nature microbiology

Article

https://doi.org/10.1038/s41564-023-01465-0

Life history strategies of soil bacterial communities across global terrestrial biomes

Received: 22 July 2022

Accepted: 8 August 2023

Published online: 5 October 2023

Check for updates

Gabin Piton[1,2] ✉, Steven D. Allison[1,3], Mohammad Bahram[4,5], Falk Hildebrand[6,7], Jennifer B. H. Martiny[3], Kathleen K. Treseder[3] & Adam C. Martiny[1,3]

The life history strategies of soil microbes determine their metabolic potential and their response to environmental changes. Yet these strategies remain poorly understood. Here we use shotgun metagenomes from terrestrial biomes to characterize overarching covariations of the genomic traits that capture dominant life history strategies in bacterial communities. The emerging patterns show a triangle of life history strategies shaped by two trait dimensions, supporting previous theoretical and isolate-based studies. The first dimension ranges from streamlined genomes with simple metabolisms to larger genomes and expanded metabolic capacities. As metabolic capacities expand, bacterial communities increasingly differentiate along a second dimension that reflects a trade-off between increasing capacities for environmental responsiveness or for nutrient recycling. Random forest analyses show that soil pH, C:N ratio and precipitation patterns together drive the dominant life history strategy of soil bacterial communities and their biogeographic distribution. Our findings provide a trait-based framework to compare life history strategies of soil bacteria.

Bacteria impact carbon (C) and nutrient cycling on a global scale[1]. Soil bacterial communities contain enormous, functionally uncharacterized genetic diversity[2,3], which hinders progress in predicting soil microbial responses to global change[4,5]. One approach to describing functional biodiversity is collapsing its complexity into one or more dimensions that capture the dominant associations and trade-offs between traits[6–10]. This multivariate trait space, or life history strategy scheme, provides a framework to compare broad organismal strategies[6,8,10].

While the trait dimensions shaping plant life history strategies are now well established[6], trait associations for soil microorganisms remain less clear. Initially, studies applied the 'competitor', 'stress tolerant' and 'ruderal' (CSR) strategies proposed for plants[7] to soil bacteria[1,11]. This scheme emphasizes trade-offs often observed between traits related to maximizing resource capture (competitor, C), persisting under low resource and stressful condition (stress tolerant, S) and responding rapidly to exploit growing window between disturbances (ruderals, R)[1,7]. Building on the CSR scheme, Malik et al. emphasized differences between microbial yield (Y), resource acquisition (A) and stress tolerance (S) traits as important for soil carbon cycling[12]. While these theoretical papers provide valuable hypotheses on which traits

[1]Department of Earth System Science, University of California, Irvine, Irvine, CA, USA. [2]Eco&Sols, University Montpellier, CIRAD, INRAE, Institut Agro, IRD, Montpellier, France. [3]Department of Ecology and Evolutionary Biology, University of California, Irvine, Irvine, CA, USA. [4]Department of Ecology, Swedish University of Agricultural Sciences, Uppsala, Sweden. [5]Institute of Ecology and Earth Sciences, University of Tartu, Tartu, Estonia. [6]Gut Microbes and Health, Quadram Institute Bioscience, Norwich Research Park, Norwich, Norfolk, UK. [7]Digital Biology, Earlham Institute, Norwich Research Park, Norwich, Norfolk, UK. ✉e-mail: gabin.piton@inrae.fr

https://doi.org/10.1038/s41564-023-01465-0

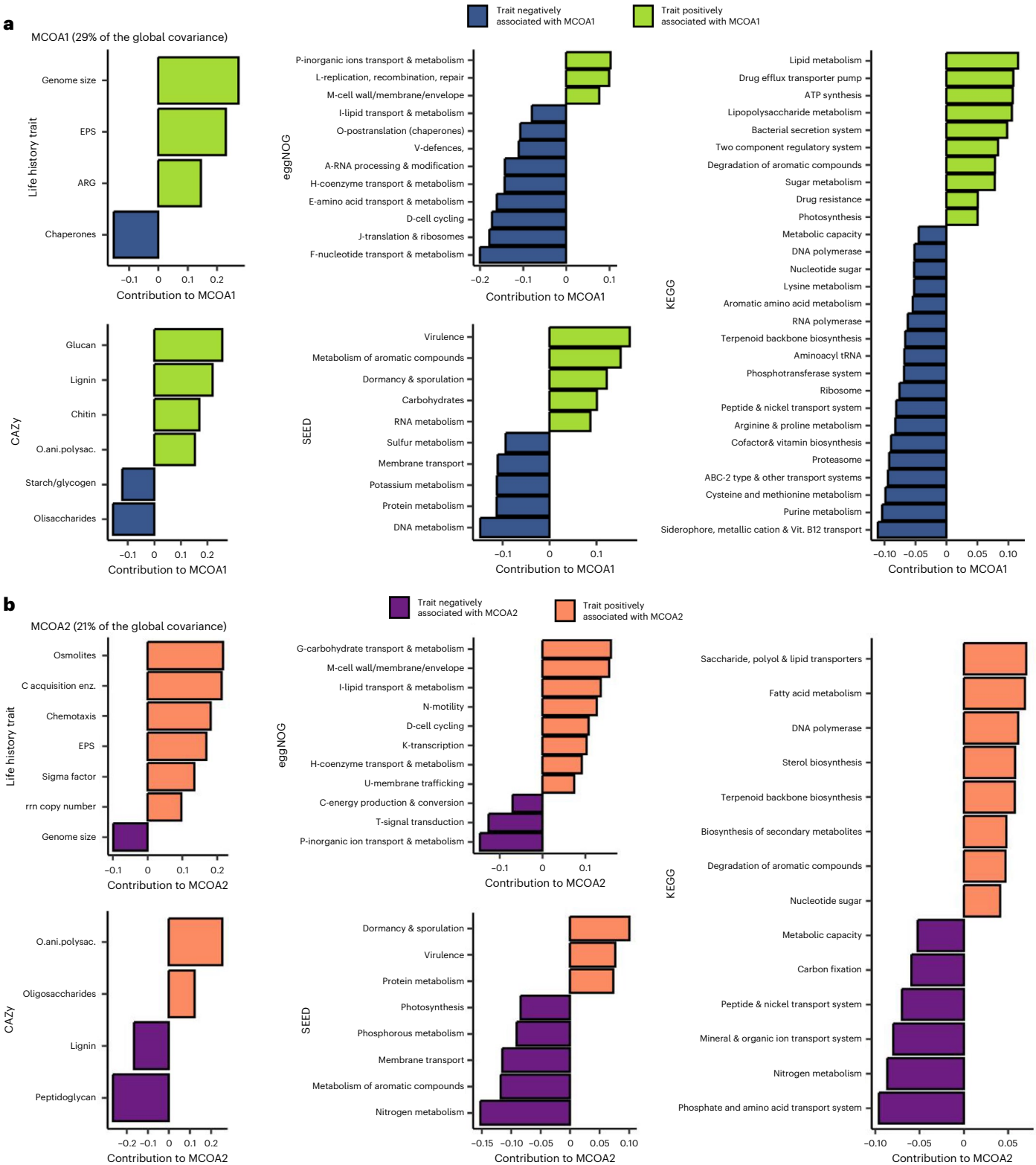


Fig. 1 | Global trait dimensions of soil bacteria metagenomes. Variable contributions to the MCOA summarizing in a common structure (MCOA dimensions 1 and 2) the information shared by 5 CAT databases (Life history trait, CAZy, eggNOG, SEED and KEGG). Only the most important variables with significant correlation ($P < 0.001$) with each dimension are reported in this figure. **a**,**b**, Variable contributions to MCOA dimensions 1 and 2, respectively. Bar colours indicate the direction of the associations between the variable and the MCOA dimensions.

are probably central to soil microbial adaptation, no clear consensus has emerged on the trait dimensions that shape life history strategies of soil bacteria[1,11,12] (Extended Data Table 1). Recently, bacterial cultures isolated from diverse habitats were analysed for genomic and phenotypic traits[13]. This analysis revealed a primary dimension associated with metabolic versatility that was highly correlated with genome size. A secondary dimension separated differences in maximum growth rate and was correlated with variation in ribosomal

gene copy number[14]. However, there is a lot of variation in how well bacterial cultures represent in situ community biodiversity[15–17]. Thus, it remains to be tested whether the life history strategies of soil bacterial communities match either the theoretical or culture-based predictions of key trait dimensions.

One advantage of studying the traits of microorganisms vs those of larger organisms is the ease with which collections of their traits can be measured at the community level. Community aggregated traits (CATs)[18] represent the average functional profile of the community emerging from the combination of organisms' traits and community composition (similar to the idea of community-weighted means of traits proposed for plants)[19,20]. Hence, it is important to note that while suggestive, such CAT patterns do not directly inform on within-organism trade-offs. Nevertheless, CATs described using metagenomic sequences offer a way to characterize shifts in organismal strategies dominating bacterial communities in situ (for example, refs. 21,22) and thus offer an approach to test theoretical life history strategy schemes to in situ microbial communities. In addition, information on the dominant strategy in a bacterial community might be used to predict the response of this key group to environmental changes for global biogeochemical cycles[4,18]. Elucidating the trait dimensions that shape the dominant life history strategies of soil bacteria would thus provide a framework for comparing soil bacterial communities and developing generic predictions in soil microbial ecology[14].

In this study, we used a global dataset of soil metagenomic sequences from major biomes to quantify key trait dimensions of soil bacterial communities. We then identified primary environmental factors partitioning the trait dimensions and projected global biogeography. Finally, we compared the emergent life history strategies with theoretical and culture-based predictions.

Results

The trait dimensions of soil bacterial communities

Using a multitable co-inertia analysis (MCOA), we found that two dimensions captured half of the overall variation in metagenomic community aggregated traits (CATs). MCOA1 and MCOA2 captured 29% and 21% of metagenomic trait variation (Fig. 1 and Extended Data Fig. 1), while MCOA 3 and MCOA4 explained 16% and 10% of this variation, respectively (Extended Data Figs. 1 and 2). The MCOA revealed the most important associations between traits (Figs. 1 and 2), including traits previously associated with life history strategies (Figs. 1–3).

Average genome size had the highest contribution to MCOA1 (Fig. 1a) with an R^2 of 0.64 for the positive correlation between average genome size and MCOA1 (Extended Data Fig. 3a). Mapping coverage decreased along this dimension (Extended Data Fig. 4). The lower end of this dimension was characterized by bacterial communities with higher relative abundance of genes for primary metabolism (that is, essential process for survival and growth) and C acquisition machinery (Fig. 1). In these communities, carbon acquisition enzymes involved in depolymerization of oligosaccharides were favoured over enzymes targeting polysaccharides. This oligosaccharide-degradation enzyme class was dominated by the beta-glucosidases GH1, GH2 and GH3 CAZy families. Finally, chaperones were overrepresented. Thus, the lower end of MCOA1 was defined by communities with a streamlined metabolism (Fig. 2).

The upper end of MCOA1 defined bacterial communities with a large genome and more complex metabolism and resource acquisition strategies (Figs. 1 and 2). The enriched genes allowed for degradation of complex polysaccharides from fungi, animals and plant lignin. There was also overrepresentation of genes for direct plant pathogenic interactions and negative interactions with other microorganisms. Finally, communities carried a higher proportion of genes encoding for exopolysaccharides (EPS) production, dormancy and sporulation, membrane and DNA repair (Figs. 1 and 2). These functions were generally present in lower relative abundance in communities with small genomes at the opposite end of MCOA1. Thus, the first trait dimension captured functional variation associated with genome size and expanded 'metabolic capacities' (Fig. 2).

Bacterial communities differentiated along a second dimension (MCOA2) but only when they increased their 'metabolic capacities' along the first trait dimension (MCOA1), shaping a triangle (Fig. 2). This distribution indicated that bacterial communities with low 'metabolic capacities' and small average genome size are constrained along the second dimension. MCOA2 separated communities according to genomic traits for 'environmental responsiveness' and 'nutrient recycling' (Fig. 2). Communities associated with the lower end of MCOA2 were enriched in mineral and organic N and P assimilation genes (Figs. 1 and 2). Furthermore, there were also higher relative frequencies of genes encoding for bacterial necromass degradation including peptidoglycan. Communities at the upper end of MCOA2 were defined by an ability to respond to a complex set of environmental cues. This was manifested by an increased presence of genes encoding for activity regulation, resistance to environmental stress, foraging of beneficial conditions, fast growth (rrn copy), and building and repairing the cell membrane (Figs. 1 and 2). The communities were also enriched in genes encoding for carbohydrates metabolism of simple substrates such as starch, glycogen and oligosaccharides. Thus, the second trait dimension captured a gradient in average environmental responsiveness that was positively associated with specialization in simple carbon substrate metabolism and negatively associated with nutrient assimilation and recycling capacities (Fig. 2).

Drivers of trait dimensions

Using random forest analyses, we next found that common soil environmental factors distributed the soil bacterial community along global trait dimensions. Random forest models based on soil pH, precipitation and C:N ratio could predict most of the variation in MCOA1 and MCOA2, with an R^2 of 0.80 and 0.58, respectively (Extended Data Fig. 5). Mean decrease in mean square error (%MSE) and R^2 calculated on the basis of a 10-fold cross-validation of the random forests indicated that soil pH and annual precipitation are the most important predictors for both MCOA1 and MCOA2. However, the two dimensions showed different response patterns to these variables, with MCOA1 decreasing with soil pH but increasing with annual precipitation whereas MCOA2 decreased with both soil pH and annual precipitation, leading to unique positions along MCOA1 and MCOA2 depending on the combination of pH and annual precipitation (Figs. 3 and 4). MCOA1 and MCOA2 were also driven by precipitation seasonality, whereas soil C:N ratio controls only MCOA1 (Figs. 3 and 4, and Extended Data Fig. 5). Next, we projected the global variation in the trait dimensions using these random forests (Fig. 4b,d) and global soil and climate databases. It is worth noting that this broad spatial resolution map, using averaged conditions across large spatial units, showed high consistency with values observed locally in our samples (Extended Data Fig. 6). Thus, the identified trait dimensions showed a clear global biogeography.

The first trait dimension (MCOA1) mainly separated arid alkaline regions from more acidic and wet ones. More precisely, bacterial communities characterized by a small genome size (that is, low MCOA1 value) were enriched under neutral to alkaline pH, low C:N ratio, low annual precipitation but high precipitation seasonality (Fig. 4a). Conversely, communities with larger genome sizes (high MCOA1 value) were found in more acidic soils as well as soil with higher C:N ratio and climate with elevated stable precipitation (Fig. 4a). Globally, these environmental controls predicted low MCOA1 coordinates (<−1) under arid and semi-arid climates at tropical and subtropical latitudes as well as in the steppe zones of central Asia and North America (Fig. 4b). Conversely, high MCOA1 coordinates (>1) were seen in equatorial forests as well as some temperate zones in northern Europe,

https://doi.org/10.1038/s41564-023-01465-0

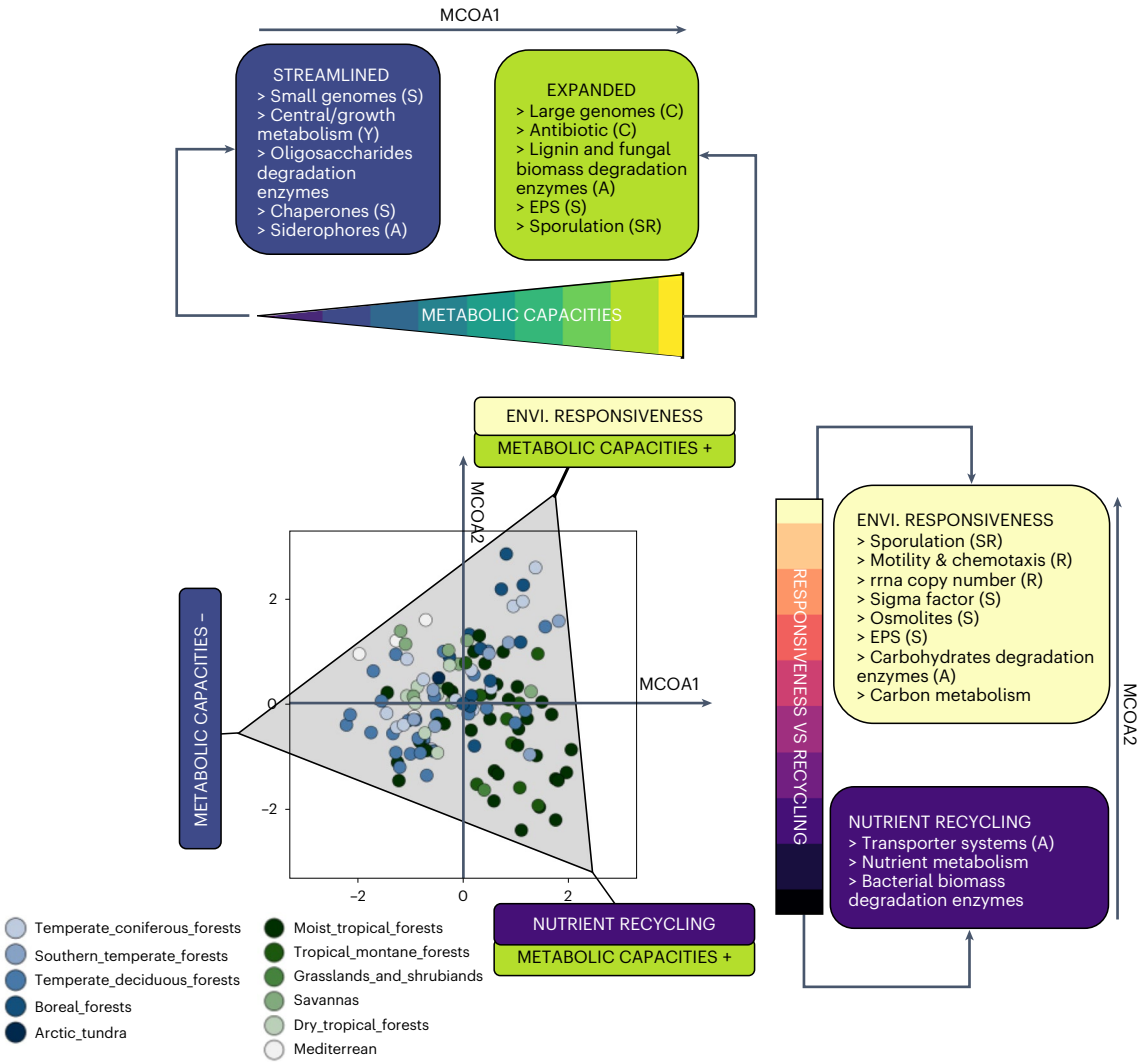


Fig. 2 | The global life history strategies of soil bacterial communities. Two-dimensional trait space from a MCOA depicting trait associations across soil bacterial communities, with traits inferred from enriched genes in bacteria metagenomes. Dots represent the positions of the 128 bacterial communities used in this study along these two dimensions. In the trait lists, letters in brackets represent how CSR (competitor, stress tolerant, ruderal) and YAS (high yield, resource acquisition, stress tolerance) strategies have been associated with these traits in previous theoretical works (Extended Data Table 1).

Western Canada, New Zealand and south Chile. Steep MCOA1 gradients were estimated to occur in regions separating arid and wet zones, and medium coordinates (−1 < MCOA dimension 1 < 1) also covered most of temperate and high latitudinal regions (Fig. 4b).

The second trait dimension (MCOA2) separated regions with high but stable precipitation from places with more seasonal climate and extremely acidic soils. The lower end of MCOA2 covered most high precipitation regions (>2,500 mm) including equatorial zones of South America and Asia, and wet Europe and North American temperate zones. Medium-high coordinates (0 < MCOA2 < 1) covered most of the globe, characterizing all tropical-dry, semi-arid and subarctic regions. The projection of this dimension (Fig. 4d) predicts very high coordinates (MCOA2 > 1) under limited regions of subtropical and high latitudes combining low annual precipitation (<1,000 mm) and very acidic pH (<4).

Finally, we found that trait differences (defined on the basis of Euclidian distances along the two first dimensions of the MCOA) were significantly correlated with Unifrac phylogenetic distances (R^2 = 0.32, Extended Data Fig. 7). Communities with average genome size below their median values depicted a correlation between trait and phylogenetic distances that is significantly steeper (slope difference: P = 0.00116) and tighter (R^2 = 0.46) than that of communities with larger genomes (R^2 = 0.15, Extended Data Fig. 7).

Discussion

Our study describes two dominant dimensions of community aggregated traits variation across soil bacterial communities (Figs. 2 and 3). In this trait space, communities are constrained in a triangle of three opposing life history strategies: low metabolic capacities; metabolic capacities expanded for environmental responsiveness; metabolic capacities expanded for nutrient recycling. These life history strategies incorporate traits previously identified as CSR strategies[1,11,12] (Extended Data Table 1). Moreover, these fit into a triangle, similar to the original CSR model[7,23] (Figs. 2 and 3), which suggests that the constraints on bacterial strategies might scale up to community level. Also consistent with CSR theory, both trait dimensions of our study capture competitor traits that trade-off with traits of the other strategies. However, while one strategy generally dominates the traits of each end of the trait dimensions, our aggregated profiles often combine traits that had been associated with different strategies. In particular, one or more stress tolerance traits are part of all profiles (Figs. 2 and 3). We hypothesize that these combinations indicate either that the

https://doi.org/10.1038/s41564-023-01465-0

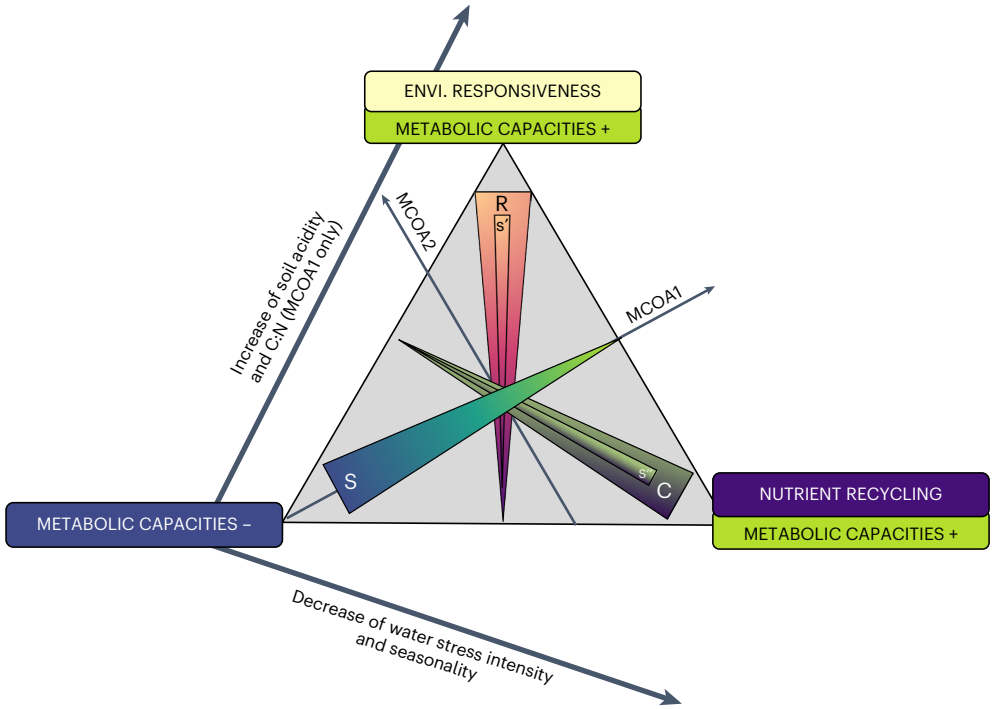


Fig. 3 | Hypothesized role of C, R and S traits in shaping the life history strategy observed at the community level and associated environmental gradients. Details of CSR traits association are provided in Figs. 1 and 2. S, S′ and S″ show that different S traits are associated with each dimension and detailed in Figs. 1 and 2.

communities are composed of taxa with different strategies or that the majority of bacteria living in soil need stress tolerance traits to survive in this challenging environment.

Bacteria with streamlined metabolism dominate the low end of the 'metabolic capacity' dimension. The genomic traits of these bacterial communities with small average genome size have only few matches with previous description of stress tolerance strategy (Extended Data Table 1)[1,11,12]. However, the clear association to arid biomes that we observed suggests that the streamlined bacteria are associated with stress tolerance strategy. This is consistent with recent studies showing that genome streamlining can play a role in adaptation to environmental stressful conditions (for example, refs. 24,25). In particular, ref. 24 used a joint species distribution model to show that soil bacteria with small genomes are selected under arid environments, as seen here. Moreover, these streamlined communities were associated with some low environmental constraints on resource acquisition (low soil C:N ratio and pH near neutrality as observed in ref. 26) that might also reduce fitness benefits for gaining new capabilities[27]. Thus, genome streamlining and associated change in gene frequency might be central in soil bacteria stress tolerance, especially in arid biomes.

Cells with larger genomes and a more complex metabolism dominate the other end of the 'metabolic capacity' dimension. The associated variation in the functional gene frequency that we observed is also consistent with previous studies reporting that genome expansion in free-living bacteria is driven by gene additions encoding for new metabolic capabilities or regulation[14,28]. Large genomes, high catabolic diversity and antibiotic resistance genes observed for this life history strategy were previously attributed to a competitor strategy (Extended Data Table 1)[1,11]. This supports the idea that complex substrates acquisition is a key trait of competitors as suggested by Malik et al.[12]. Consistent with competitor traits, these attributes are favoured under stable and wet climates, reducing the benefits of desiccation stress traits and possibly leading to intense resource competition[7]. We also detected an enrichment in traits associated with sporulation and exopolysaccharides production, two traits often associated with stress tolerance or ruderality (Extended Data Table 1) that might also improve tolerance to antimicrobial compounds or nutritional constraints for such competitor profile[29,30]. Together, the first trait dimension appears to represent a gradient from stress tolerant communities with small genomes to communities dominated by bacteria with increased 'metabolic capacities' associated with other strategies, especially competitors.

When average genome size increases, bacterial communities differentiate along the second dimension with opposing profiles of increased capacities for either 'environmental responsiveness' or 'nutrient recycling'. At the high end of this dimension, communities with high 'environmental responsiveness' shared numerous genomic features tied to both the ruderal and stress tolerance strategies (Extended Data Table 1). These include traits to resist stress, sense favourable environmental conditions, activate fast growth and for C acquisition. The reduced and fluctuating precipitation patterns associated with this profile are also consistent with original descriptions of these strategies[1,7]. At the opposite end of this second dimension, bacteria specialized in 'nutrient recycling' show a resource acquisition strategy with a high number of transporters and bacterial biomass (peptidoglycan) recycling, and a higher investment towards nitrogen and phosphorus metabolism compared with carbon metabolism. Microbial mineralization activity and biomass turnover release nutrients and necromass into soil, which this profile seems optimized to recycle. Such traits might reflect a strategy that emphasizes resource use efficiency and increased competitiveness for nutrients[11,12]. Further, the environmental parameters associated with this life history strategy (medium-low pH, high precipitation and low seasonality) are the most favourable for resource acquisition[31], biomass turnover and yield[32,33], reinforcing potential selection for competitor traits[7]. In summary, the second trait dimension reflects communities with increased metabolic capacities associated with either a combination of stress tolerance and ruderal traits that maximize their responsiveness or a reinforcement of competitor traits that favour nutrient recycling.

Overall, our dimension of 'metabolic capacities' matches the versatility dimension described in ref. 14 across cultured bacterial taxa, with both studies supporting the notion that genome size plays a central role in differentiating bacterial strategies. Our dimension

https://doi.org/10.1038/s41564-023-01465-0

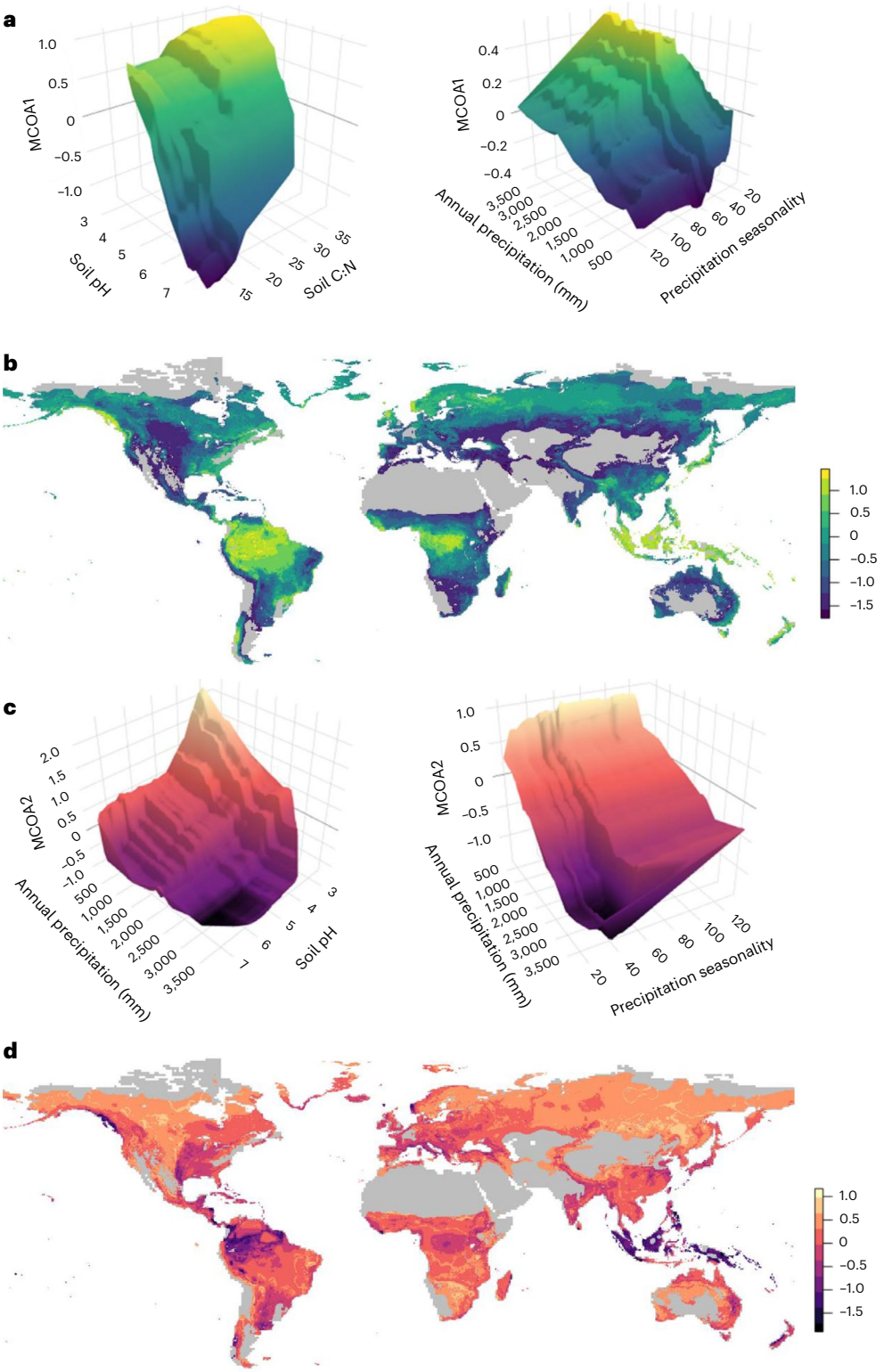


Fig. 4 | Environmental control and global-scale projection of bacterial communities' coordinates along MCOA dimensions 1 and 2. a,**c**, Random forest partial dependence plots describing relationships between bacterial communities' coordinates along MCOA dimension 1 (**a**) and 2 (**c**) and their most significant environmental predictors (Extended Data Fig. 5). **b**,**d**, Random forest predictions for MCOA dimension 1 (**b**) and 2 (**d**) projected across the globe using a broad resolution map of mean soil and climate conditions (1,600 × 1,200 pixel), with land out of the dataset range in grey. Colour bars represent the predicted coordinates along MCOA dimension 1 (**b**) and MCOA dimension 2 (**d**). SoilGrids v.2.0 was used for soil properties and Worldclim2 for climate variables. Accuracy of the prediction was verified by 10-fold cross-validation of the random forest (Extended Data Fig. 5) and by comparing the predicted values of the broad resolution projection with local observations (Extended Data Fig. 6).

opposing 'environmental responsiveness' and 'nutrient recycling' also shows some consistencies with the second trait dimension described in ref. 14 capturing a rate–yield trade-off, with rrn copy number as a principal trait. Indeed, as discussed above, the traits of the 'nutrient recycling' profile may favour growth yield, and high 'environmental responsiveness' is associated with higher rrn copy number. However, these variations in rrn copy numbers have only a limited importance in the second trait dimension of our study, contrasting with the observations in ref. 14 for cultured bacteria from diverse habitats. This could be explained by the constraint range of this trait in soil. Indeed, variation in average rrn copy number observed across communities in our study is highly constrained (1 to 1.5 copies, Extended Data Fig. 3). These observations are consistent with a previous report stating that most soil bacteria have less than 2 rrn copies, whereas bacteria from other environments can have up to 15 copies[34]. Further, variation in the average rrn copy number of whole communities will be more constrained than variation across individual isolates within the community; indeed, some bacteria with more copies may be present in the soil community, with their populations increasing during resource flushes (for example, ref. 35). In the oligotrophic environment of the soil, our results suggest that increased capacity to recycle resources efficiently, sense favourable conditions and survive or escape stressful ones represent more common adaptations for bacteria than growing more rapidly. Investigating variation in these traits across taxa in soil and their distribution within communities represents a challenging but fascinating perspective to disentangle how trait dimensions across taxa scale up to the community level. Overall, the life history strategies of soil bacteria that we described using aggregated traits at the community level show some important consistencies with life history strategies described across bacterial taxa from various habitats, but also highlights some specificities and challenges associated with the soil environment.

Soil bacteria remain poorly characterized, with a limited number of reference genomes and gene functional characterization[36,37]. This reduces annotation coverage of metagenomic data and can limit analysis conclusions. In our study, the proportion of reads annotated (between 5 and 15%, depending on the database) were in the range of what is commonly obtained from soil metagenomes[38]. Our usage of stringent quality filtering criteria in the annotation[2] also reduced the annotation coverage but increased annotation confidence. Finally, the proportion of unannotated reads is increased by sequencing error and our usage of short-read sequencing technology and read-based profiling (as opposed to assembly-based profiling with better annotation but very limited representativity of the community). Our annotation coverage also showed a decrease with genome sizes, as reported across taxa[36,37]. However, unannotated genes probably belong to accessory genes and not to core metabolism genes that are well represented in current databases[37]. Thus, we can expect that increased annotation of large genomes would have accentuated evidence for our conclusion that our first trait dimension captured an increase in metabolic capacities. Overall, our trait dimensions are expected to capture at least the functional variations associated with core metabolism and provide some first elements about functional genes associated with expansion of metabolic capacities.

We showed that communities with similar life history strategies tend to be phylogenetically closer, supporting a certain phylogenetic conservatism of the genomic traits shaping life history strategies[39]. However, this relationship weakens as genome size and metabolic capacities expand (Extended Data Fig. 7). This suggests that metabolic expansion during different evolutionary histories can converge to similar life history strategy[40]. Hence, phylogenetic distance becomes a poorer predictor of difference in life history strategies for soil bacterial communities with large genomes.

The biogeography of dominant life history strategies in soil bacterial communities is mainly driven by the combinations of soil pH and precipitation patterns across the globe. These environmental factors impact stress and competition intensity for soil bacteria, either through direct effect on their physiology and interaction[41–43] or indirectly through their modification of abiotic (for example, solubilization of toxic ions Al^{3+}) and biotic (for example, plant and fungal communities) characteristics of the ecosystem[44–46]. The environmental distribution of life history strategies suggests that bacteria expand their metabolic capacities to deal with conditions associated with increasing soil acidity and annual precipitation until a certain level (Fig. 3). Then, expansion of 'metabolic capacities' increases either 'environmental responsiveness' to survive under more extreme pH and fluctuating precipitation or 'nutrient recycling' to be competitive under higher precipitation levels. These global effects of pH and precipitation are consistent with previous studies of soil bacteria biogeography[3,26,47] and provide some new information on the traits associated with these environmental factors.

Our global projection (Fig. 3b,d) aims at giving a picture of the general biogeographic patterns in the functional profiles of soil bacterial communities. However, it is important to note that transposition of our trait dimensions at local scale will need further investigation. Values predicted for these broad resolution maps can be dissociated from the local situation if its conditions highly differ from the regional mean (Extended Data Fig. 6) and should be used with caution. Despite outstanding issues that remain open, our study demonstrates how metagenomic approaches can provide substantial advance in our understanding of microbial community functioning. Altogether, our results suggest that land use and climate changes impacting soil pH and precipitation gradients at biogeographic scale might be central in shaping future functional potential of soil bacterial communities and thus global biogeochemical cycles.

Methods

Soil sampling and characteristics

We analysed a global dataset of 128 metagenomes each from unique soil samples distributed across continents and latitude (Extended Data Fig. 8)[2]. We selected this dataset for our analysis because of its coverage and its use of a highly standardized protocol that: (1) sampled topsoils in spatially independent sites across the globe selected to represent all of the most important vegetation types; (2) analysed soil chemistry and metagenomes[2]. All samples were processed using similar standardized protocols for their chemistry (carbon, nitrogen, phosphorus content and pH_{H20}) and metagenome (see ref. 2 for protocol details). We checked the global environmental coverage by comparing variation of the main environmental variables (mean annual temperature (MAT), mean annual precipitation (MAP), soil pH and net primary productivity (NPP)) in our dataset with global variation from the Atlas of the Biosphere (https://sage.nelson.wisc.edu/data-and-models/atlas-of-the-biosphere/). This showed an almost complete global coverage, with only extreme MAT at very high latitude (below −11.33 °C) and in Sahelian Africa (above MAT 27.97 °C) as well as very high pH (higher than 7.76) characterizing some parts of North Africa, West Asia and Himalaya missing in our dataset (Extended Data Fig. 8). As far as we know, when we conducted this analysis, this dataset was the only one available with such precise characterization of the soil environment done on the same sample as that used for shotgun metagenomic analysis, making this dataset the most robust for our objective of assessing environmental drivers of metagenomic profiles. Nevertheless, potential to extend environmental range by adding all (excluding agricultural and contaminated) soil metagenomes available (accession date 28 January 2021) from the main sequence repositories MG-RAST[48] and IMG:M[49] was also tested. This indicated that adding these data would not have extended environmental range (except for a few samples from very cold sites with mean annual temperature lower than −11.5 °C available on MG-RAST) and would have greatly decreased precision of soil properties characterization (Extended Data Fig. 9).

Metagenomic and amplicon sequencing data

DNA extraction, sequencing (Illumina with RTA v.1.18.54 and bcl2fastq v.1.8.4), trimming and mapping approaches are detailed in ref. 2. In this study, four community aggregated trait databases were built, corresponding to metagenomic reads mapping on different functional annotation systems in ref. 2. An additional database was made for this study with genomic traits previously associated with bacterial life history strategies (see details below). Data from 16S rRNA gene amplicon sequencing were also used to characterize phylogenetic distances between bacterial communities using the Unifrac metric[50].

Bacterial community aggregated trait calculation

Ref. [2] mapped reads to the functional databases KEGG, eggNOG and CAZy. Data were aggregated at the (1) pathway (KEGG), (2) functional categories (eggNOG) levels, (3) SEED functional modules and (4) glycolysis hydrolases (GH) and auxiliary activities (AA) gene families from CAZy[51]. All read mapping was done competitively against both prokaryotic and eukaryotic functional databases and best bit score in the alignment, and the taxonomic annotation was used to retrieve only reads annotated as bacteria.

We used output data from these four annotation processes to provide complementary classification of functional genes (for example, eggNOG categories include motility, cell envelopes and defence which are not included in SEED, whereas SEED classes include dormancy and sporulation, stress response, virulence, carbon, nitrogen and phosphorus metabolism which are not included in eggNOG). The eggNOG annotation also differed from KEGG and SEED in the construction of orthologous groups, with eggNOG using non-supervised construction increasing coverage, whereas KEGG used supervised construction increasing annotation robustness. To obtain a more precise picture of C acquisition strategy, the CAZy annotated reads abundances were aggregated on the basis of their targeted substrates (cellulose, chitin, glucan, lignin, peptidoglycan, starch/glycogen, xylan, other animal polysaccharides, other plant polysaccharides, oligosaccharides) using a curated database (Supplementary Table 1) based on previous works[52–54]. After mapping, the relative abundance of each gene (or aggregated group of genes) was normalized by the total number of bacteria reads annotated for this sample on the same database. Such normalization corrects for variation between samples in the quantity of annotated reads and avoids biases induced by contamination and sequencing error[55]. The obtained relative abundances inform on the relative importance of a gene (or gene group) compared to all the other annotated functions.

Life history trait calculation

An additional database was built with genomic traits previously associated with bacteria life history strategies (Extended Data Table 1). For this database, nine life history traits were calculated. Seven traits were calculated by summing the relative abundances of genes associated with Sigma factor[56], exopolysaccharides[57], chaperons[12,58], chemotaxis and osmolytes[12,59–62] antibiotic resistance[2] or carbohydrates degradation enzymes (CAZyme). In addition, average genome size was calculated using MicrobeCensus[63] and rrn copy number using the method described in ref. 64. All sequences were used as input for average genome size and rrn copy number, after verifying that eukaryotic sequences were negligible (less than 2% of annotated reads for all databases verified for all samples) and therefore, that the samples mostly captured bacteria.

Statistical analysis

To identify the multivariate axes that best explain the global-scale variation in metagenomic community aggregated traits of soil bacteria, we used a MCOA, an exploratory analysis that leverages together the information from the 5 databases (Life history traits, eggNOG categories, SEED modules, KEGG pathway, CAZy types). This method identifies co-relationships between the different databases and uses a covariance optimization criterion to summarize in a common structure the information shared by multiple multivariate (for example, omic) tables[65–67]. All variables (CATs) were log transformed ($\log X+1$) before the analysis to improve normality[67], and standardized to a mean of zero and a variance of 1. The R package ade4 was used for the MCOA analysis[68].

Sample coordinates on the first and second dimension of the MCOA were extracted and used as latent variables representing bacterial community positions in the global trait space. Random forest models were then used to identify predictors of these coordinates among potential environmental drivers, which were the soil properties measured on the same sample used for metagenome analysis (see Soil sampling and characteristics) and climatic variables extracted from Worldclim2: BIO1, annual mean temperature; BIO4, temperature seasonality (standard deviation); BIO12, annual precipitation; and BIO15, precipitation seasonality (standard deviation). First, we verified that all selected environmental drivers had spearman correlation coefficients lower than 0.7 to mitigate collinearity problems as recommended in ref. 69. Second, a variable selection process was carried out using the method implemented in the VSURF R package[70]. The number of predictors randomly tested at each node of the random forest tree (mtry) was optimized on the basis of randomForest's tuneRF algorithm and the number of trees set to 1,000. Third, the random forest models selected following the VSURF selection process were trained using 10-fold cross-validation (100 repetitions) implemented in the caret package[71], and model performance was assessed on the basis of root mean square error and R^2. Finally, random forest predictive models were used to project a broad resolution map of trait dimension global biogeography, using environmental maps (1,600 × 1,200 pixels) as predictors. For this projection, we used the latest map (June 2022) released by ISRIC's World Soil Information Service (https://files.isric.org/soilgrids/latest/data_aggregated/) based on SoilGrids v.2.0 (ref. 72). Worldclim2 (https://www.worldclim.org/) was used for climatic variables. The raster R package was used for the spatial predication and projection. To validate the relevance of this broad resolution map in representing average local values, we tested the correlation between local observations and the predicted value of the cell in which the local observation was done.

Finally, we tested the relationship between phylogenetic composition of the bacterial communities and their positions in the MCOA trait space using linear correlation between Euclidean distances along the first two dimensions of the MCOA and Unifrac phylogenetic distance. The influence of average genome size on this relationship was then assessed by comparing the correlation coefficients for communities below and above the median average genome size in the dataset.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The five CAT databases used to build the trait dimensions and the associated environmental variables are available on the Figshare repository at https://doi.org/10.6084/m9.figshare.22620025. All the original sequences are available in the European Bioinformatics Institute Sequence Read Archive database: soil metagenomes, accession numbers PRJEB18701 (ERP020652); 16S metabarcoding sequences, accession numbers PRJEB19856 (ERP021922).

Code availability

Access to the code used in the analyses done for this research is available by request to the corresponding author.

References

1. Fierer, N. Embracing the unknown: disentangling the complexities of the soil microbiome. *Nat. Rev. Microbiol.* **15**, 579–590 (2017).
2. Bahram, M. et al. Structure and function of the global topsoil microbiome. *Nature* **560**, 233–237 (2018).
3. Delgado-Baquerizo, M. et al. A global atlas of the dominant bacteria found in soil. *Science* **359**, 320–325 (2018).
4. Crowther, T. W. et al. The global soil community and its influence on biogeochemistry. *Science* **365**, eaav0550 (2019).
5. Wieder, W. R., Bonan, G. B. & Allison, S. D. Global soil carbon projections are improved by modelling microbial processes. *Nat. Clim. Change* **3**, 909–912 (2013).
6. Diaz, S. et al. The global spectrum of plant form and function. *Nature* **529**, 167–171 (2016).
7. Grime, J. P. Evidence for the existence of three primary strategies in plants and its relevance to ecological and evolutionary theory. *Am. Nat.* **111**, 1169–1194 (1977).
8. Wright, I. J. et al. The worldwide leaf economics spectrum. *Nature* **428**, 821–827 (2004).
9. Southwood, T. R. Habitat, the templet for ecological strategies? *J. Anim. Ecol.* **46**, 337–365 (1977).
10. Reich, P. B. et al. The evolution of plant functional variation: traits, spectra, and strategies. *Int. J. Plant Sci.* **164**, S143–S164 (2003).
11. Krause, S. et al. Trait-based approaches for understanding microbial biodiversity and ecosystem functioning. *Front. Microbiol.* **5**, 251 (2014).
12. Malik, A. A. et al. Defining trait-based microbial strategies with consequences for soil carbon cycling under climate change. *ISME J.* https://doi.org/10.1038/s41396-019-0510-0 (2019).
13. Madin, J. S. et al. A synthesis of bacterial and archaeal phenotypic trait data. *Sci. Data* **7**, 170 (2020).
14. Westoby, M. et al. Trait dimensions in bacteria and archaea compared to vascular plants. *Ecol. Lett.* **24**, 1487–1504 (2021).
15. Steen, A. D. et al. High proportions of bacteria and archaea across most biomes remain uncultured. *ISME J.* **13**, 3126–3130 (2019).
16. Martiny, A. C. High proportions of bacteria are culturable across major biomes. *ISME J.* **13**, 2125–2128 (2019).
17. Martiny, A. C. The '1% culturability paradigm' needs to be carefully defined. *ISME J.* **14**, 10–11 (2020).
18. Fierer, N., Barberán, A. & Laughlin, D. C. Seeing the forest for the genes: using metagenomics to infer the aggregated traits of microbial communities. *Front. Microbiol.* **5**, 614 (2014).
19. Garnier, E. et al. Plant functional markers capture ecosystem properties during secondary succession. *Ecology* **85**, 2630–2637 (2004).
20. Violle, C. et al. Let the concept of trait be functional! *Oikos* **116**, 882–892 (2007).
21. Fierer, N. et al. Comparative metagenomic, phylogenetic and physiological analyses of soil microbial communities across nitrogen gradients. *ISME J.* **6**, 1007–1017 (2012).
22. Sorensen, J. W., Dunivin, T. K., Tobin, T. C. & Shade, A. Ecological selection for small microbial genomes along a temperate-to-thermal soil gradient. *Nat. Microbiol.* **4**, 55–61 (2019).
23. Grime, J. P. & Pierce, S. *The Evolutionary Strategies That Shape Ecosystems* (John Wiley & Sons, 2012).
24. Liu, H. et al. Warmer and drier ecosystems select for smaller bacterial genomes in global soils. *iMeta* https://doi.org/10.1002/imt2.70 (2023).
25. Simonsen, A. K. Environmental stress leads to genome streamlining in a widely distributed species of soil bacteria. *ISME J.* **16**, 423–434 (2021).
26. Chuckran, P. F. et al. Edaphic controls on genome size and GC content of bacteria in soil microbial communities. *Soil Biol. Biochem.* **178**, 108935 (2023).
27. Guieysse, B. & Wuertz, S. Metabolically versatile large-genome prokaryotes. *Curr. Opin. Biotechnol.* **23**, 467–473 (2012).
28. Konstantinidis, K. T. & Tiedje, J. M. Trends between gene content and genome size in prokaryotic species with larger genomes. *Proc. Natl Acad. Sci. USA* **101**, 3160–3165 (2004).
29. Paul, C. et al. Bacterial spores, from ecology to biotechnology. *Adv. Appl. Microbiol.* **106**, 79–111 (2019).
30. Singh, S., Datta, S., Narayanan, K. B. & Rajnish, K. N. Bacterial exo-polysaccharides in biofilms: role in antimicrobial resistance and treatments. *J. Genet. Eng. Biotechnol.* **19**, 140 (2021).
31. Sinsabaugh, R. L. & Follstad Shah, J. J. Ecoenzymatic stoichiometry and ecological theory. *Annu. Rev. Ecol. Evol. Syst.* **43**, 313–343 (2012).
32. Buckeridge, K. M. et al. Environmental and microbial controls on microbial necromass recycling, an important precursor for soil carbon stabilization. *Commun. Earth Environ.* **1**, 36 (2020).
33. Zheng, Q. et al. Growth explains microbial carbon use efficiency across soils differing in land use and geology. *Soil Biol. Biochem.* **128**, 45–55 (2019).
34. Gao, Y. & Wu, M. Free-living bacterial communities are mostly dominated by oligotrophs. Preprint at *bioRxiv* https://doi.org/10.1101/350348 (2018).
35. Li, J. et al. Predictive genomic traits for bacterial growth in culture versus actual growth in soil. *ISME J.* **13**, 2162–2172 (2019).
36. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833–844 (2017).
37. Lobb, B., Tremblay, B. J.-M., Moreno-Hagelsieb, G. & Doxey, A. C. An assessment of genome annotation coverage across the bacterial tree of life. *Microb. Genom.* **6**, e000341 (2020).
38. Coelho, L. P. et al. Towards the biogeography of prokaryotic genes. *Nature* **601**, 252–256 (2022).
39. Martiny, J. B., Jones, S. E., Lennon, J. T. & Martiny, A. C. Microbiomes in light of traits: a phylogenetic perspective. *Science* **350**, aac9323 (2015).
40. Allison, S. D. & Martiny, J. B. Resistance, resilience, and redundancy in microbial communities. *Proc. Natl Acad. Sci. USA* **105**, 11512–11519 (2008).
41. Jones, D. L., Cooledge, E. C., Hoyle, F. C., Griffiths, R. I. & Murphy, D. V. pH and exchangeable aluminum are major regulators of microbial energy flow and carbon use efficiency in soil microbial communities. *Soil Biol. Biochem.* **138**, 107584 (2019).
42. Fernández-Calviño, D. & Bååth, E. Growth response of the bacterial community to pH in soils differing in pH. *FEMS Microbiol. Ecol.* **73**, 149–156 (2010).
43. Auger, C. et al. Metabolic reengineering invoked by microbial systems to decontaminate aluminum: implications for bioremediation technologies. *Biotechnol. Adv.* **31**, 266–273 (2013).
44. Bruelheide, H. et al. Global trait–environment relationships of plant communities. *Nat. Ecol. Evol.* **2**, 1906–1917 (2018).
45. Tedersoo, L. et al. Regional-scale in-depth analysis of soil fungal diversity reveals strong pH and plant species effects in Northern Europe. *Front. Microbiol.* **11**, 1953 (2020).
46. Bagousse-Pinguet, Y. L. et al. Testing the environmental filtering concept in global drylands. *J. Ecol.* **105**, 1058–1069 (2017).
47. Lauber, C. L., Hamady, M., Knight, R. & Fierer, N. Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. *Appl. Environ. Microbiol.* **75**, 5111–5120 (2009).
48. Meyer, F. et al. The metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9**, 386 (2008).
49. Chen, I.-M. A. et al. IMG/M v. 5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Res.* **47**, D666–D677 (2019).

50. Lozupone, C. & Knight, R. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* **71**, 8228–8235 (2005).
51. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* **42**, D490–D495 (2014).
52. Nguyen, L. T. et al. Responses of the soil microbial community to nitrogen fertilizer regimes and historical exposure to extreme weather events: flooding or prolonged-drought. *Soil Biol. Biochem.* **118**, 227–236 (2018).
53. Berlemont, R. & Martiny, A. C. Genomic potential for polysaccharide deconstruction in bacteria. *Appl. Environ. Microbiol.* **81**, 1513–1519 (2015).
54. López-Mondéjar, R. et al. Metagenomics and stable isotope probing reveal the complementary contribution of fungal and bacterial communities in the recycling of dead biomass in forest soil. *Soil Biol. Biochem.* **148**, 107875 (2020).
55. Nayfach, S. & Pollard, K. S. Toward accurate and quantitative comparative metagenomics. *Cell* **166**, 1103–1116 (2016).
56. Chávez, J., Devos, D. P. & Merino, E. Complementary tendencies in the use of regulatory elements (transcription factors, sigma factors, and riboswitches) in bacteria and archaea. *J. Bacteriol.* **203**, 413–20 (2020).
57. Cania, B. et al. Site-specific conditions change the response of bacterial producers of soil structure-stabilizing agents such as exopolysaccharides and lipopolysaccharides to tillage intensity. *Front. Microbiol.* **11**, 568 (2020).
58. Finn, D., Yu, J. & Penton, C. R. Soil quality shapes the composition of microbial community stress response and core cell metabolism functional genes. *Appl. Soil Ecol.* **148**, 103483 (2020).
59. Sharma, M. P. et al. Deciphering the role of trehalose in tripartite symbiosis among rhizobia, arbuscular mycorrhizal fungi, and legumes for enhancing abiotic stress tolerance in crop plants. *Front. Microbiol* **11**, 509919 (2020).
60. Yaakop, A. S. et al. Characterization of the mechanism of prolonged adaptation to osmotic stress of *Jeotgalibacillus malaysiensis* via genome and transcriptome sequencing analyses. *Sci. Rep.* **6**, 33660 (2016).
61. Wargo, M. J. Homeostasis and catabolism of choline and glycine betaine: lessons from *Pseudomonas aeruginosa*. *Appl. Environ. Microbiol.* **79**, 2112–2120 (2013).
62. Boch, J., Kempf, B., Schmid, R. & Bremer, E. Synthesis of the osmoprotectant glycine betaine in *Bacillus subtilis*: characterization of the gbsAB genes. *J. Bacteriol.* **178**, 5121–5129 (1996).
63. Nayfach, S. & Pollard, K. S. Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. *Genome Biol.* **16**, 51 (2015).
64. Pereira-Flores, E., Glöckner, F. O. & Fernandez-Guerra, A. Fast and accurate average genome size and 16S rRNA gene average copy number computation in metagenomic data. *BMC Bioinformatics* **20**, 453 (2019).
65. Chessel, D. & Hanafi, M. Analyses de la co-inertie de *K* nuages de points. *Rev. Stat. Appl.* **44**, 35–60 (1996).
66. Piton, G. et al. Using proxies of microbial community-weighted means traits to explain the cascading effect of management intensity, soil and plant traits on ecosystem resilience in mountain grasslands. *J. Ecol.* **108**, 876–893 (2020).
67. Meng, C., Kuster, B., Culhane, A. C. & Gholami, A. M. A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics* **15**, 162 (2014).
68. Dray, S., Dufour, A. B. & Chessel, D. The ade4 package-II: two-table and K-table methods. *R News* **7**, 47–52 (2007).
69. Dormann, C. F. et al. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* **36**, 27–46 (2013).
70. Genuer, R., Poggi, J.-M. & Tuleau-Malot, C. VSURF: an R package for variable selection using random forests. *R J.* **7**, 19–33 (2015).
71. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw.* https://doi.org/10.18637/jss.v028.i05 (2008).
72. Poggio, L. et al. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. *Soil* **7**, 217–240 (2021).

Acknowledgements

We thank L. Tedersoo and P. Bork who conceived and supervised the acquisition of the global dataset used in this study with M. Bahram and F. Hildebrand; all their collaborators who contributed to this global data acquisition effort; A. Larkin and L. Ustick for guidance in the bioinformatic analysis conducted in this study. G.P., S.D.A., J.B.H.M., K.K.T. and A.C.M. were supported by the US Department of Energy, Office of Science, Office of Biological and Environmental Research grants DE-SC0016410 and DE-SC0020382. F.H. was supported by the European Research Council H2020 StG (erc-stg-948219, EPYC) and by the Biotechnology and Biological Sciences research Council (BBSrC) Institute Strategic Program (ISP) Food Microbiome and Health BB/X011054/1 and its constituent project BBS/E/F/000PR13631; Earlham DECODE ISP BBX011089/1 and its constituent work package BBS/E/ER/230002A.

Author contributions

Data collection was designed and supervised by M.B. Initial bioinformatics analysis to obtain functional genes abundance tables (eggNOG, KEGG, SEED, CAZy) was designed and performed by F.H. The idea of this new analysis was conceived by G.P. with inputs from A.C.M., S.D.A, J.B.H.M. and K.K.T. New quantification of genomic traits, Unifrac and data analyses were performed by G.P. Writing of the first draft and subsequent editing was performed by G.P. with inputs from all co-authors.

Competing interests

The authors declare no competing interests.

Additional information

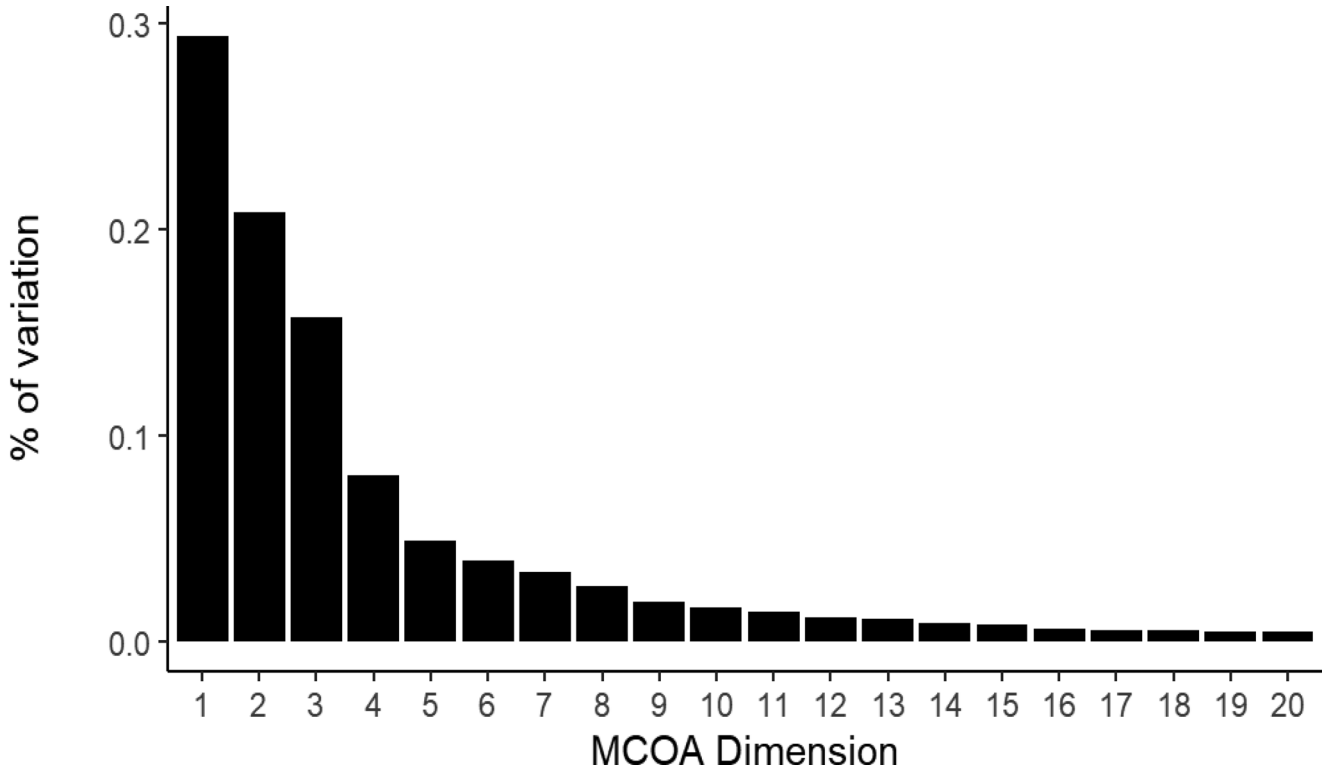
Extended data is available for this paper at https://doi.org/10.1038/s41564-023-01465-0.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41564-023-01465-0.

Correspondence and requests for materials should be addressed to Gabin Piton.

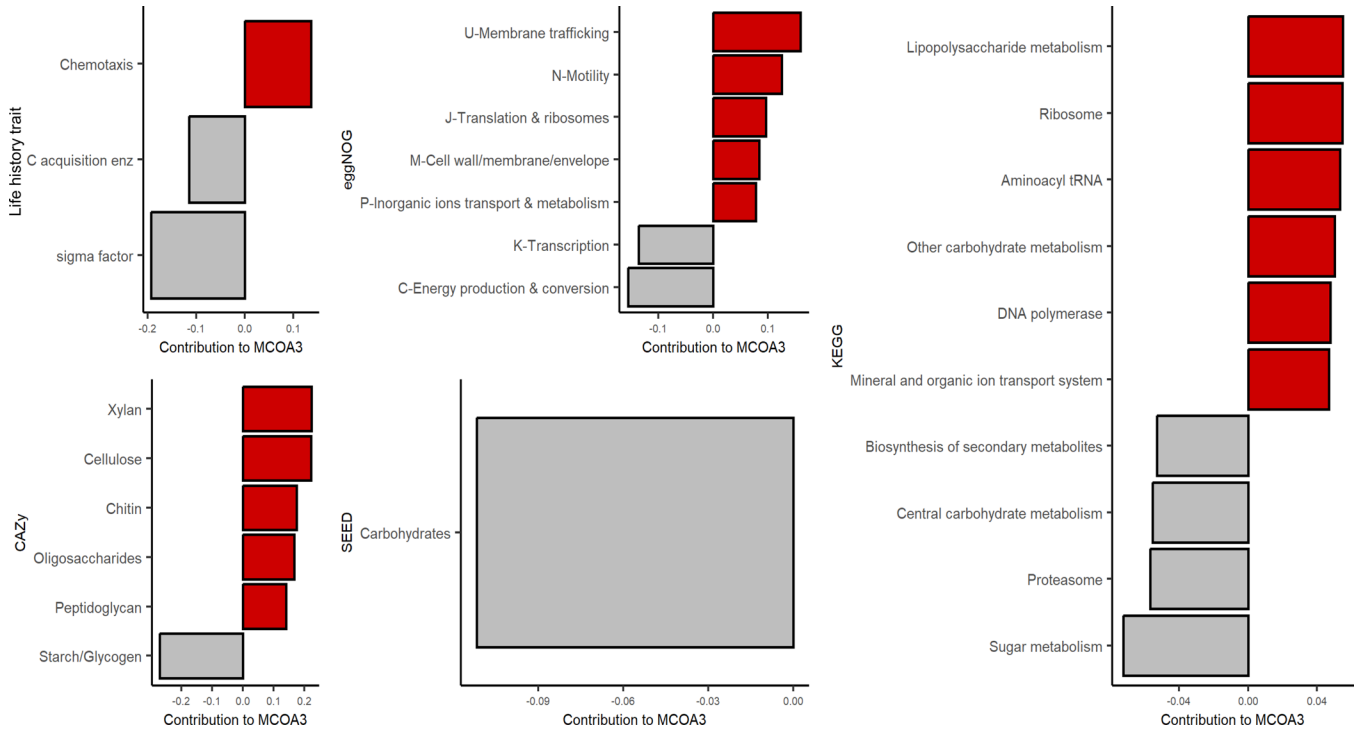
Peer review information *Nature Microbiology* thanks Tess Brewer, Kate Buckeridge and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s41564-023-01465-0



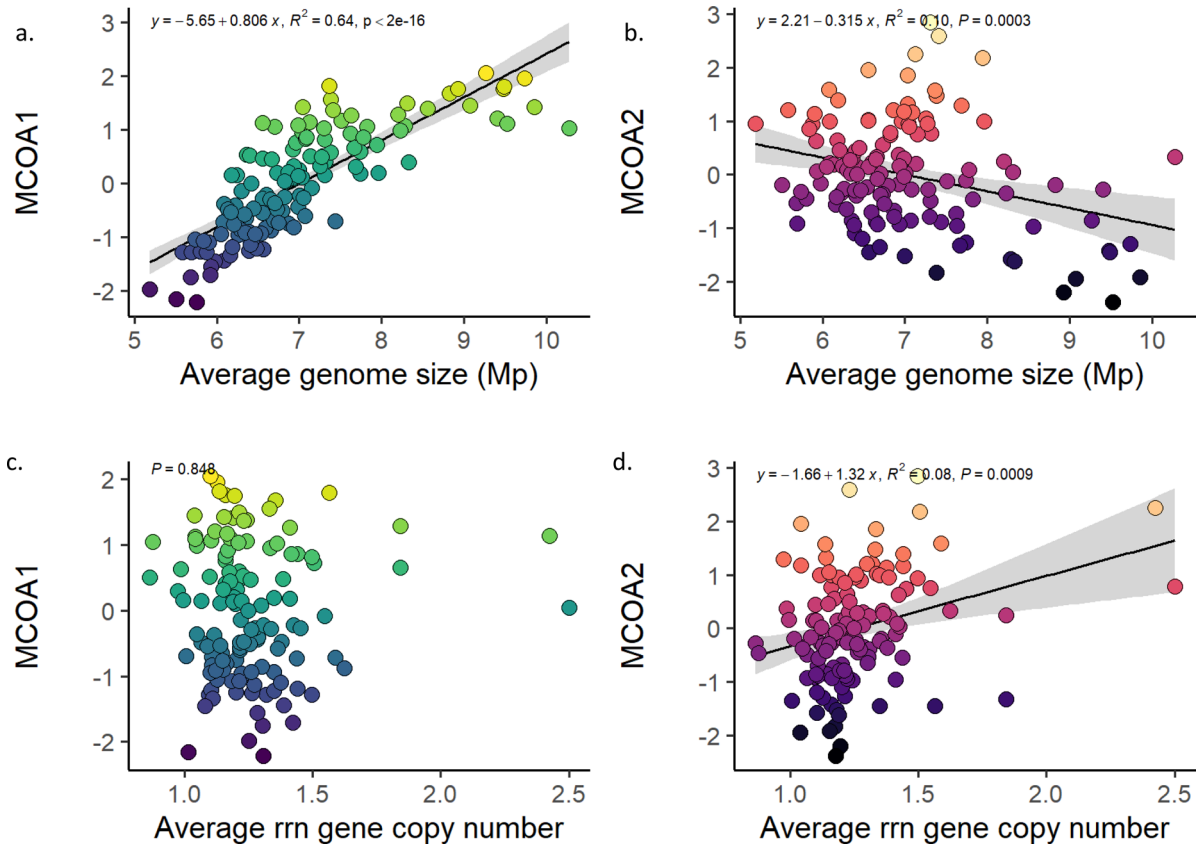
Extended Data Fig. 1 | Stress plot representing the % of variation of the global dataset captured by each dimension of the MCOA.

https://doi.org/10.1038/s41564-023-01465-0



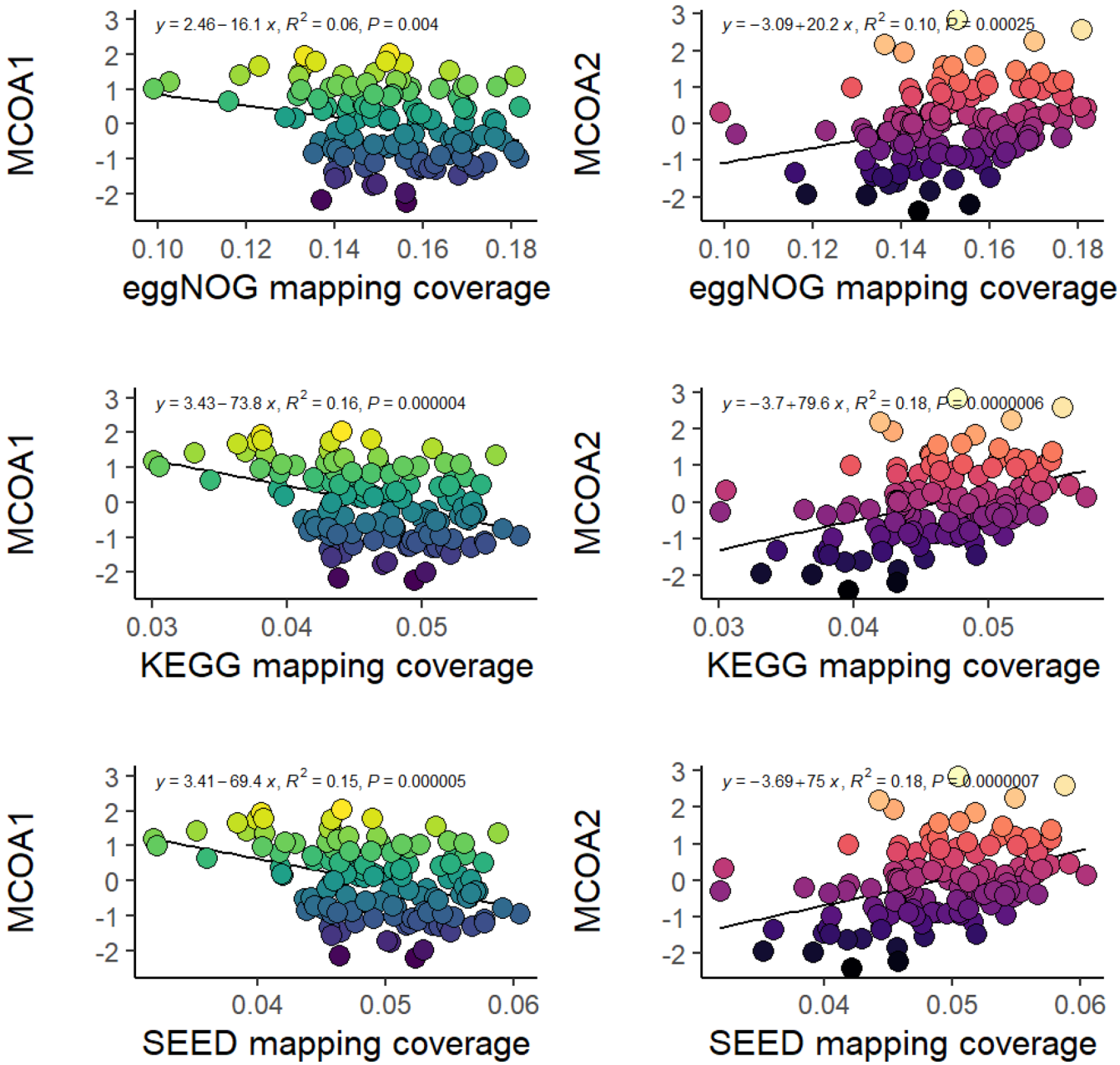
Extended Data Fig. 2 | Variable contributions to the third trait dimension of the multiple co-inertia analysis (MCOA). The MCOA summarizes in a common structure the information shared by 5 community aggregated trait (CAT) databases (Genomic trait, CAZy, eggNOG, SEED and KEGG). Only the most important variables with significant correlation ($p < 0.001$) with each dimension are reported in this figure.

https://doi.org/10.1038/s41564-023-01465-0



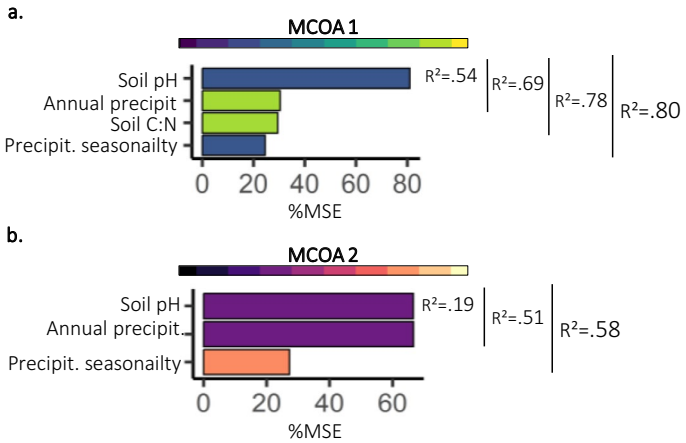
Extended Data Fig. 3 | Correlations between average genome size (a,b) or average rrn gene copy number (c,d) and the coordinates along dimensions 1 and 2 of the MCOA. The P value indicates the significance of the regression slope obtained using a t-test. Shade represents the estimated 95% confidence interval. Color gradients follow MCOA dimensions and match with Figs. 1 and 3 in the main text.

https://doi.org/10.1038/s41564-023-01465-0



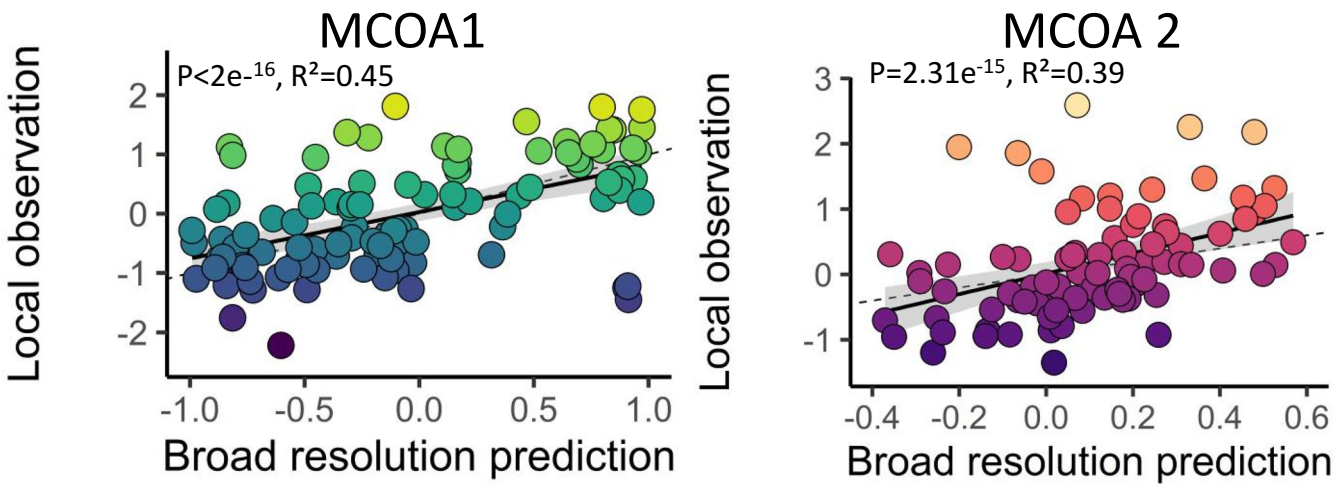
Extended Data Fig. 4 | Correlations between MCOA dimensions (MCOA1 and MCOA2) and mapping coverages on the 3 general databases (eggNOG, KEGG, SEED) used in this study. The P value indicates the significance of the regression slope obtained using a t-test.

https://doi.org/10.1038/s41564-023-01465-0



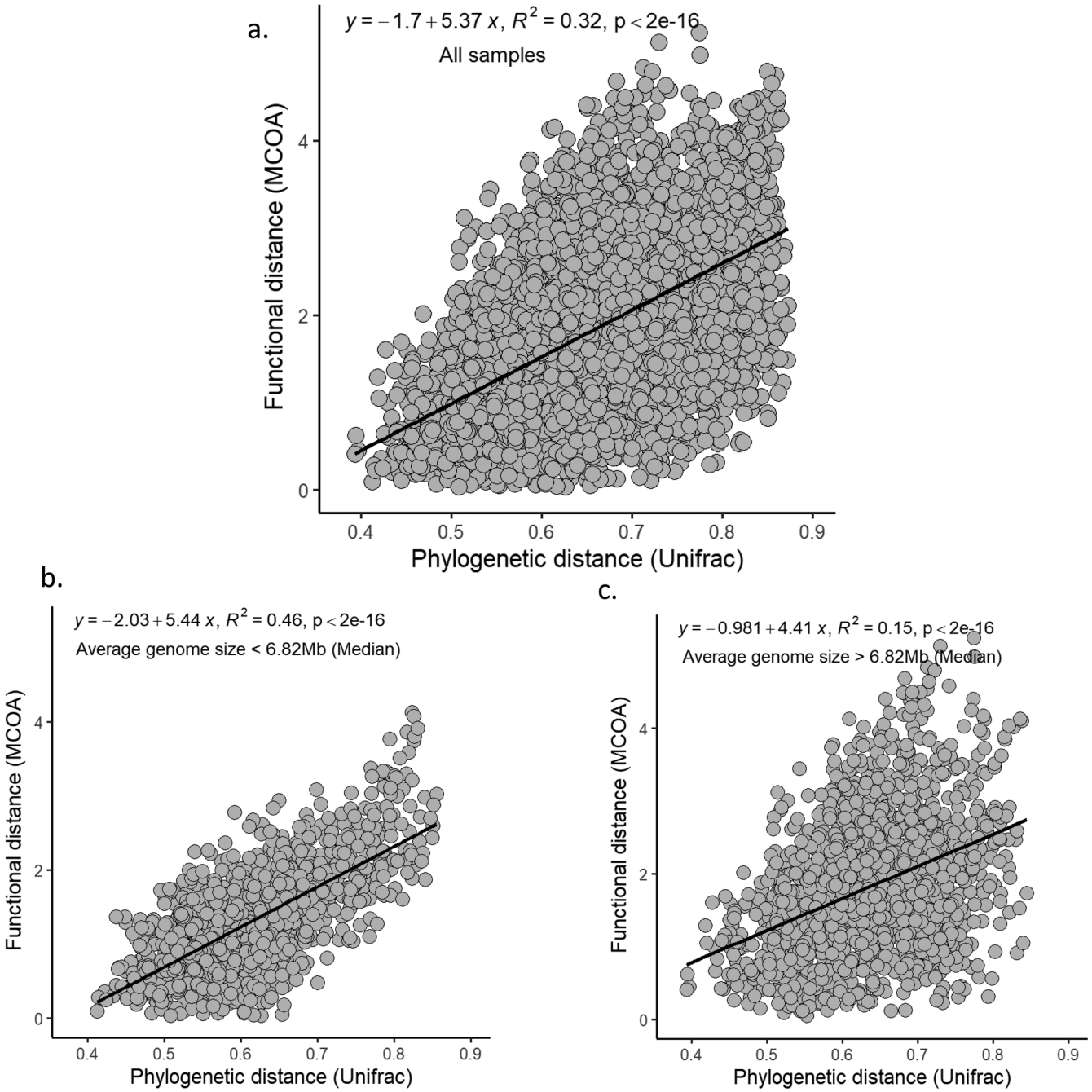
Extended Data Fig. 5 | Environmental drivers of the bacterial community trait dimensions. Environmental variable importances are represented as the mean decrease in mean square error (%MSE) and R squared in random forest models predicting MCOA Dimension 1 (**a**) and 2 (**b**). Bar colours indicate which end of the dimension (Figs. 1 and 3) is positively correlated with the variable.

https://doi.org/10.1038/s41564-023-01465-0



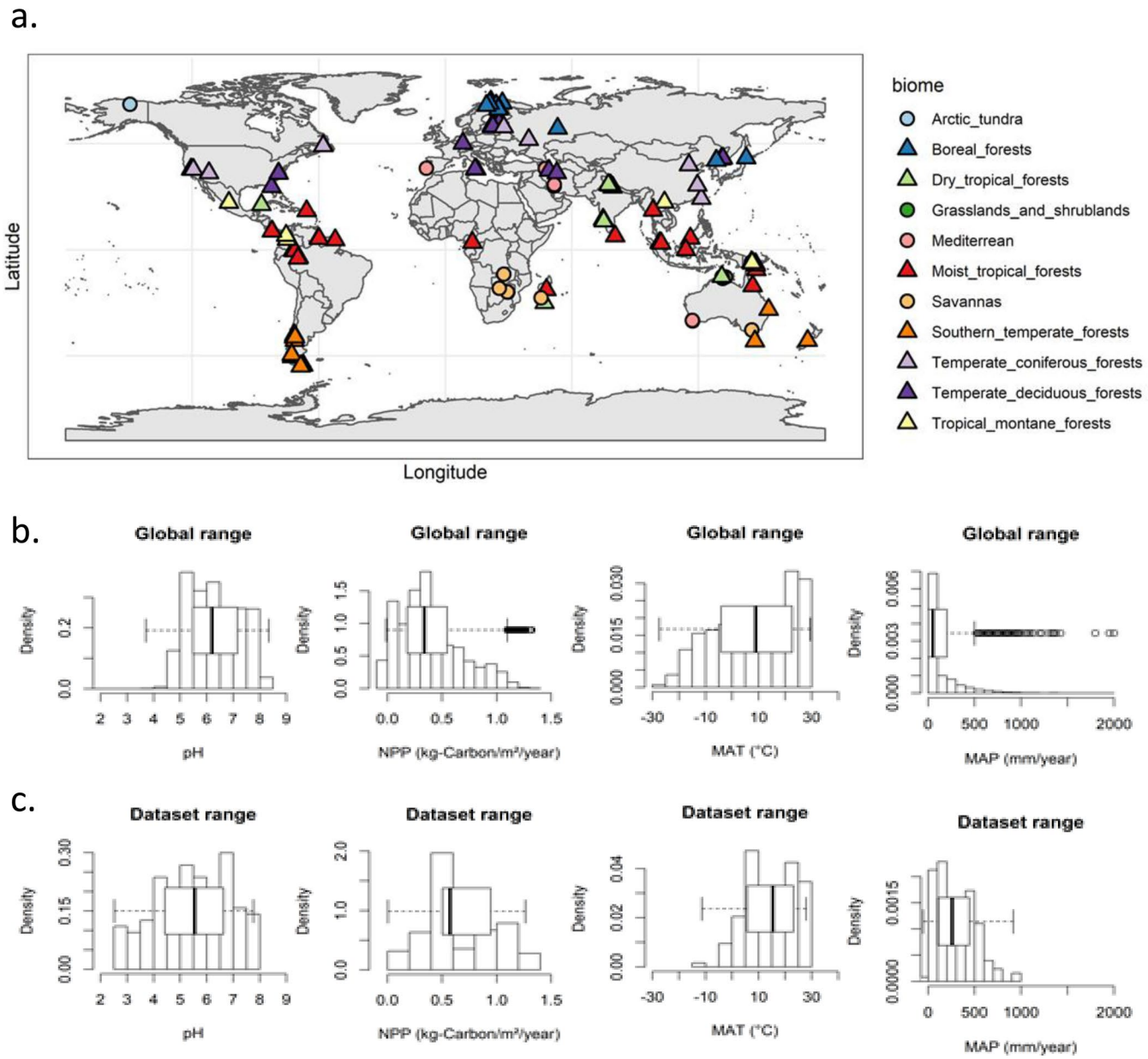
Extended Data Fig. 6 | Correlations between local trait dimension observations and global spatial prediction. Correlations between local observations of bacterial community positions along the first and second trait dimensions from the MCOA (Figs. 1–2) and the predicted value of the global map cell (Fig. 4) corresponding to where the local observations have been done. Dashed line represents a 1:1 correlation. The P value indicates the significance of the regression slope obtained using a t-test. Shade represents the estimated 95% confidence interval. Color gradients follow MCOA dimension and match with Figs. 1,2 and 4 in the main text.

https://doi.org/10.1038/s41564-023-01465-0



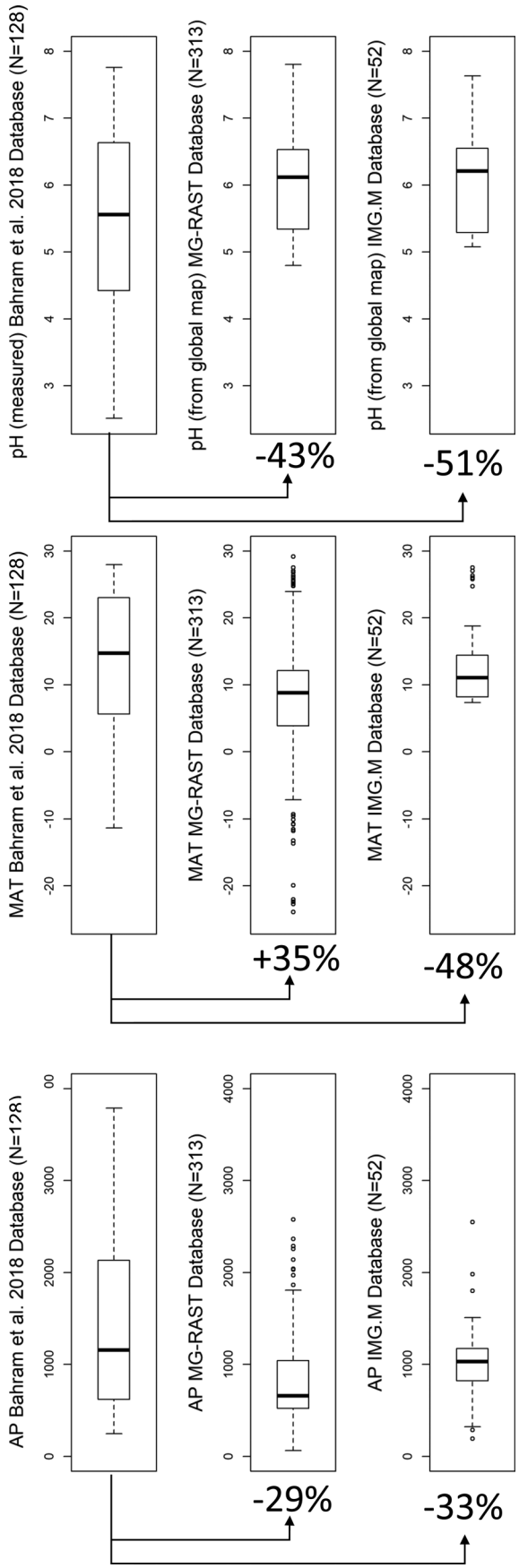
Extended Data Fig. 7 | Correlation between phylogenetic distance (Unifrac metric) and functional distance (Euclidian distance in MCOA space using coordinates of the two principal dimensions). Correlation for all samples (**a**) and restricted to samples with average genome size below (**b**) and above (**c**) its median value in the dataset. The P value indicates the significance of the regression slope obtained using a t-test.

https://doi.org/10.1038/s41564-023-01465-0



Extended Data Fig. 8 | Dataset distribution and environmental coverage. **a**. Sample localisations and associated biomes **b**, **c**. Comparison between global range of environmental variables from the Atlas of the Biosphere (b) and the environmental coverage of dataset (n = 128) used in this study (c). Boxplot elements: Center line=median; box limits=upper and lower quartiles; whiskers = 1.5x interquartile range; points=outliers. World map was done with rnaturalearth R package (https://github.com/ropensci/rnaturalearth).

https://doi.org/10.1038/s41564-023-01465-0



Extended Data Fig. 9 | See next page for caption.

https://doi.org/10.1038/s41564-023-01465-0

Extended Data Fig. 9 | Environmental coverage comparison between the database used in this study from Barham et al.[2] and databases from the main metagenomes repositories (MG-RAST and IMG:M). N corresponds to the number of metagenomes available in each database. MAT=Mean Annual Temperature, AP=Annual Precipitation. Boxplot elements: Center line=median; box limits=upper and lower quartiles; whiskers = 1.5× interquartile range; points=outliers.

https://doi.org/10.1038/s41564-023-01465-0

Extended Data Table 1 | Life history traits used in this study

Life history traits	Associated metagenomic community aggregated traits used in this study	CSR [1]	CSR [2]	YAS [3]
Amino acid, fatty acid and nucleotide synthesis [3]	> eggNOG category, KEGG pathway and SEED modules associated with amino acid, lipid and nucleotide metabolism			Y
Chaperons [3]	> Chaperons genes : GroEL (COG0459), dnaK (COG0443) and dnaJ (COG0484) (Malik et al. 2020, Finn et al. 2020)			S
Siderophores [1,3]	> KEGG pathway "Metallic cation iron siderophore and vitamin B12 transport system "	C		A
Oligosaccharides degradation enzymes	> Genes associated with Oligosaccharides degradation among other GH and AA genes			
Carbohydrate central metabolism [3]	> KEGG pathway "Central_carbohydrate_metabolism"			Y
Primary metabolism	> eggNOG categories : F-Nucleotide transport and metabolism, J-Translation and ribosomes, D-Cell cycling, E-Amino acid transport and metabolism, H-Coenzyme transport and metabolism and A-RNA processing and modification, SEED modules : DNA and protein metabolisms. KEGG pathways: Purine, Cysteine, Methionine, Arginine, Proline and Lysine metabolism, Proteosome, cofactors and vitamins metabolisms and Ribosome, Aminoacyl tRNA, RNA and DNA polymerase and Nucleotide sugars			
Genome size [1,2]	> Average genome size (Nayfach and Pollard 2015)	C	R	
Complex polymers degradation enzymes [3]	> Genes associated with Lignin degradation among other GH and AA genes			A
Fungal biomass degradation enzymes	> Genes associated with Chitin and Glucan degradation among other GH and AA genes			
Antibiotic [1,2]	> Antibiotic Resistance Genes	C	C	
Pathogenic interactions with plants	> SEED module : Virulence			
Sporulation [1,2]	> SEED module "Dormancy_and_Sporulation"	R	S	
EPS [1,2,3]	> EPS genes : WcaB (COG1596), WcaF (COG0110), Wza (COG1596), KpsE and RkpR(COG3524) and wcaK(COG2327) (Cania et al. 2020)	S	S	S
Membrane synthesis and repair [3]	> eggNOG categories : L-Replication, recombination & repairs, M-Cell wall, membrane and envelope, KEGG pathways : Lipid and lipopolysaccharide metabolism			S
rRNA gene copies [1,2]	> Average rRNA copy number (Pereira-Flores et al. 2019)	R	C	
Motility [2,3]	> eggNOG category : "N-Motility"		R	A
Chemotaxis [2,3]	> Genes associated with chemotaxis : CheA (COG0643), CheY (COG0784), CheW (COG0835), CheB (COG2201), CheX (COG1406), CheD(COG1871), Methyl-accepting chemotaxis proteins (COG0840, COG1352)		R	A
Sigma factor [3]	> σ factor genes : σD, σS and σH (COG0568), σF and σB (COG1191), σN (COG1508) and extracytoplasmic function σ factors (COG1595) (Chávez et al. 2020)			S
Osmolytes [3]	> Genes associated with Trehalose and glycine betaine (Malik et al. 2020, Sharma et al. 2020, Suriaty Yaakop et al. 2016, Bochet al. 1996, Wargo et al. 2013)			S
Exoenzymes (All) [2,3]	> GH and AA genes in global metabolism		S	A
Bacterial biomass degradation enzyme	> Genes associated with Peptidoglycan degradation			
Uptake system [2,3]	> KEGG pathway and SEED modules associated with transport systems		S	A

Traits were selected on the basis of their previous association with CSR ('Competitor', 'Stress tolerant', and 'Ruderal') strategies by Fierer. [1] [1] or Krause et al.[11] [2] or YAS strategies ('Yield','Resource acquisition', and 'Stress tolerant') by Malik et al. [3][12]. Cells associated with CSR and YAS have been greyed based on the strategy to facilitate comparisons between references. Same gray has been used for C and A, and for R and Y strategies as they have some important theoretical linkages[12].

nature portfolio

Corresponding author(s): Gabin Piton

Last updated by author(s): Jul 20, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Original data were collected by Bahram et al. (2018) and provided by the authors. For DNA sequencing, they used illumina technology with RTA Version 1.18.54 and bcl2fastq v1.8.4. Bahram, M., F. Hildebrand, S. K. Forslund, J. L. Anderson, N. A. Soudzilovskaia, P. M. Bodegom, J. Bengtsson-Palme, S. Anslan, L. P. Coelho, H. Harend, and others. 2018. Structure and function of the global topsoil microbiome. Nature 560:233–237.
Data analysis	R version 4.1.2 was used for statistical analyzes with packages VSURF v1.2.0., raster v3.6-11, caret v6.0-93 and rnaturalearth v0.3.2, ade4 v1.7-20

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All metagenomics sequences are available in the European Bioinformatics Institute-Sequence Read Archive database under accession numbers PRJEB18701 (ERP020652). The final dataset that support the finding of this study and the original annotation table are deposited on figshare : https://doi.org/10.6084/m9.figshare.22620025

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	n/a
Reporting on race, ethnicity, or other socially relevant groupings	n/a
Population characteristics	n/a
Recruitment	n/a
Ethics oversight	n/a

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This study uses metagenomic data to characterize the trait dimensions of soil bacteria communities based on relationships between their metagenomic attributes (metagenomic profile from eggNOG, SEED, KEGG and CAZy database) and a set of genomic traits associated with life history strategy (eg. genome size, RNA gene copy number). This study next assesses environmental drivers of this trait dimensions based on soil chemistry and climate condition associated with each sample and project its global geographical distribution.
Research sample	In this study we used the metagenomics and environmental data of 128 soils distributed across continents, previously collected by Bahram et al. (2018). We selected this dataset for our analysis because of its important geographical and environmental coverage (see Extended Data Figure 8 of our study) and its use of a highly standardized protocol: 1) to sample top-soils in spatially independent sites across the globe selected to represent all the most important vegetation types; 2) to analyze their chemistry and their metagenomes. Nevertheless, potential to extend environmental range by adding all natural (Agricultural and contaminated soil excluded) soil metagenomes available (accession date January 28 2021) from the main sequence repositories MG-RAST and IMG:M was also tested (See. Extended Data Figure 9 of our study). This indicated that adding these data would not have extended environmental range and would have greatly decreased precision of soil properties characterization. In the metagenomics analysis, we focus on sequences assigned to bacteria as this group dominates soil metagenomes. Bahram, M., F. Hildebrand, S. K. Forslund, J. L. Anderson, N. A. Soudzilovskaia, P. M. Bodegom, J. Bengtsson-Palme, S. Anslan, L. P. Coelho, H. Harend, and others. 2018. Structure and function of the global topsoil microbiome. Nature 560:233–237.
Sampling strategy	The 128 samples used in this study was selected as representative sites for different biomes and latitudes among 1450 sites worldwide, with these 128 samples covering most of the environmental range worldwide (see Supplementary Figure 8 of our study). These samples were also selected to minimize spatial autocorrelation (see Bahram et al. 2018 for details). Sample size of this study was determined by the number of samples from the global sampling of Bahram et al. 2018 for which both metagenomes and soil

	chemistry were available (n=128). This sample size is comparable with other global analyzes of soil microbial communities and was thus considered as sufficient.
Data collection	Data collection procedure are detailed in the initial study of Bahram et al. 2018. To sum up, initial metagenome sequencing and soil chemistry characterization of Bahram et al. 2018 were done at the Estonian Genomics Center and the Estonian University of Life Sciences respectively. Analyzes presented in this study were done by Gabin Piton at the University of California Irvine based on data shared by Bahram et al., and additional metagenomics output obtained by reprocessing sequences from European Bioinformatics Institute-Sequence Read Archive.
Timing and spatial scale	Data collection timing is detailed in the initial study of Bahram et al. 2018. Samples were collected between 2011-2016 depending on the availability of collaborators to the Bahram et al. 2018 study with no specific season preselected.
Data exclusions	All samples from Bahram et al. 2018 for which both metagenomes and soil chemistry were available were used, no sample were excluded.
Reproducibility	Cross validation were used where appropriate.
Randomization	No randomization was necessary as all analyzes were correlative and sampling was design to have spatially independant observations.
Blinding	This study rely on analyzes of environmental samples, thus blinding was not relevant for this study.

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging