Life history strategies of soil bacterial communities across global terrestrial biomes

Gabin Piton, Steven Allison, Mohammad Bahram, Falk Hildebrand, Jennifer Martiny, Kathleen Treseder, Adam Martiny

Life history strategies of soil bacterial communities across global terrestrial biomes

Gabin Piton[1,2,*], Steven D Allison[1,3], Mohammad Bahram[4,5], Falk Hildebrand[6,7], Jennifer BH Martiny[3], Kathleen K Treseder[3], Adam C Martiny[1,3]

Author information

Affiliations

1. Department of Earth System Science, University of California, Irvine, California, USA
2. Eco&Sols, INRAE-IRD-CIRAD-SupAgro, University Montpellier, Montpellier, France
3. Department of Ecology and Evolutionary Biology, University of California, Irvine, California, 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 & Health, Quadram Institute Bioscience, Norwich Research Park, Norwich, Norfolk NR4 7UA, UK
7. Digital Biology, Earlham Institute, Norwich Research Park, Norwich, Norfolk NR4 7UA, UK.

* Corresponding author: gabin.piton@inrae.fr

Abstract

The life history strategies of soil microbes determine their metabolic potential and their response to environmental changes. Yet they remain poorly understood. Here we use shotgun metagenomes from terrestrial biomes to characterise overarching covariations of the genomic traits that captures 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, bacteria communities increasingly differentiate along a second dimension that reflects a tradeoff between increasing capacities for environmental responsiveness or nutrient recycling. Random forest analyses shows that soil pH, C:N and precipitation patterns together drive the dominant life history strategy of soil bacteria communities and its biogeographic distribution. Our findings provide a trait-based framework to compare life history strategies of soil bacteria.

Introduction

Bacteria impact carbon (C) and nutrient cycling on a global scale[1]. Soil bacterial communities contain enormous, functionally uncharacterized genetic diversity[2,3] that hinders progress in predicting soil microbial responses to global change[4,5]. One approach to describe functional biodiversity is to collapse its complexity into one or more dimensions that capture the dominant associations and trade-offs between traits[6–10]. This multivariate trait space - o 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 is 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 emphasises 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. (2020)[12] emphasised 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 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, Westoby and co-workers (2021) analysed bacterial cultures isolated from diverse habitats 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 if the life history strategies of soil bacterial communities matches either the theoretical or culture-based predictions of key trait dimensions.

One advantage of studying the traits of microorganisms over those of larger organisms is the ease of which collections of their traits can be measured on 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 the within-organism tradeoffs. Nevertheless, CATs described using metagenomic sequences offer a way to characterize shifts in the organismal strategies dominating bacterial communities *in situ* (eg. [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 to environmental changes of this key group 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 the 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 multi-table co-inertia analysis (MCOA), we found that two dimensions captured half of the overall variation in metagenomic community aggregated traits (CATs). MCOA1 and MCOA 2 captured 29% and 21% of metagenomic trait variation (Figure 1 and Extended Data Figure 1), while MCOA 3 and MCOA4 explained 16% and 10% of this variation, respectively (Extended Data Figure 1 and 2). The MCOA revealed the most important associations between traits (Figure 1-2) including traits previously associated with life history strategies (Figure 1-3).

Average genome size had the highest contribution to MCOA1 (Figure 1A) with a R^2 of 0.64 for the positive correlation between average genome size and MCOA1 (Extended Data Figure 3A). Mapping coverage decreased along this dimension (Extended Data Figure 4). The lower end of this dimension was characterised by bacterial communities with higher relative abundance of genes for primary metabolism (ie. essential process for survival and growth) and C acquisition machinery (Figure 1). In these communities, carbon acquisition enzymes involved in depolymerization of oligosaccharides were favored 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 were defined by communities with a streamlined metabolism (Figure 2).

The upper end of MCOA1 defined bacterial communities with a large genome and more complex metabolism and resource acquisition strategies (Figure 1-2). The enriched genes allowed for degradation of complex polysaccharides from fungi, animals and plant lignin. There was also a gene overrepresentation for direct plant pathogenic interactions and negative interactions with other microorganisms. Finally, communities carried a higher proportion of genes encoding for EPS production, Dormancy and Sporulation, membrane, and DNA repair (Figure 1-2). 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 (Figure 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 (Figure 2). This distribution indicated that bacteria communities with low *metabolic capacities* and small average genome size are constrained along the second dimension. The MCOA2 separated communities according to genomic traits for *environmental responsiveness* and *nutrient recycling* (Figure 2). Communities associated with the lower end of MCOA2 were enriched in mineral and organic N and P assimilation genes (Figure 1-2). 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 (Figure 1-2). The communities were also enriched in genes encoding for carbohydrates metabolism of simple substrates like starch, glycogen, and oligosaccharides. Thus, the second trait dimension captured a gradient in the average environmental responsiveness that was positively associated with a specialisation in simple carbon substrate metabolism and negatively with nutrient assimilation and recycling capacities (Figure 2).

Drivers of the 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 could predict most of the variation in MCOA1 and MCOA2 with a R^2 of 0.80 and 0.58, respectively (Extended Data Figure 5). Mean decrease in mean square error (%MSE) and R squared calculated based on a ten-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 position along MCOA1 and MCOA2 depending on the combination of pH and annual precipitation (Figure 3-4). MCOA1 and MCOA2 were also driven by precipitation seasonality whereas soil C:N controls only MCOA1 (Figure 3-4, Extended Data Figure 5). Next, we projected the global variation in the trait dimensions using these random forests (Figure 4 B and 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 Figure 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 characterised by a small genome size (i.e., low MCOA1 value) were enriched under neutral to alkaline pH, low C:N, low annual precipitation but high precipitation seasonality (Figure 4A). Conversely, communities with larger genome sizes (high MCOA1 value), were found in more acidic soils as well as soil with higher C:N and climate with elevated stable precipitation (Figure 4A). Globally, these environmental controls predicted low MCOA1 coordinates (< -1) under arid and semi-arid climates at tropical and subtropical latitude as well as in the steppe zones of central Asia and North America (Figure 4B). Conversely, high MCOA1 coordinates (>1) were seen in equatorial forests as well as some temperate zones in northern Europe, 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 (Figure 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 (>2500mm) including equatorial zones of South-America and Asia and wet Europe and North America temperate zones. Medium-high coordinates (0< MCOA2 < 1) covered most of the globe, characterising all tropical-dry, semi-arid and subarctic regions. The projection of this dimension

(Figure 4D) predicts very high coordinates (MCOA2 >1) under limited regions of subtropical and high latitudes combining low annual precipitation (<1000mm) and very acidic pH (<4).

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

Discussion

Our study describes two dominant dimensions of community aggregated traits variation across soil bacteria communities (Figure 2-3). 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 incorporates traits previously identified as CSR strategies[1,11,12] (Extended Data Table 1). Moreover, it fits into a triangle like the original CSR model[7,23] (Figure 2-3) 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 tradeoff 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 (Figure 2-3). We hypothesise that these combinations indicate either that the communities are composed of taxa with different strategies or that the majority of bacteria living in soil need stress tolerant 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 (eg.[24,25]). In particular, Liu et al. (2023) 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 and pH near neutrality as observed in [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 the 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. (2020). Consistent with competitor traits, these attributes are favoured under stable and wet climates, that reduce 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, bacteria communities differentiate along the second dimension with opposing profiles of either increased capacities for *environmental responsiveness* or for *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 tolerant strategies (Extended Data Table 1). This includes traits to resist stress, sensing favourable environmental conditions, activate fast growth, and 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 specialised 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 to carbon metabolism. Microbial mineralisation activity and biomass turnover release nutrients and necromass into soil that this profile seems optimised to recycle. Such traits might reflect a strategy that emphasises 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 maximise their responsiveness or a reinforcement of competitor traits that favour nutrient recycling.

Overall, our dimension of *metabolic capacities* matches the versatility dimension described by Westoby et al. (2021) across cultured bacterial taxa, with both studies supporting that genome size plays a central role in differentiating bacteria strategies**.** Our dimension opposing *environmental responsiveness* and *nutrient recycling* also shows some consistencies with the second trait-dimension described by Westoby et al. (2021) capturing a rate-yield tradeoff, with rrn copy number as principal trait. Indeed, as discussed above, the traits of the *nutrient recycling* profile might favour growth yield, and high *environmental responsiveness* is associated with higher rrn copy number. However, these variations of rrn copy numbers have only a limited importance in the second trait dimension of our study, contrasting with the

observations of Westoby et al. (2021) 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 Figure 3). These observations are consistent with Gao and Wu (2018) reporting 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 might be present in the soil community, with their populations increasing during resource flushes (eg.[35]). In the oligotrophic environment of soil, our results suggest that increased capacity to recycle resources efficiently, to sense favourable conditions and to survive or escape stressful ones represent more common adaptations for bacteria than growing more rapidly. Investigating the variation of these traits across taxa in soil and their distribution within communities represents a challenging, but fascinating perspective to disentangle how the trait dimensions across taxa scales up to the community level. Overall, 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 soil environment.

Soil bacteria remain poorly characterised 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 the 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 likely belong to accessory genes and not to core metabolism 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 (Expended Data Fig 7). This suggests that metabolic expansion during different evolutionary histories can converge to similar life history strategy[40]. Hence, phylogenetic distance become 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 (eg. solubilization of toxic ions Al3+) and biotic (eg. plant and fungal communities) characteristics of the ecosystem[44–46]. The environmental distribution of the 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 (Figure 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 (Figure 3B and 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 (Sup Figure 8) 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 Figure 8)[2]. We selected this dataset for our analysis because of its coverage and its use of a highly standardised protocol that: 1) sampled top-soils in spatially independent sites across the globe selected to represent all the most important vegetation types; 2) analysed soil chemistry and metagenomes[2]. All samples were processed using similar standardised protocols for their chemistry (carbon, nitrogen, phosphorus content and pH_{H20}) and metagenome (See[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://nelson.wisc.edu/sage/data-and-models/atlas/maps.php). This showed an almost complete global coverage, with only extreme MAT of very high latitude (below -11.33°C) and Sahelian Africa (above MAT 27.97°C) as well as very high pH (higher than 7.76) characterising some parts of North Africa, West Asia and Himalaya missing in our dataset (Extended Data Figure 8). As far as we know, when we conducted this analysis, this dataset was the only available with such precise characterization of soil environment done on the same sample as shotgun metagenomic analysis, making this dataset the most

robust for our objective to assess environmental drivers of metagenomic profiles. Nevertheless, potential to extend environmental range by adding all (excluding agricultural and contaminated) soil metagenomes available (accession date January 28 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 (excepted a few samples from very cold sites with mean annual temperature lower than -11.5°C available on MG-RAST) and this would have greatly decreased precision of soil properties characterization (Extended Data Figure 9).

Metagenomic and amplicon sequencing data

DNA extraction, sequencing (Illumina with RTA Version 1.18.54 and bcl2fastq v1.8.4), trimming and mapping approaches are detailed in Bahram et al. (2018). In this study, four community aggregated trait databases were built, corresponding to metagenomic reads mapping on different functional annotation systems by Bahram et al. (2018). 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 characterise phylogenetic distances between bacterial communities using the Unifrac metric[50]

Bacterial community aggregated trait calculation

Bahram et al. (2018) 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 (e.g. eggNOG categories include Motility, Cell envelopes and Defense 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 abundance 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 2) based on previous works[52–54]. After mapping, the relative abundance of each gene (or aggregated group of genes) was normalised by the total number of bacteria-reads annotated for this sample on the same database. Such normalisation 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 (EPS)[57], Chaperons[12,58], Chemotaxis, and Osmolytes[59–62], antibiotic resistance and carbohydrates degradation enzymes (CAZyme). In addition, average genome size was calculated using MicrobeCensus[63] and rrn copy number using the method described in [64]. All sequences were used as input for average genome size and rrn copy number, after a verification 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 multi-table co-inertia analysis (MCOA), an exploratory analysis that leverages together the information from the 5 databases (genomic 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 summarise in a common structure the information shared by multiple multivariate (eg. omic) tables[65–67]. All variables (CATs) were log transformed (log X +1) before the analysis to improve normality[67] and standardised to a mean of zero and a variance of 1. The R package ade4 was used for the MOCA 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 as metagenome (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 [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 optimised based on randomForest's tuneRF algorithm and the number of trees set to 1000. Third, the random forest models selected following the VSURF selection process were trained using ten-fold cross-validation (100 repetitions) implemented in the caret package[71] and model performance was assessed based on Root Mean Square Error (RMSE) and R squared. Finally, random forest predictive models were used to project a broad resolution map of trait dimension global biogeography using environmental maps (1600x1200 pixel) as predictors. For this projection, we used the the latest map (June 2022) released by ISRIC's World Soil Information Service (https://files.isric.org/soilgrids/latest/data_aggregated/) based on SoilGrids version 2.0[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 to represent 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 two first 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.

Data availability

The five CAT databases used to build the trait dimensions and the associated environmental variables are available on figshare repository : 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.

Acknowledgements

We thank Leho Tedersoo and Peer Bork who conceived and supervised the acquisition of the global dataset used in this study with Mohammad Bahram and Falk Hildebrand. We also wish to thank all their collaborators who contribute to this global data acquisition effort. We also thank Alyse Larkin and Lucas Ustick for their guidance in the bioinformatic analysis conducted in this study. Finally, we thank Kate Buckeridge and two anonymous referees for their insightful reviews. GP, SA, JM, KT and AM were

supported by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research grants DE-SC0016410 and DE-SC0020382. FH was supported by the European Research Council H2020 StG (erc-stg-948219, EPYC) and the Biotechnology and Biological Sciences research Council (BBSrC) Institute Strategic Program Gut Microbes and Health BB/r012490/1 and its constituent project BBS/e/F/000Pr10355.

Authors 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. Idea of this new analysis was conceived by G.P. with inputs from A.M., S.A, J.M. and K.T. New quantification of genomic traits, Unifrac and data analyses were performed by G.P. First draft and following editing was conducted by G.P. with inputs from all co-authors.

Competing interests

The authors declare no competing interests.

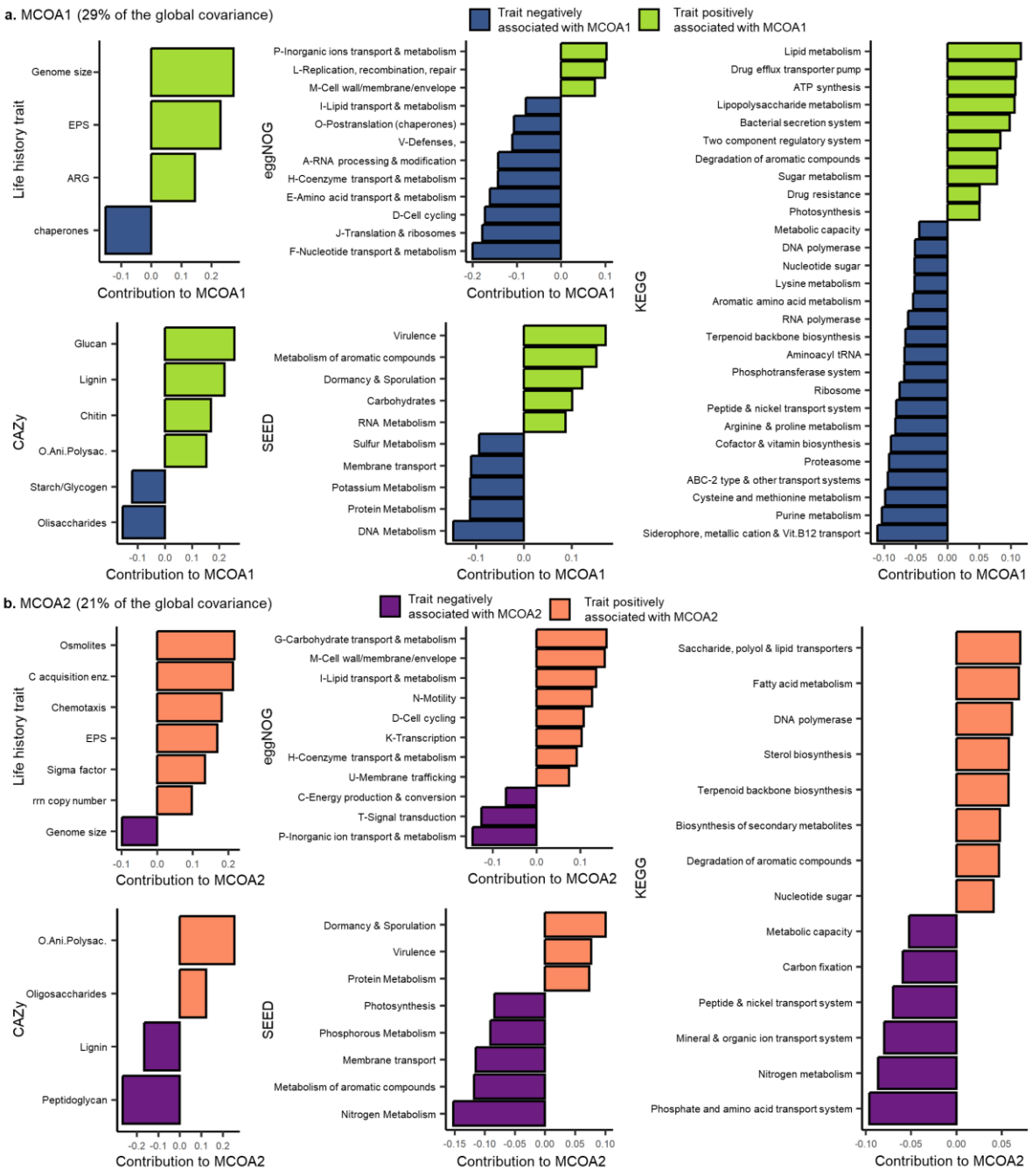


Figure 1. Global trait dimensions of soil bacteria metagenomes. Variable contributions to the multiple co-inertia analysis (MCOA) summarising in a common structure (MCOA dimensions 1 and 2) the information shared by 5 community aggregated trait (CAT) databases (Life history trait, CAZy, eggNOG, SEED and KEGG). Only the most important variables with significant correlation ($p<0.001$) with each dimension are reported in this figure. ***a*** *and* ***b*** *panels present variable contributions to MCOA*

Dimension 1 and 2 respectively. Bar colours indicate the direction of the associations between the variable and the MCOA dimensions.

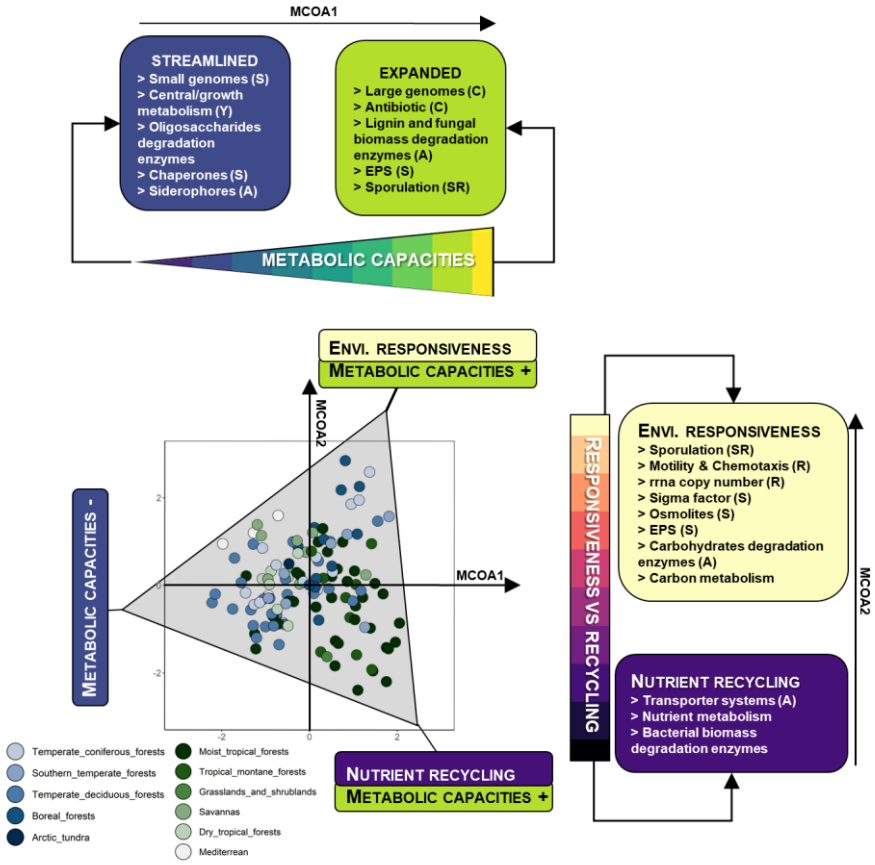
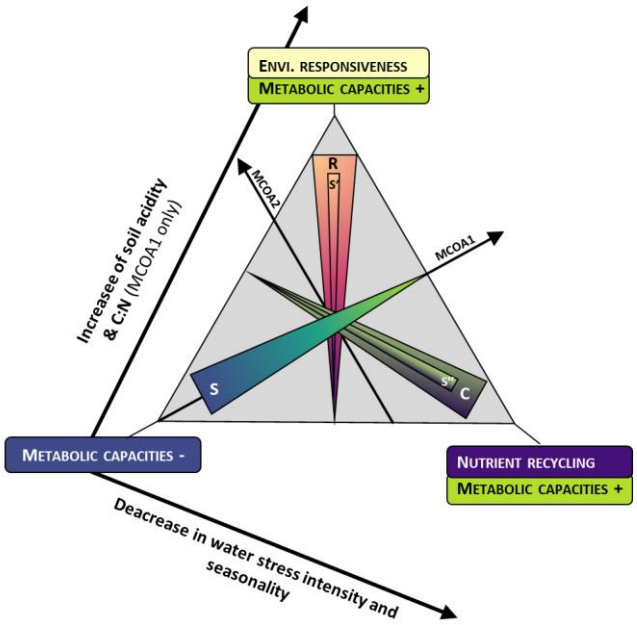
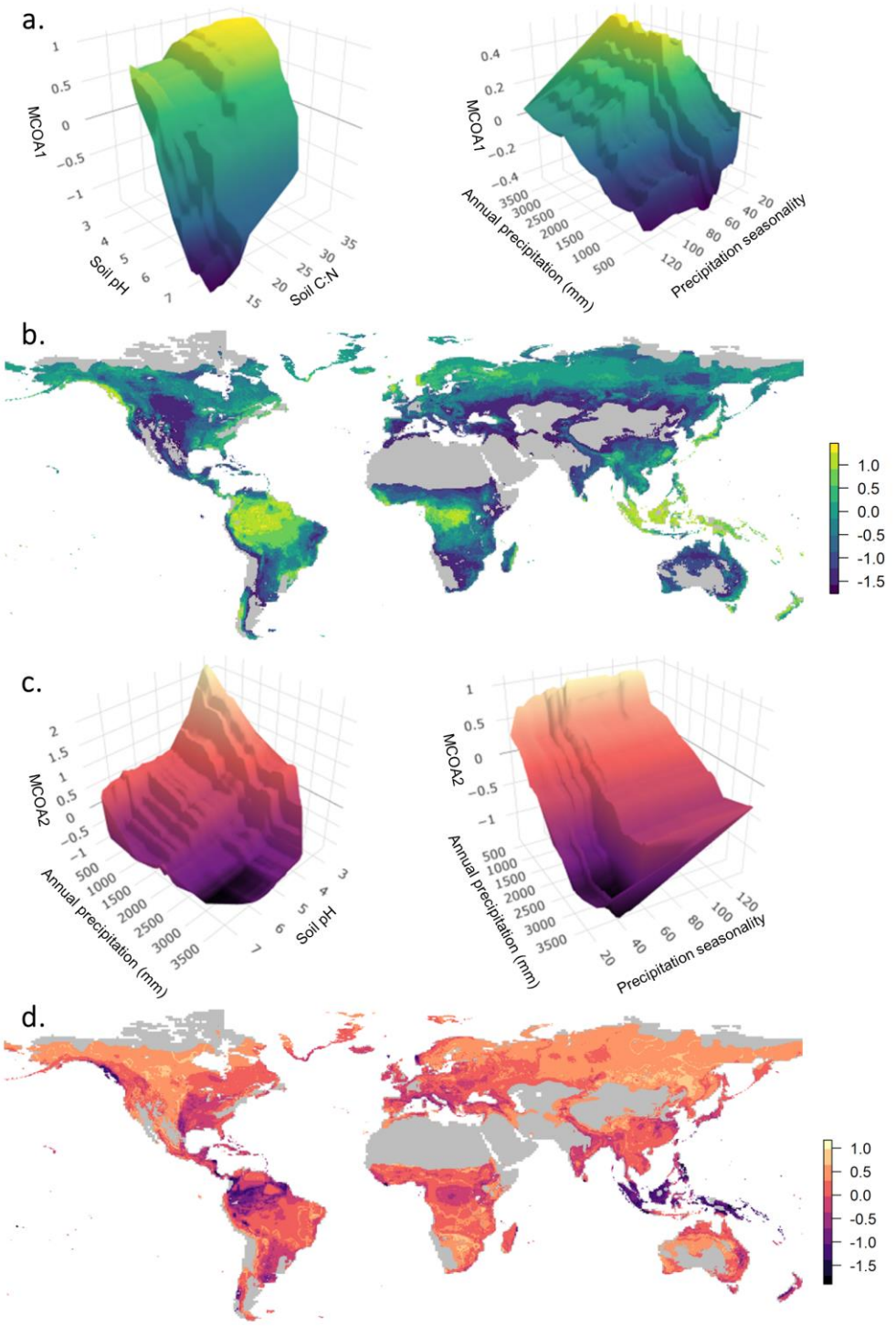


*Figure 2. The global life history strategies of soil bacteria communities. **T**wo-dimensional trait space from a MCOA depicting trait associations across soil bacteria communities, with traits inferred from enriched genes in bacteria metagenomes. Dots represent the positions of the 128 bacterial communities used in this study along these two dimensions. In the trait lists, letters in brackets represent how CSR (Competitors, Stress tolerant, Ruderal) and YAS (High Yield, resource Acquisition, Stress tolerance) strategies have been associated with these traits in previous theoretical works (Extended Data Table1).*



428

429 *Figure 3. Hypothesised role of competitor (C), Ruderal (R) and Stress tolerant (S) traits in shaping the*
430 *life history strategy observed at the community level and associated environmental gradients. Details*
431 *of CSR traits association are provided in Figure 1 and 2. S, S' and S'', represent the different S traits*
432 *associated with each dimension and detailed in Figure 1 and 2.*



433

434 *Figure 4. Environmental control and global scale projection of bacterial communities' coordinates*
435 *along MCOA dimension 1 and 2.* ***a*** *and* ***c****, random forest partial dependence plots describing*
436 *relationships between bacterial communities' coordinates along MCOA dimension 1 (****a****) and 2 (****c****) and*
437 *their most significant environmental predictors (Extended Data Figure 5).* ***b*** *and* ***d****, random forest*

*predictions for MCOA dimension 1 (**b**) and 2 (**d**) projected across the globe using broad resolution map of mean soil and climate conditions (1600x1200 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 version 2.0 was used for soil properties and Worldclim2 for climate variables. Accuracy of the prediction was verified by ten-fold cross-validation of the random forest (Extended Data Figure 5) and by comparing the predicted values of the broad resolution projection with local observations (Extended Data Figure 6).*

References

1. Fierer, N. Embracing the unknown: disentangling the complexities of the soil microbiome. *Nature Reviews Microbiology* **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. *Nature Climate Change* **3**, 909–912 (2013).
6. Diaz, S. *et al.* The global spectrum of plant form and function. *Nature* **529**, 167 (2016).
7. Grime, J. P. Evidence for the existence of three primary strategies in plants and its relevance to ecological and evolutionary theory. *The American Naturalist* **111**, 1169–1194 (1977).
8. Wright, I. J. *et al.* The worldwide leaf economics spectrum. *Nature* **428**, 821 (2004).
9. Southwood, T. R. Habitat, the templet for ecological strategies? *Journal of animal ecology* **46**, 337–365 (1977).
10. Reich, P. B. *et al.* The evolution of plant functional variation: traits, spectra, and strategies. *International Journal of Plant Sciences* **164**, S143–S164 (2003).
11. Krause, S. *et al.* Trait-based approaches for understanding microbial biodiversity and ecosystem functioning. *Frontiers in Microbiology* **5**, (2014).
12. Malik, A. A. *et al.* Defining trait-based microbial strategies with consequences for soil carbon cycling under climate change. *The ISME Journal* **14**, 1–9 (2019).
13. Madin, J. S. *et al.* A synthesis of bacterial and archaeal phenotypic trait data. *Scientific Data* **7**, 1–8 (2020).
14. Westoby, M. *et al.* Trait dimensions in bacteria and archaea compared to vascular plants. *Ecology Letters* (2021).
15. Steen, A. D. *et al.* High proportions of bacteria and archaea across most biomes remain uncultured. *The ISME journal* **13**, 3126–3130 (2019).
16. Martiny, A. C. High proportions of bacteria are culturable across major biomes. *The ISME Journal* **13**, 2125–2128 (2019).
17. Martiny, A. C. The "1% culturability paradigm"needs to be carefully defined. *The ISME journal* **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. *Frontiers in microbiology* **5**, (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. *The ISME journal* **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. *Nature microbiology* **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* e70 (2023).
25. Simonsen, A. K. Environmental stress leads to genome streamlining in a widely distributed species of soil bacteria. *The ISME Journal* **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 Biology and Biochemistry* **178**, 108935 (2023).
27. Guieysse, B. & Wuertz, S. Metabolically versatile large-genome prokaryotes. *Current Opinion in Biotechnology* **23**, 467–473 (2012).
28. Konstantinidis, K. T. & Tiedje, J. M. Trends between gene content and genome size in prokaryotic species with larger genomes. *Proceedings of the National Academy of Sciences* **101**, 3160–3165 (2004).
29. Paul, C. *et al.* Bacterial spores, from ecology to biotechnology. *Advances in applied microbiology* **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. *Journal of Genetic Engineering and Biotechnology* **19**, 1–19 (2021).
31. Sinsabaugh, R. L. & Follstad Shah, J. J. Ecoenzymatic stoichiometry and ecological theory. *Annual Review of Ecology, Evolution, and Systematics* **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. *Communications Earth &ampmathsemicolon Environment* **1**, (2020).
33. Zheng, Q. *et al.* Growth explains microbial carbon use efficiency across soils differing in land use and geology. *Soil Biology and Biochemistry* **128**, 45–55 (2019).
34. Gao, Y. & Wu, M. Free-living bacterial communities are mostly dominated by oligotrophs. *bioRxiv* 350348 (2018).
35. Li, J. *et al.* Predictive genomic traits for bacterial growth in culture versus actual growth in soil. *The ISME journal* **13**, 2162–2172 (2019).
36. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nature biotechnology* **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. *Microbial Genomics* **6**, (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. *Proceedings of the National Academy of Sciences* **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 Biology and Biochemistry* 107584 (2019).

42. Fernández-Calviño, D. & Bååth, E. Growth response of the bacterial community to pH in soils differing in pH. *FEMS microbiology ecology* **73**, 149–156 (2010).
43. Auger, C. *et al.* Metabolic reengineering invoked by microbial systems to decontaminate aluminum: implications for bioremediation technologies. *Biotechnology advances* **31**, 266–273 (2013).
44. Bruelheide, H. *et al.* Global trait–environment relationships of plant communities. *Nature Ecology & Evolution* **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. *Frontiers in Microbiology* **11**, 1953 (2020).
46. Bagousse-Pinguet, Y. L. *et al.* Testing the environmental filtering concept in global drylands. *Journal of Ecology* **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**, 1–8 (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 research* **47**, D666–D677 (2019).
50. Lozupone, C. & Knight, R. UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and environmental microbiology* **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 research* **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 Biology and Biochemistry* **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 Biology and Biochemistry* **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. *Journal of bacteriology* **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. *Frontiers in microbiology* **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. *Applied Soil Ecology* **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. *Frontiers in microbiology* **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. *Scientific reports* **6**, 1–14 (2016).
61. Boch, J., Kempf, B., Schmid, R. & Bremer, E. Synthesis of the osmoprotectant glycine betaine in Bacillus subtilis: characterization of the gbsAB genes. *Journal of Bacteriology* **178**, 5121–5129 (1996).

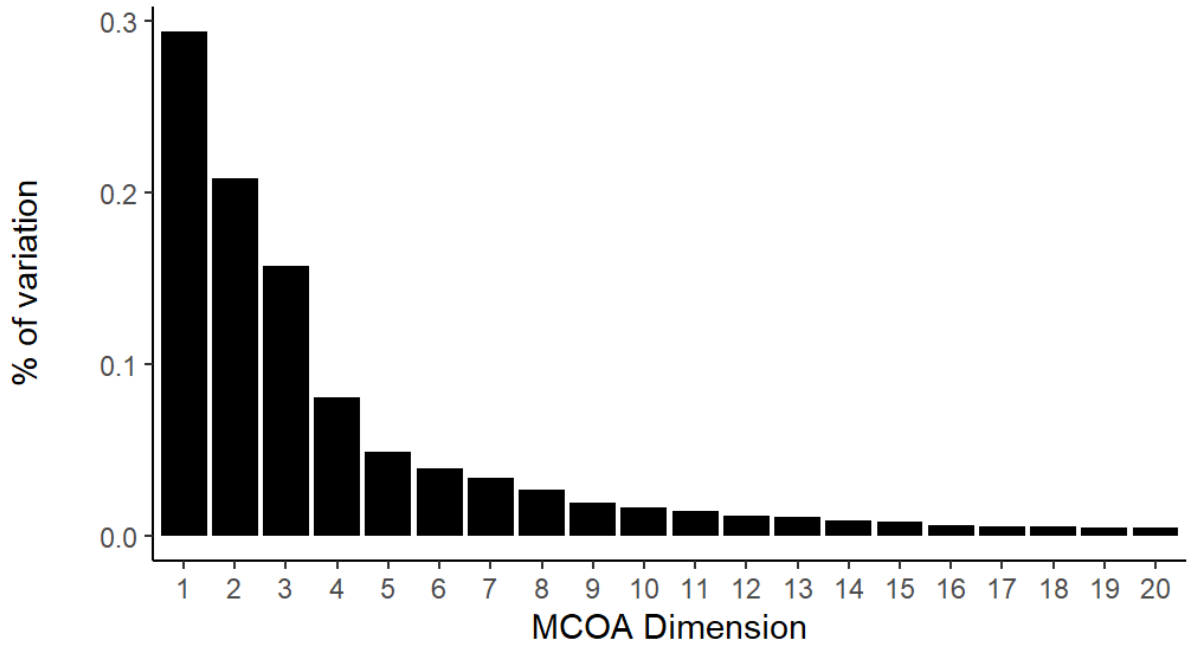
62. Wargo, M. J. Homeostasis and catabolism of choline and glycine betaine: lessons from Pseudomonas aeruginosa. *Applied and environmental microbiology* **79**, 2112–2120 (2013).
63. Nayfach, S. & Pollard, K. S. Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. *Genome biology* **16**, 1–18 (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**, 1–13 (2019).
65. Chessel, D. & Hanafi, M. Analyses de la co-inertie de K nuages de points. *Revue de statistique appliquée* **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. *Journal of Ecology* **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**, 1–13 (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. *The R Journal* **7**, 19–33 (2015).
71. Kuhn, M. Building predictive models in R using the caret package. *Journal of statistical software* **28**, 1–26 (2008).
72. Poggio, L. *et al.* SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. *Soil* **7**, 217–240 (2021).

Extended Data

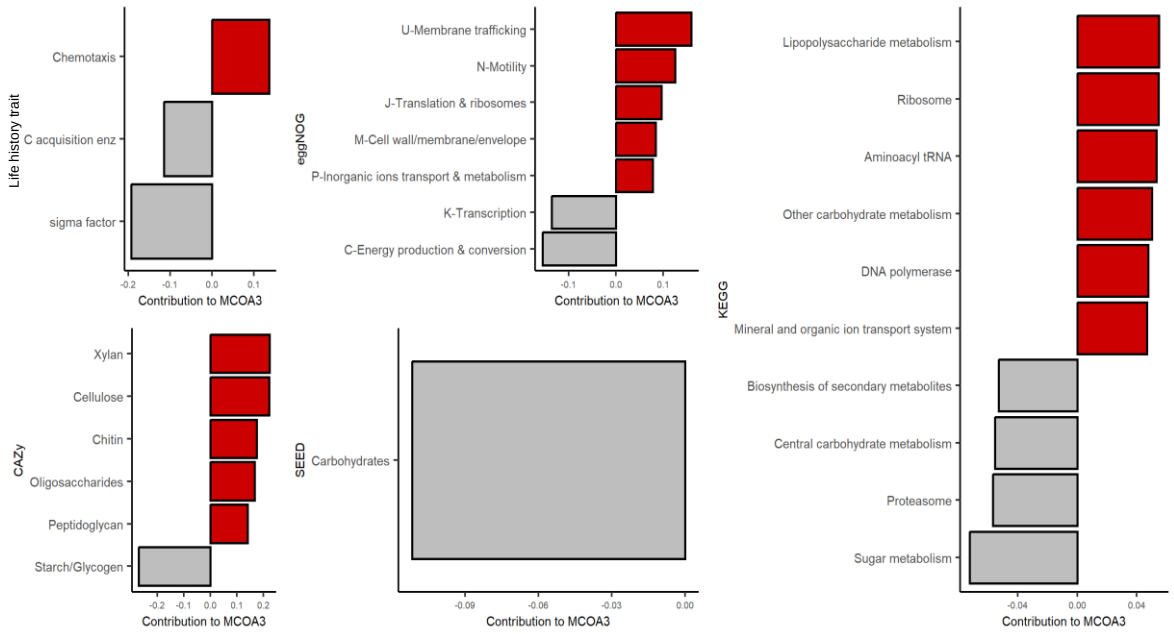
Extended Data Table 1. Life history traits used in this study. Traits were selected based on their previous association with CSR ('Competitor', 'Stress tolerant', and 'Ruderal') strategies by Fierer (2017) [1] or Krause et al. (2014) [2] or YAS strategies ("Yield","Resource acquisition", and "Stress tolerant") by Malik et al. (2020) [3]. 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 (Malik et al. 2020).

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

615

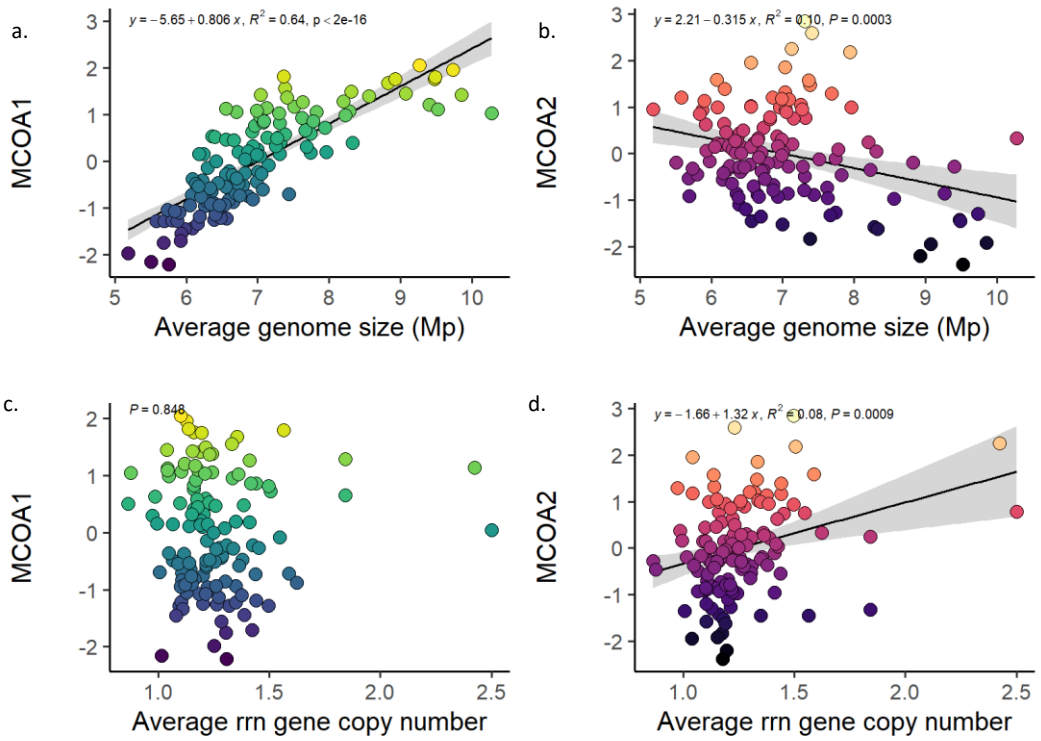


Extended Data Figure 1. Stress plot representing the % of variation of the global dataset captured by each dimension of the MCOA



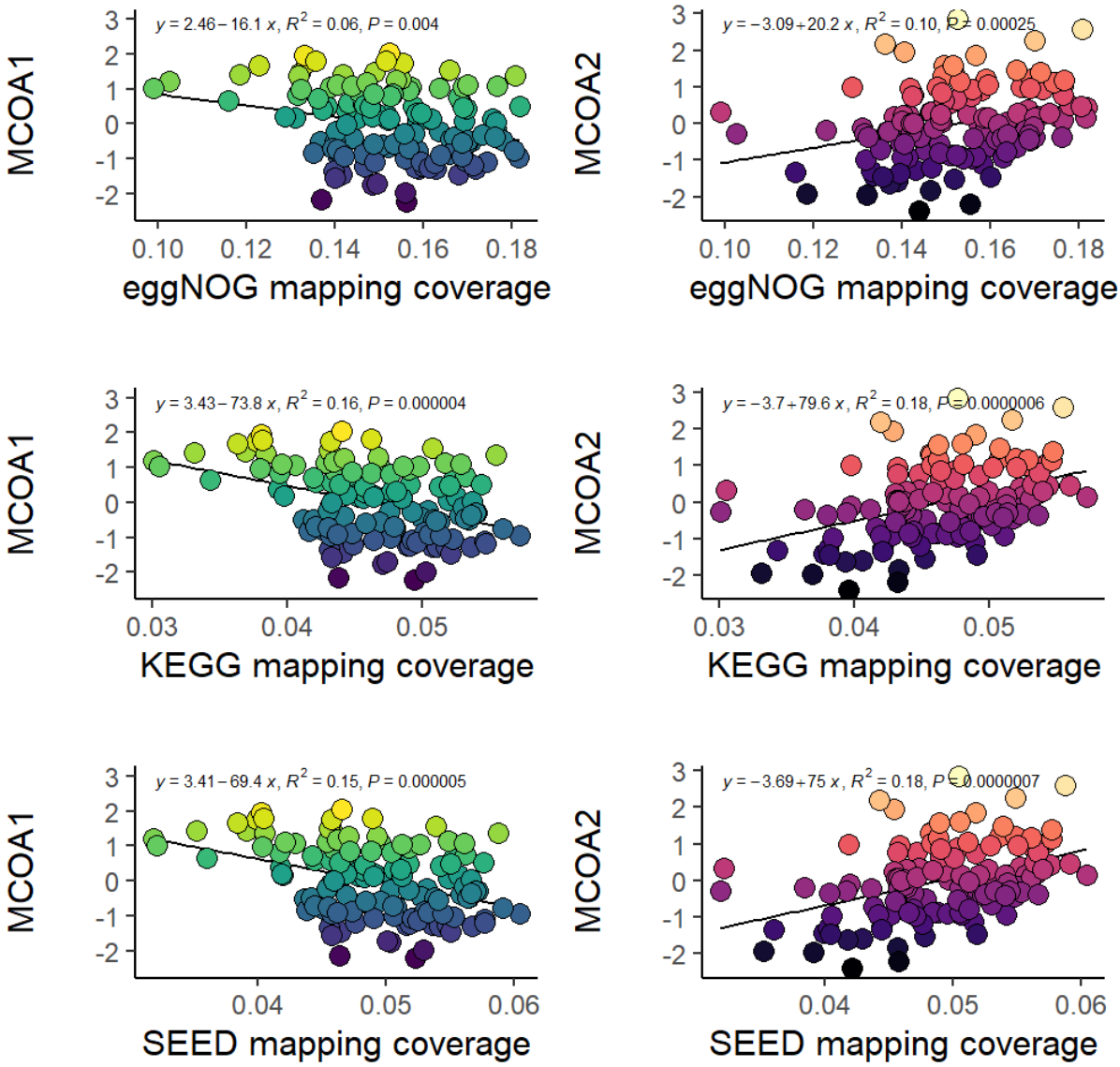
620

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



626

627 *Extended Data Figure 3. Correlations between genomic traits and coordinates along dimensions 1 and*
628 *2 of the MCOA. The P value indicates the significance of the regression slope obtained using a t-test.*
629 *Shade represents the estimated 95% confidence interval. Color gradients follow MCOA dimensions and*
630 *match with figure 1 and 3 in the main text.*



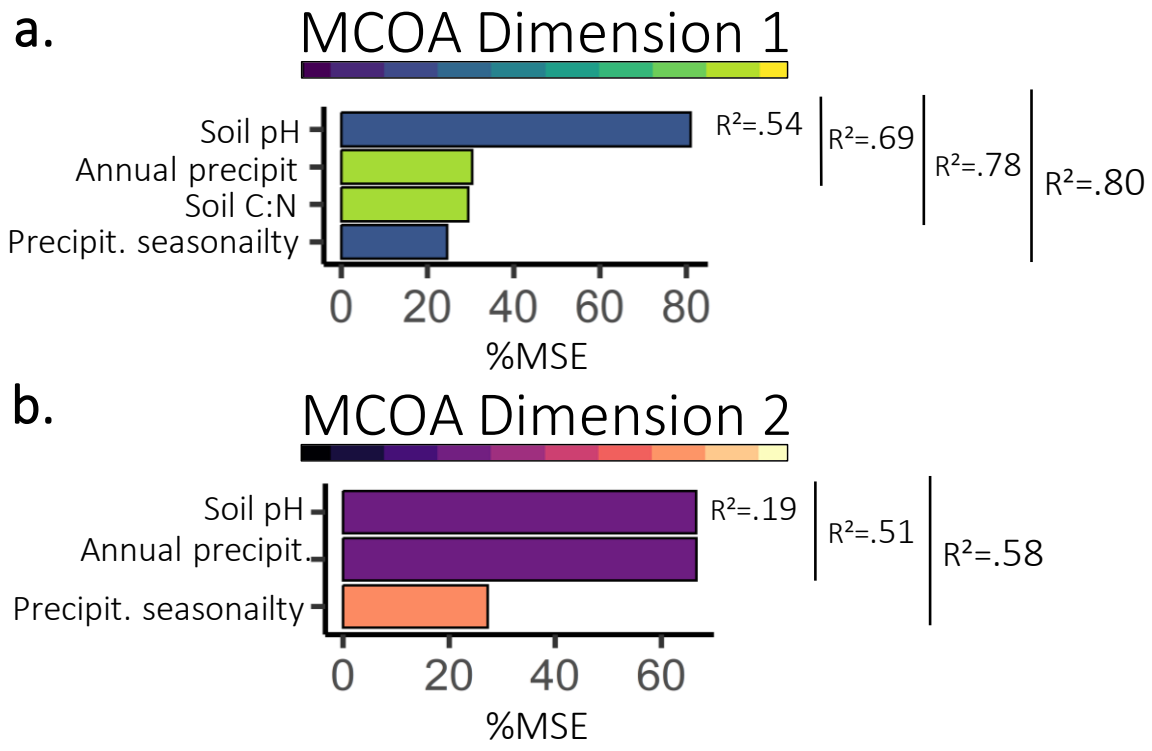
631

632 *Extended Data Figure 4. Correlations between MCOA dimensions (MCOA1 and MCOA2) and mapping*

633 *coverages on the 3 general databases (eggNOG, KEGG, SEED) used in this study. The P value indicates*

634 *the significance of the regression slope obtained using a t-test.*

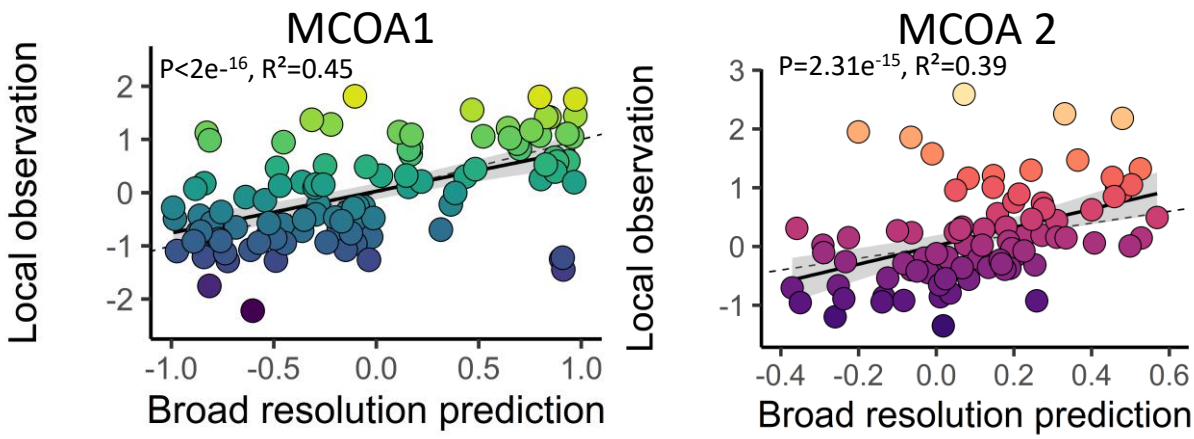
635



636

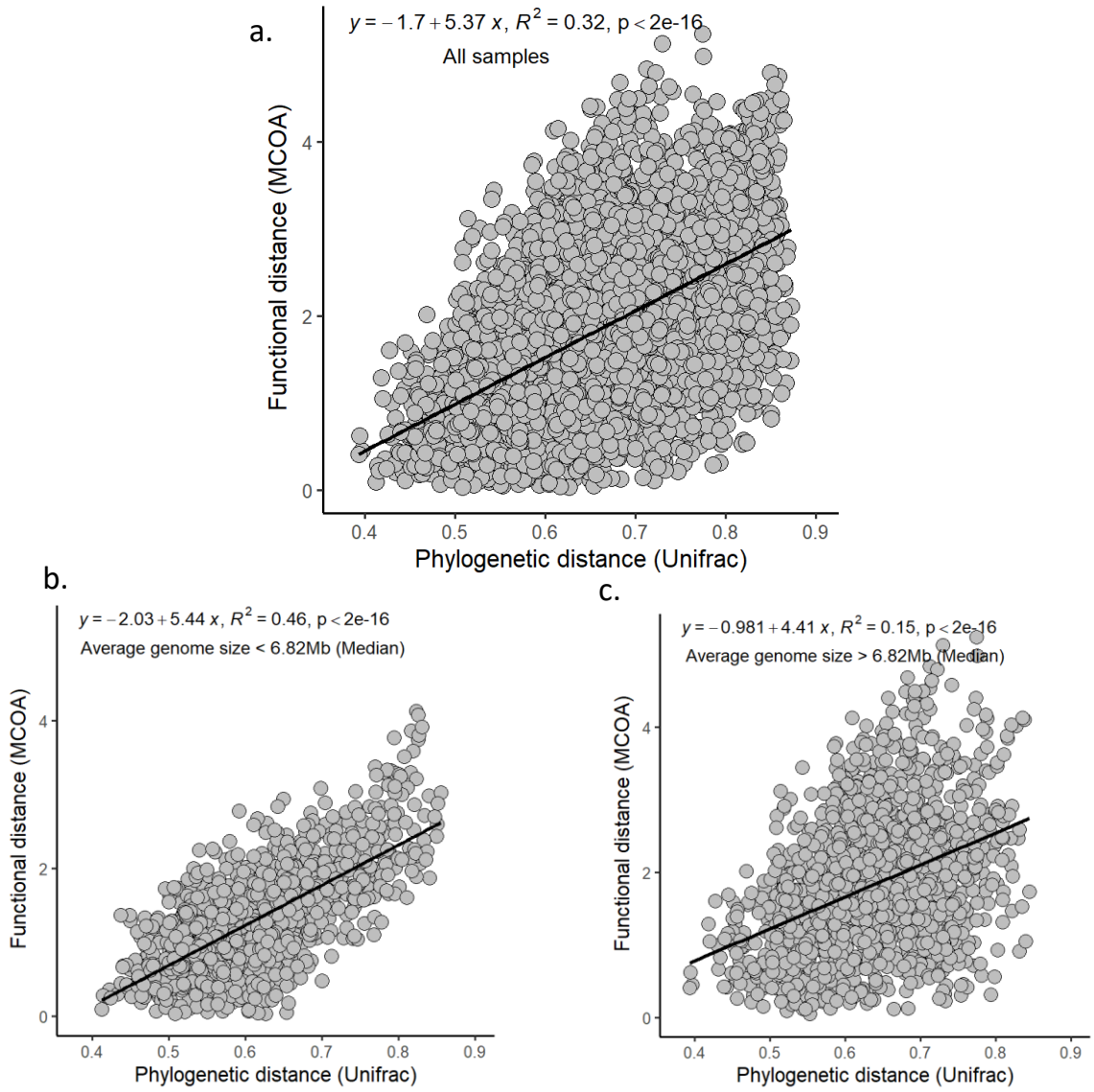
637 *Extended Data Figure 5. Environmental drivers of the bacterial community trait dimensions.*
638 *Environmental variable importances are represented as the mean decrease in mean square error*
639 *(%MSE) and R squared in random forest models predicting MCOA Dimension 1 (**a**) and 2 (**b**). Bar*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

641



642

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

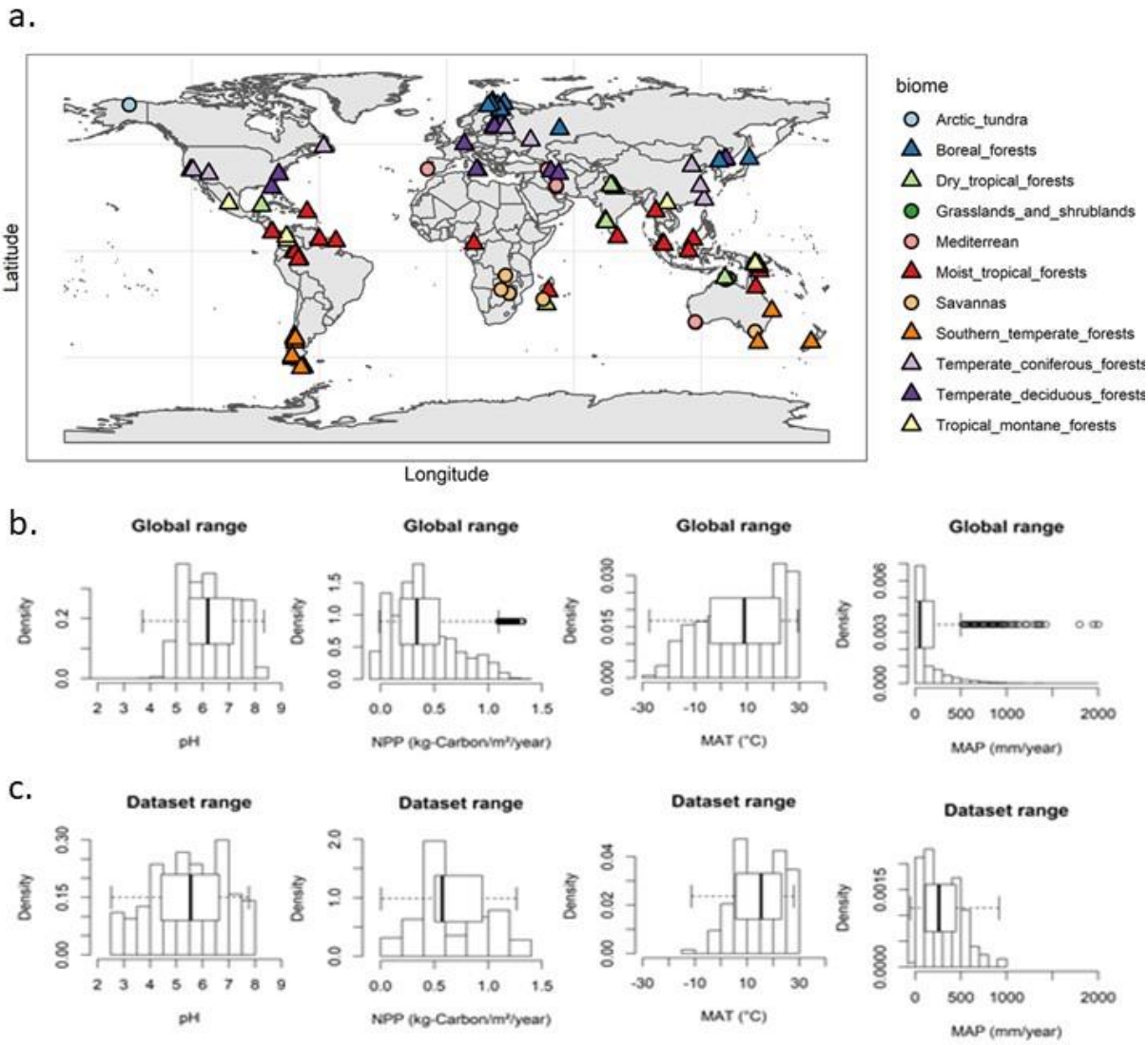


650

651

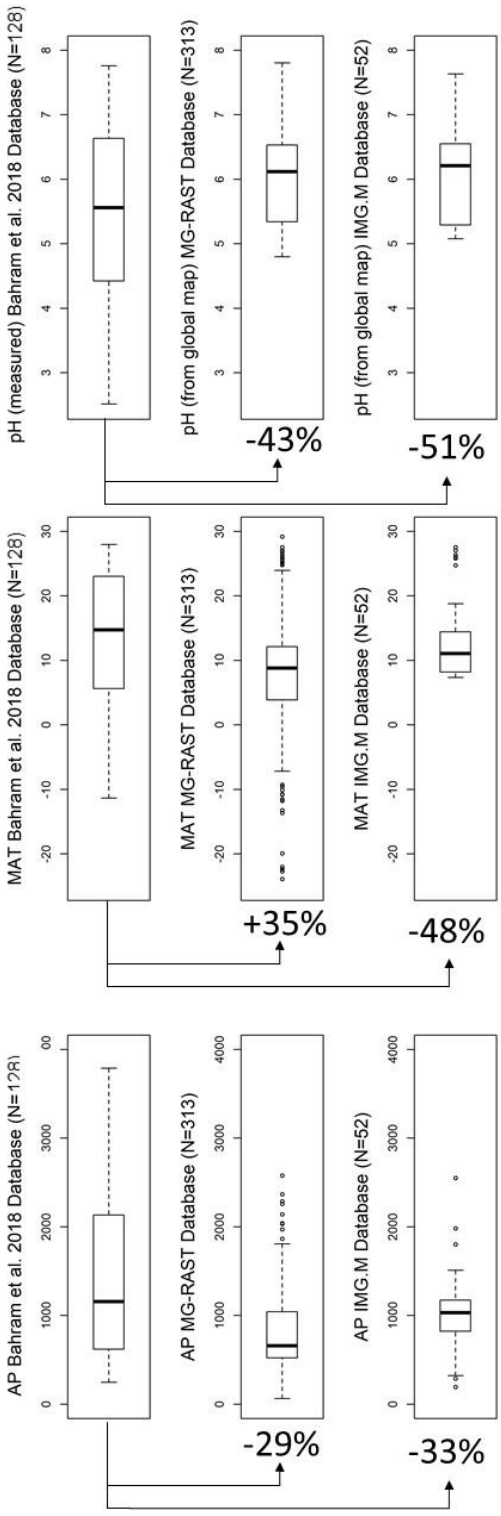
652 *Extended Data Figure 7. Correlation between phylogenetic distance (Unifrac metric) and functional*
653 *distance (Euclidian distance in MCOA space using coordinates of the two principal dimensions).*
654 *Correlation for all samples (a) and restricted to samples with average genome size below (b)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

657



658

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



Extended Data Figure 9. Environmental coverage comparison between the database used in this study from Barham et al. (2018) 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=1.5x interquartile range; points=outliers.