Environment and soil pH as a driver for soil bacteria diversity and function - A study with Earth Microbiome Project

Introduction

Bacteria are critical for the functioning of the ecosystem at different scales. Important nutrient cycles such as nitrogen, phosphorus, and carbon depend on bacteria in soils. In addition, bacteria can have symbiotic and pathogenic associations with plants that can influence plant productivity and vegetation composition. Studying soil bacteria under different environmental factors and soil conditions is important to understand ecosystem dynamics.

Moreover, Bateria can be affected by climate, vegetation, soil resources, and soil conditions. Subsequently, we can find different patterns of bacteria diversity and abundance globally. As an important factor of soil condition, soil pH has been significantly important in explaining bacteria diversity, abundance, and functions. Motivated by Bacteria's critical role in the ecosystem, here we analyzed data from the Earth Microbiome Project and observed that soil bacteria diversity and function are influenced by different environments and soil pH.

Data description

Earth Microbiome Project (EMP) used a systematic approach to characterize microbial taxonomic and functional diversity across different environments and humankind (Thomson et al., 2017). EMP comprises 27,751 samples from 97 studies with microbial data representing 16S rRNA amplicon sequencing, metagenomes, and metabolomics. For this study, we used the data that had been rarefied by EMP. The dataset was further split into an operational taxonomic unit (OTU) table, a sample table, and a metadata table. Below is the column information for each table.

(a) OTU table: ID, Sequence, Kingdom, Phylum, Class, Order, Family, Genus, and Species.
(b) Sample table: ID and Sample Name.
(c) Metadata table: 76 environmental information from each sample. For this report, we used Sample ID, Environment Biome, Environment Feature, and soil pH.

To identify functionality by taxonomy we used the FAPROTAX database (Louca et. al., 2016). FAPROTAX used bacteria and archaea genera and species taxonomic information to assign metabolic and other relevant functions. The database consists of information collected by published literature on cultured strains. In total, it contains over 80 functions, 7,600 annotations, and more than 4,600 taxa. This database was created as an alternative to expensive methodologies like shogun sequencing for functional community profiling.

Methods

(1) Data pre-processing

In order to focus on the soil environment, we subsetted for **soil and rhizosphere** samples from the EMP database. Moreover, OTUs with a prevalence of at least ten samples were chosen. The selected data was used to build an OTU-sample bipartite

adjacency matrix, named *A*. *Aij = x* represented *x* number of OTU *i* were observed in sample *j*. Then, the non-zero entries went through square-root transformation for variance stabilization. The “A” matrix had a dimension of 116,658 rows and 945 columns (i.e. 116,658 kinds of OTUs collected from 945 sample sites). Figure 1 shows the overall OTU abundance distribution. The right-skewed distribution indicated that some OTUs were more abundant than others, which is natural in microbial ecosystems.

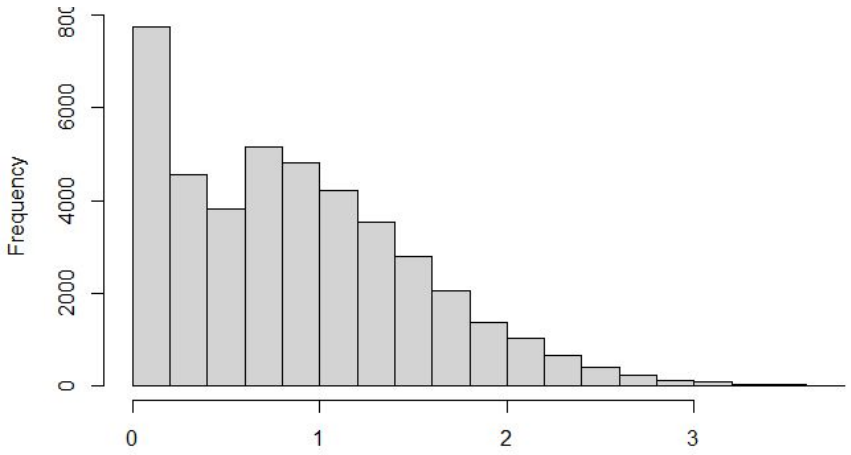

Figure 1. Histogram of OTU abundance for A matrix

(2) Clustering OTU-Sample graph into communities

We used Vintage Sparse PCA (VSP) to cluster the graph into 6 communities with the R *vsp* package. Six was chosen because there are 6 levels of biome label (defined in the next session) across the whole EMP database. As we hoped to estimate the □s in the Latent Dirichlet Allocation (LDA) model (Rohe and Zeng 2020) to obtain the relative phyla or function abundance, we operated vsp on the centered and scaled adjacency matrix .

(3) Contextualizing clusters with environmental factors

The “best feature function” or *bff* from the *vsp* R-package was used for contextualization. The bff function requires an *n* by *k* matrix of the weights that indicates how important “*i”th node* is for the “*j”th cluster*, an *n* by *d* matrix that contains features for each node in the network, and a number indicating how many features for differentiating between loadings.

It was expected that sample sites with similar OTU composition will be clustered together. Thus, *the Y* matrix was used as the *n* by *k* loading matrix, and the *n* by *d* feature matrix was created from the metadata table, which contained the environmental information of each sample site. In particular, two variables were chosen: *environmental biomes* and *environmental features*. The *biome* data classified the environment as a habitat related to plants, animals, and climate (e.g., Tropical forest, temperate grassland). The *feature* data classified the environment as a particular component of the environment where the sample was collected (e.g., soil, coral, bay).

Discussions

(1) Clusters

According to Figure 2. (unit of analysis: 1 sample), overall, radial streaks aligned well with the axis. Moreover, the B matrix was rather diagonal, indicating that *OTUs* from *OTU cluster i* (*Zi)* mainly exists in *Samples* from *sample cluster i (Yi)*. Thus, it was reasonable to conclude that there were approximately 6 communities in the bipartite graph. Moreover, from Table 1., through contextualization with environmental biomes and features, we could give each cluster a meaningful name: weathered, cold moist, dry, short vegetation, cold dry, and disturbed soil environment.

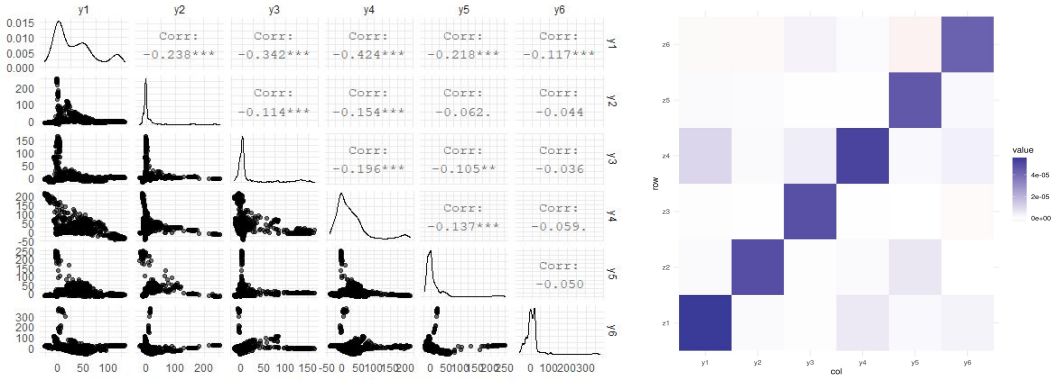

Figure 2. Cluster result diagnosis: left: Cluster pair plot; Right: B matrix

Table 1. Contextualization by environmental biome and feature

Weathered	Cold moist	Dry	Short vegetation	Cold dry	Disturbed
cultivated habitat	tundra	polar desert	vineyard	dry lake	cropland
forest soil	permafrost	cold temperature habitat	tropical shrubland	tundra	plant-associated habitat
tropical shrubland	bog	desert	volcano	tundra	cultivated habitat
volcano	tundra	dry soil	grassland soil	montane shrubland	desert
tropical moist broadleaf forest	montane shrubland	tundra	grassland	mountain	agricultural soil
forest	mountain	shrubland	desert	coniferous forest	dry soil
forest	plant-associated habitat	temperate grassland	forest soil	forest	shrubland
cropland	taiga	basin	dry soil	plant-associated habitat	bog
montane shrubland	peatland	rocky desert	montane shrubland	agricultural soil	dry lake
mountain	pasture	urban	agricultural soil	taiga	temperate grassland

(2) Soil pH distribution by cluster

To better characterize the communities, the distribution of pH values among clusters was examined in Figure 3. The red dots in each box were the average pH values. Overall, soil pH average and variability differed among clusters. “Dry” cluster had a higher pH than any other cluster. “Weathered”, “Coild moist” and “Cold dry” had acidic pH. “Disturbed” and “Short vegetation” had average neutral pH; however, “Disturbed” cluster showed great variability of pH values. The association between cluster and soil pH indicated that the role of pH may vary with circumstances. Specifically, the “Dry” cluster could be significantly influenced by soil pH.

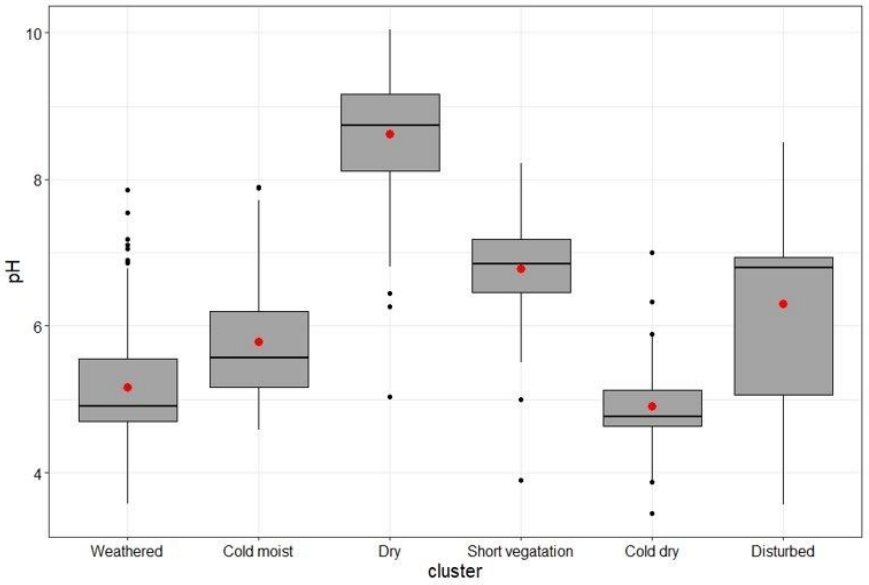

Figure 3. pH distribution of each cluster

(3) Species richness related to pH per cluster

To see if pH was an influential factor to diversity, we compared the species richness of each cluster with pH, as displayed in Figure 4. None of the clusters showed a linear relation between richness and pH. However, if we group the richness value by general biomes we could see some pattern that could explain the variability of richness within each cluster. In the short vegetation cluster, anthropogenic biomes had high richness, while shrublands had low richness. In the weathered cluster, anthropogenic biomes had high variability of richness but mostly in acidic pH values, while forests and shrublands had average richness in neutral pH.

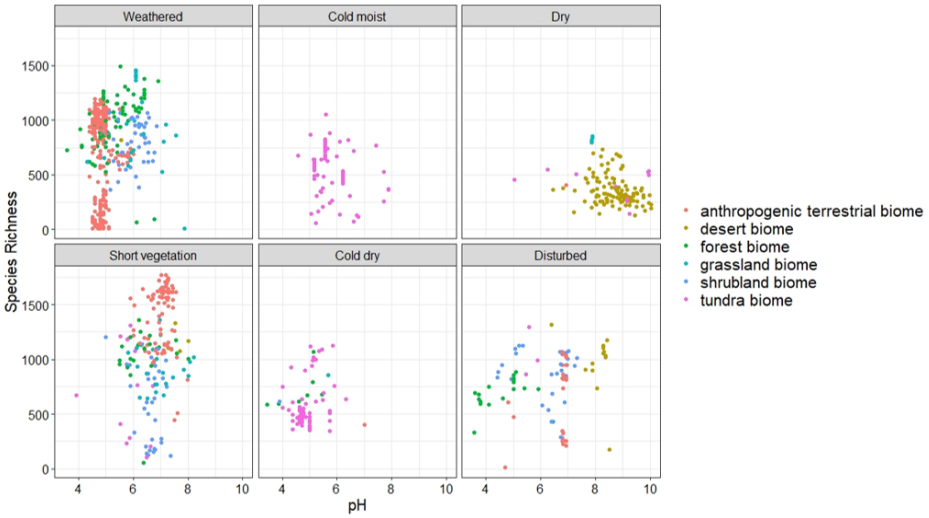

Figure 4. pH-Species Richness by cluster

(4) Most abundant phyla for each cluster and their relation with pH

To further investigate the microbial community structure, instead of examining the overall species richness, phyla abundance may provide more insights, as shown in Figure 5. Across clusters, the Acidobacteria and Proteobacteria were of the top abundant phyla, corresponding to the fact that these two phyla are commonly seen in nature across various environments.

Aside from the commonality, there were phyla that are relatively abundant under specific conditions. For example, under "Short vegetation", Planctomycetes wes more abundant, possibly relating to the fact that they can participate in degrading plant-derived polymers which are rich in short vegetation soil. Under "Cold dry", Verrucomicrobia were more prominent, aligning with the reason that Verrucomicrobia are one of the few phyla that still thrive under extreme icy conditions.

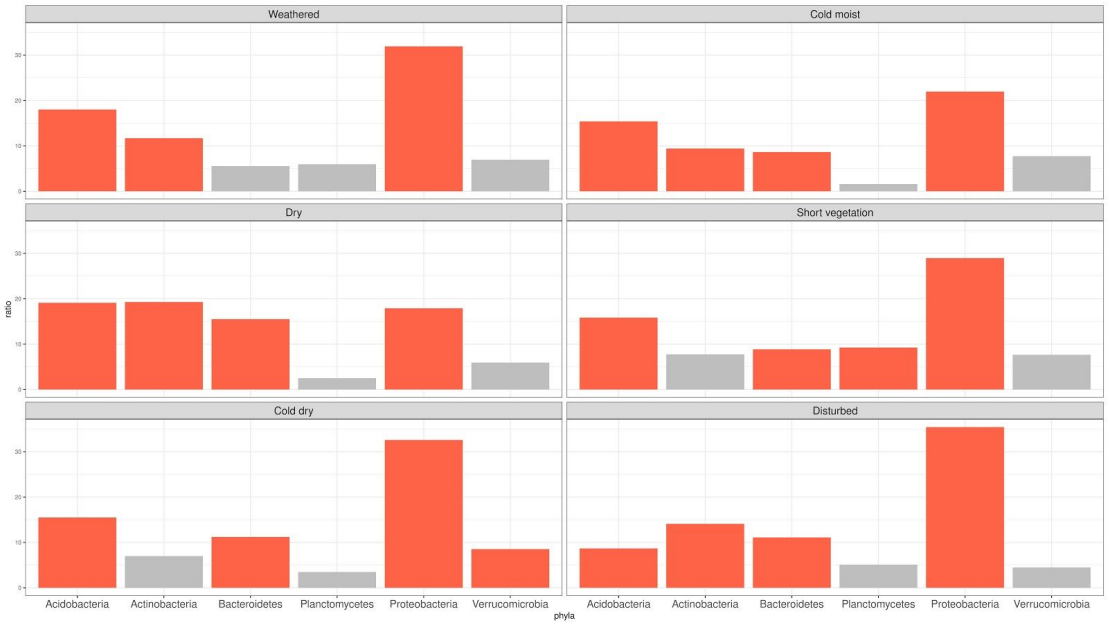

Figure 5. Phyla abundance of each cluster
(bars with abundance > 8% are encoded in red)

As phyla with low abundance were not of our interest, for each cluster, Figure 6 only showed the relationship between top most abundant phyla and pH. For example, under the "Weathered" environment, more abundant Bacteria (defined as > 8%) were Acidobacteria, Actinobacteria, and Proteobacteria; thus, only these were shown.

Overall, Acidobacteria, Actinobacteria, and Bacteroidetes thrived across pH and various environments due to the fact that these three phyla consist of highly diverse bacteria, and thus can live under all kinds of soil conditions. In particular, Actinobacteria abundance and pH were positively correlated. Other obvious trends between pH and phyla abundance were observed under the "Disturbed" environment. Acidobacteria and pH were negatively correlated, while Bacteroidetes and pH were positively correlated.

Figure 6. pH-phyla abundance by cluster

(5) Function for each cluster and their relation with pH

With the taxonomic path information, microbial function can be further mapped. Table 2. displayed the top related functions for each cluster. Common microbial functions across soil environments were those responsible for organic molecules decomposition, e.g. aerobic chemoheterotrophy. “Weathered”, “Short-vegetation”, and “Disturbed” environments had microbiomes that influence nutrient cycling (e.g. functions prefixed with nitrate, nitrite, nitrogen, and ammonia) and decomposition of carbon to carbon dioxide.

On the other hand, some clusters had unique specializations. “Cold moist” soil consisted of bacteria involved in anoxic carbon decomposition, e.g. methanotrophy, and fermentation. Specifically, methane is a green-house gas that is 84 times more potent than carbon dioxide in terms of heating up the earth. Thus, it is beneficial that under the cold moist environment, one of the vulnerable places deeply influenced by climate change, have these bacteria working to balance the methane concentration. Alternatively, under “Cold Dry” environments, aside from methane-related bacteria, there exist bacteria that live in plants as parasites or symbionts. Another obvious specialty is the "Dry" cluster, which contained bacteria that are dependent on sunlight for energy, e.g. photosynthetic cyanobacteria and photoheterotrophy.

Table 2. Contextualization by Function

Weathered	Cold moist	Dry	Short vegetation	Cold dry	Disturbed
nitrate denitrification	nitrate denitrification	aerobic chemoheterotrophy	nitrate denitrification	iron respiration	aerobic chemoheterotrophy
aerobic chemoheterotrophy	nitrogen fixation	chitinolysis	aerobic ammonia oxidation	fermentation	aerobic ammonia oxidation
nitrogen fixation	iron respiration	aerobic ammonia oxidation	aerobic chemoheterotrophy	nitrate denitrification	nitrate denitrification
aerobic ammonia oxidation	methanotrophy	photosynthetic cyanobacteria	aerobic nitrite oxidation	nitrogen fixation	methanol oxidation
xylanolysis	methanogenesis	manganese oxidation	chitinolysis	aerobic chemoheterotrophy	xylanolysis
chitinolysis	fermentation	nitrogen fixation	nitrogen fixation	methanotrophy	chitinolysis
aerobic nitrite oxidation	methanol oxidation	photoheterotrophy	iron respiration	methanogenesis	ureolysis
invertebrate parasites	methanogenesis	xylanolysis	manganese oxidation	animal parasites/symbionts	manganese oxidation

For each cluster, Figure 7. showed the relationship between pH and more abundant microbial functions. Under specific circumstances, there were some functions that

had clearer relationships with pH value. For example, under “Short-vegetation” and “Disturbed” environments, aerobic ammonia oxidation function was richer in alkaline soil (i.e. higher pH). The reason may be that this function requires the consumption of ammonia, and ammonia can have a higher concentration in an alkaline environment. However, under the “Dry” environment, the trend was not observed. One of the reasons may be that ammonia will volatize, and thus cancel out its higher concentration in the alkaline soil.

While ammonia oxidation consumes ammonia, nitrate denitrification assists the formulation of ammonia. It follows that under “Disturbed” environment, we could observe a negative correlation between pH and nitrate denitrification abundance. Aerobic chemoheterotrophy on the other hand was commonly more abundant with the increase of pH, and the relationship could be seen in “Disturbed”, “Cold dry”, and “Cold moist” environments. Overall, under the “Disturbed” environment, pH could be a key driver for multiple functions.

Figure 7. pH-function abundance by cluster

Conclusion

Different environments and soil pH can influence microbial community functions. This is important because microbial functions can affect ecosystem dynamics, including plant productivity and nutrient cycling. Any change in the environment or soil pH may affect microbial diversity or functionality and lead to changes in ecosystem processes. By understanding microbial diversity and functional changes in different environments and soil pH, we can better understand changes in nutrient cycling like nitrogen, phosphorus and carbon that could be influencing global climate changes

Reference

[1] Thompson, L. et al. “A communal catalogue reveals Earth’s multiscale microbial diversity.” *Nature* 551 (2017).

[2] Louca, S. et al. “Decoupling function and taxonomy in the global ocean microbiome.” *Science* 353 (2016)

[3] Rohe, K., and Zeng, M. “Vintage Factor Analysis with Varimax Performs Statistical Inference.” *arXiv: Methodology* (2020).