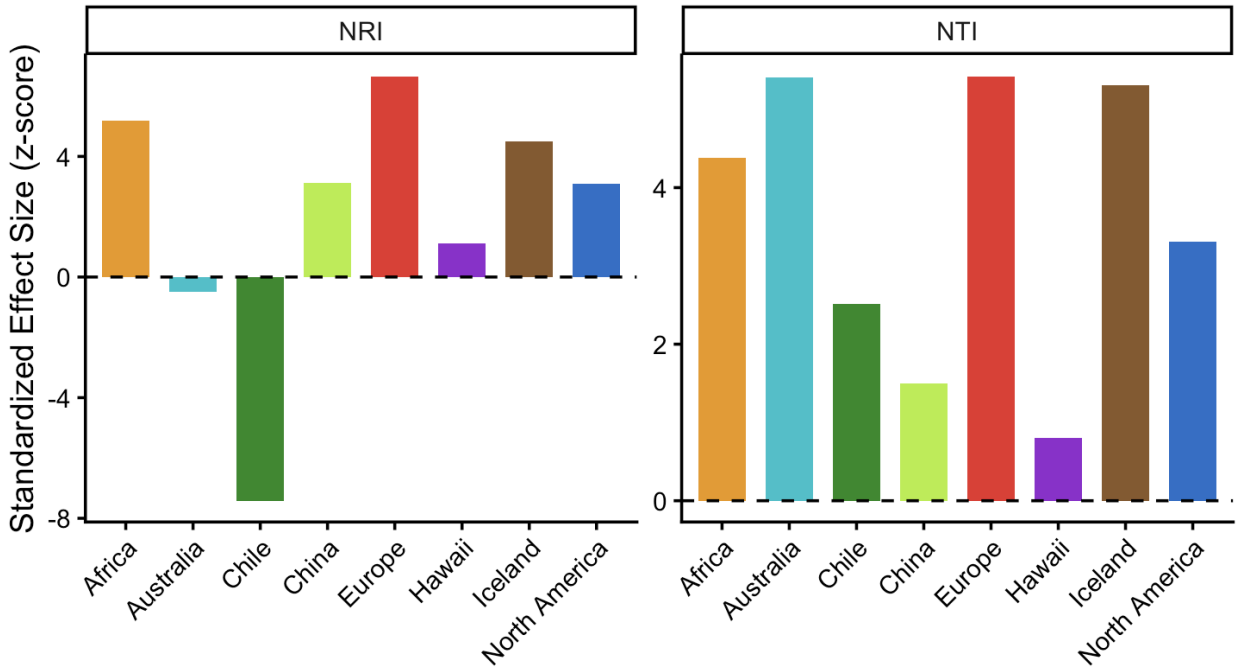

Fig. S10 - Phylogenetic structure of abundant *Legionella* ASVs across global regions. Net Relatedness Index (NRI) and Nearest Taxon Index (NTI) were calculated as standardized effect sizes (SES z-scores) relative to null model communities. Positive values indicate phylogenetic clustering, whereas negative values indicate phylogenetic overdispersion. Bars are colored by geographic region.

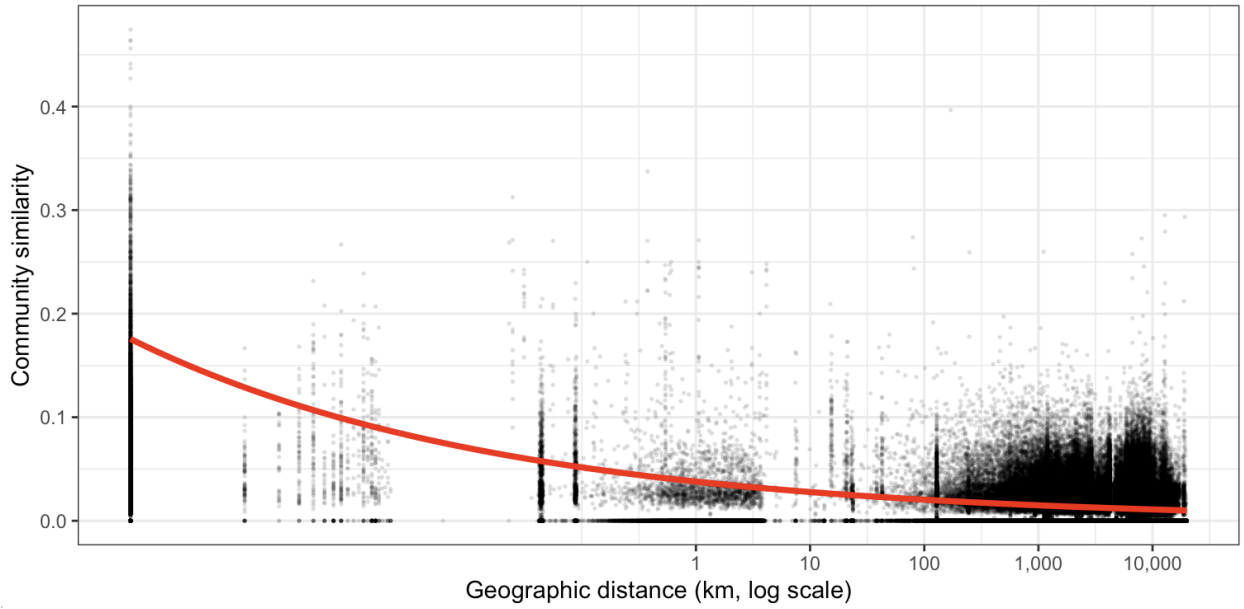

Fig. S11 - Distance-decay relationship of *Legionella* community composition. Pairwise community similarity among samples (1-Bray-Curtis dissimilarity) is plotted against geographic distance (km). Points represent pairwise comparisons between samples, and the red line indicates the fitted distance-decay relationship.

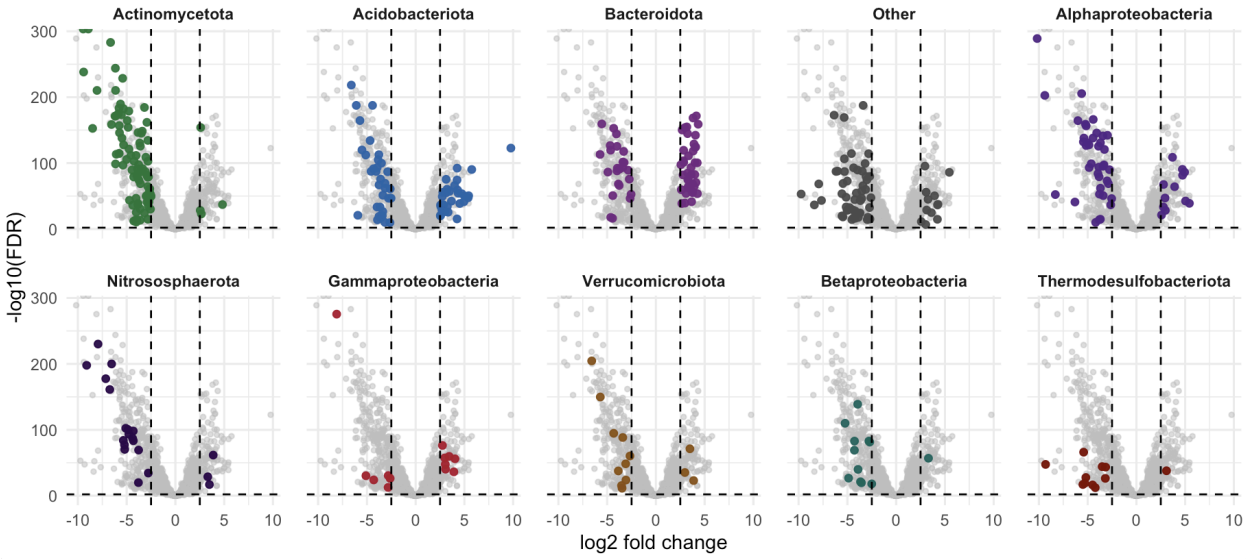

Fig. S12 - Differentially abundant microbial taxa associated with *Legionella* abundance. Volcano plot showing microbial community features (97% OTUs) enriched in *Legionella*-rich versus *Legionella*-poor samples. Samples were classified into upper (>0.062%) and lower (<0.009%) quartiles of *Legionella* relative abundance. Points represent OTUs with $\log_2$ fold change on the x-axis and $-\log_{10}$(FDR) on the y-axis. Positive fold changes indicate enrichment in *Legionella*-rich samples, whereas negative fold changes indicate enrichment in *Legionella*-poor samples. Panels show major taxonomic groups, and colored points indicate OTUs significantly enriched in either direction ($|\log_2 \text{ fold change}| \geq 2.5$, FDR < 0.01).

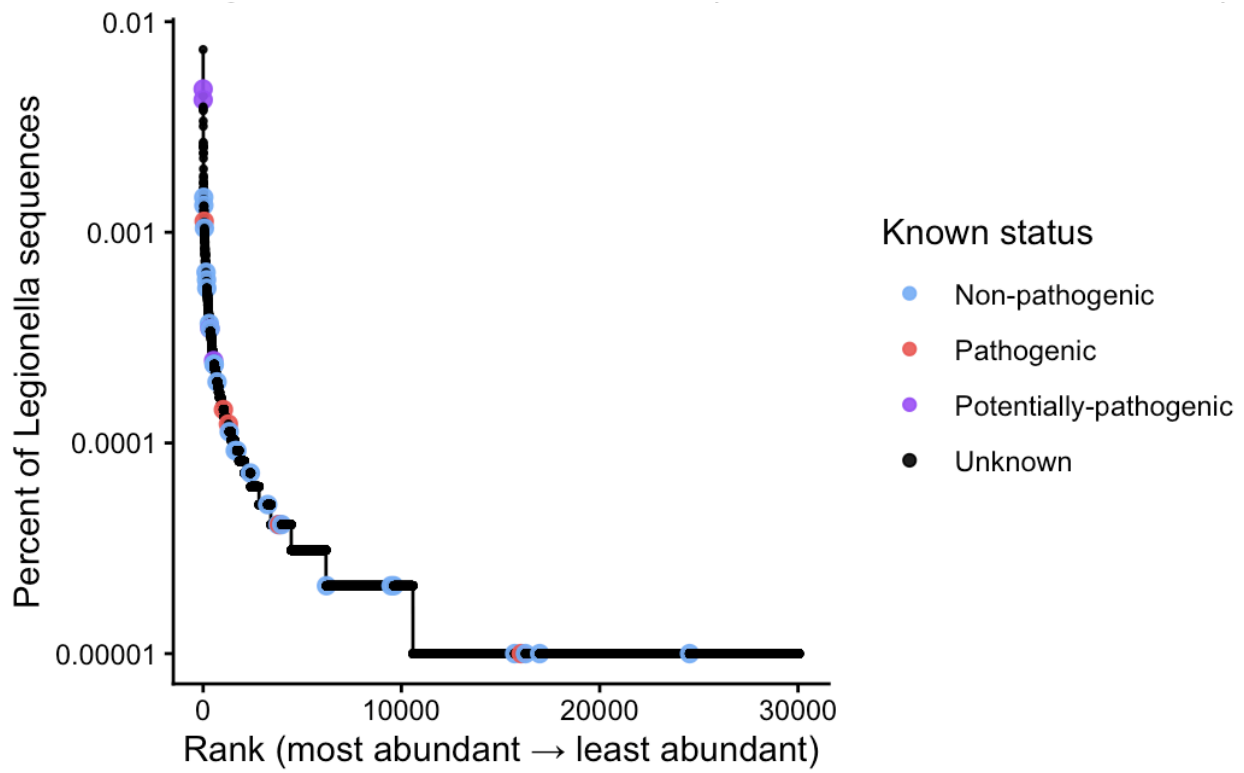

Fig. S13 - Rank abundance distribution of *Legionella* ASVs. Rank abundance plot showing the relative abundance of all *Legionella* ASVs recovered. ASVs are ordered from most abundant to least abundant along the x-axis, and the y-axis indicates the percentage of total *Legionella* sequences represented by each ASV. Points are colored according to pathogenic status based on matches to characterized species (non-pathogenic, pathogenic, potentially pathogenic, or unknown).

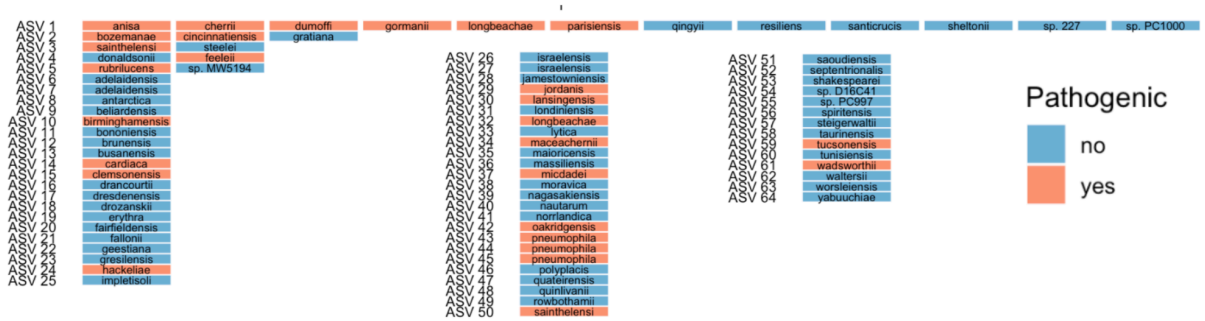

Fig. S14 - Clustering of characterized *Legionella* species based on 16S rRNA gene sequence identity. Clustering of characterized *Legionella* species based on the full 515F-806R 16S rRNA gene amplicon region. Species sharing identical sequences across this region are grouped into the same ASV. Boxes are colored according to whether the species is a known human pathogen.