This appendix present in greater details the datasets used in this study and the additional results

Figure S1: Datasets used in this study

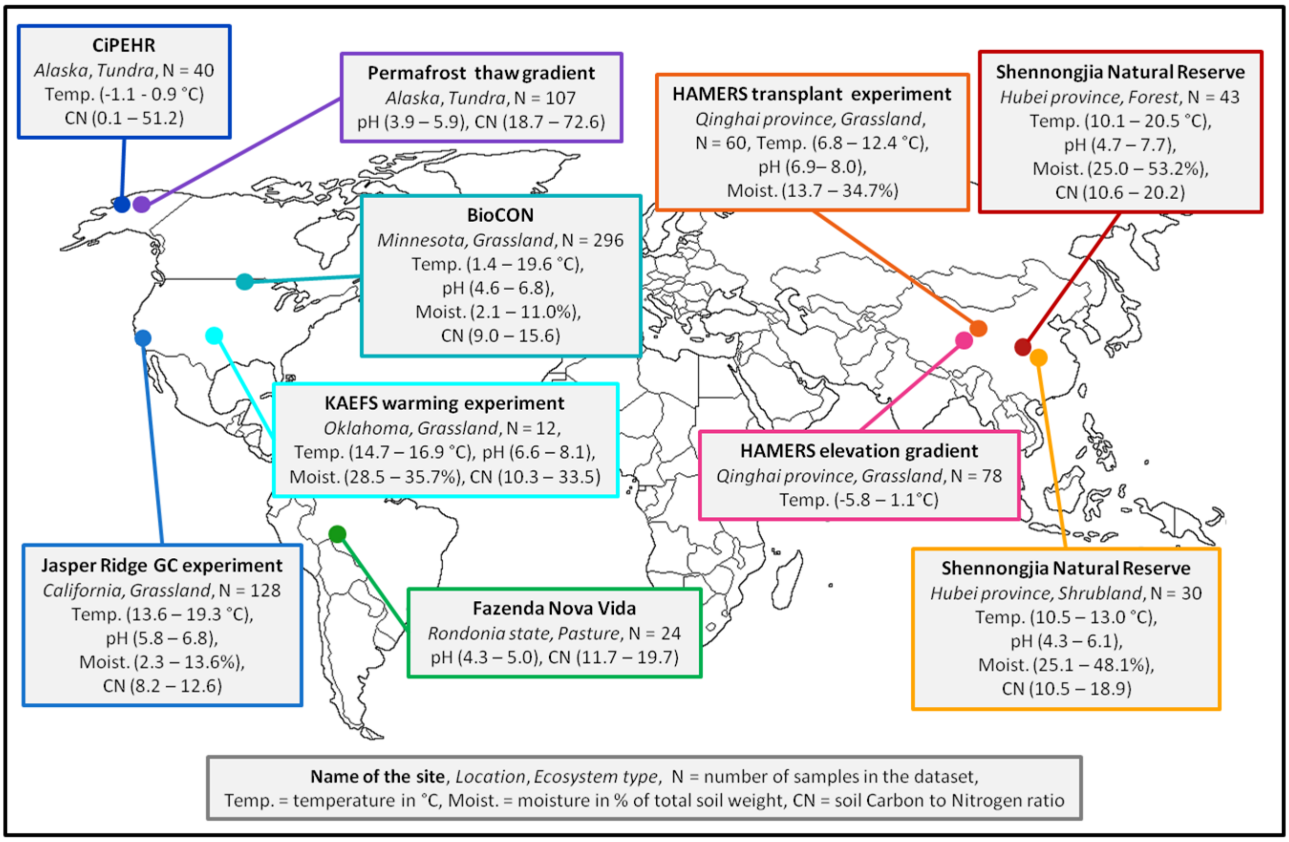

CiPEHR
Alaska, Tundra, N = 40
Temp. (-1.1 - 0.9 °C)
CN (0.1 – 51.2)
Permafrost thaw gradient
Alaska, Tundra, N = 107
pH (3.9 – 5.9), CN (18.7 – 72.6)
HAMERS transplant experiment
Qinghai province, Grassland,
N = 60, Temp. (6.8 – 12.4 °C),
pH (6.9– 8.0),
Moist. (13.7 – 34.7%)
Shennongjia Natural Reserve
Hubei province, Forest, N = 43
Temp. (10.1 – 20.5 °C),
pH (4.7 – 7.7),
Moist. (25.0 – 53.2%),
CN (10.6 – 20.2)
BioCON
Minnesota, Grassland, N = 296
Temp. (1.4 – 19.6 °C),
pH (4.6 – 6.8),
Moist. (2.1 – 11.0%),
CN (9.0 – 15.6)
KAEFS warming experiment
Oklahoma, Grassland, N = 12,
Temp. (14.7 – 16.9 °C), pH (6.6 – 8.1),
Moist. (28.5 – 35.7%), CN (10.3 – 33.5)
HAMERS elevation gradient
Qinghai province, Grassland, N = 78
Temp. (-5.8 – 1.1°C)
Jasper Ridge GC experiment
California, Grassland, N = 128
Temp. (13.6 – 19.3 °C),
pH (5.8 – 6.8),
Moist. (2.3 – 13.6%),
CN (8.2 – 12.6)
Fazenda Nova Vida
Rondonia state, Pasture, N = 24
pH (4.3 – 5.0), CN (11.7 – 19.7)
Shennongjia Natural Reserve
Hubei province, Shrubland, N = 30
Temp. (10.5 – 13.0 °C),
pH (4.3 – 6.1),
Moist. (25.1 – 48.1%),
CN (10.5 – 18.9)
Name of the site, *Location, Ecosystem type*, N = number of samples in the dataset,
Temp. = temperature in °C, Moist. = moisture in % of total soil weight, CN = soil Carbon to Nitrogen ratio

Table S1: Description of the ten data sets composing the database

Site name	Short name	# samples	Ecosystem	Continent	Country	Location	Latitude	Longitude	Origin of the dataset
Shennongjia Natural Reserve-Forest	SNNRf	43	Forest	Asia	China	Hubei province	31.456	110.343	Shennongjia National Natural Reserve
Shennongjia Natural Reserve-Shrubland	SNNRs	30	Shrubland	Asia	China	Hubei province	31.459	110.270	Shennongjia National Natural Reserve
HAMERS elevation gradient	CE	78	Grassland	Asia	China	Qinghai province	35.202	99.089	Haibei Alpine Meadow Ecosystem Research Station (HAMERS)
HAMERS transplant experiment	TB	60	Grassland	Asia	China	Qinghai province	37.617	101.317	Haibei Alpine Meadow Ecosystem Research Station (HAMERS)
CiPHER	AKC	40	Tundra	North America	USA	Alaska	63.883	-149.226	Carbon in Permafrost Experimental Heating Project (CiPEHR)
Permafrost thaw gradient	AKG	107	Tundra	North America	USA	Alaska	63.883	-149.226	Permafrost Thaw Gradient
Jasper Ridge GC experiment	JRGCE	128	Grassland	North America	USA	California	37.403	-122.242	Jasper Ridge Global Change Experiment (JRGCE)
BioCON	BC	296	Grassland	North America	USA	Minnesota	45.403	-93.189	Biodiversity, CO2 and Nitrogen experiment (BioCON)
KAEFS warming experiment	OK	12	Grassland	North America	USA	Oklahoma	34.983	-97.517	Kessler Atmospheric and Ecological Field Station (KAEFS)
Fazenda Nova Vida	AM	24	Pasture	South America	Brazil	Rondonia state	-10.164	-62.783	Fazenda Nova Vida

Table S2: Distribution of functional genes varants in the various levels of resolution used in this study

Broad category	Gene family	Genes	Variants
Antibiotic resistance	3	10	1908
Carbon cycling	4	54	8871
Energy process	1	4	635
Metal resistance	15	38	5641
Nitrogen cycling	6	17	4934
Phosphorus cycling	1	3	1127
Stress	12	46	11944
Sulphur cycling	4	10	2671
Virulence	11	12	1950

Rationale of the analytical framework

The “importance” of a function in a bin was quantified using the weight index, which was defined in order to take into account the particularities of FGA data. Indeed, FGA are so called closed format metagenomic approach (Zhou et al. 2015), which has two main consequences. First, that is the information retrieved from the analysis of a microbial community DNA is predetermined by the FGA design, in other terms we won’t be able to detect a gene for which there was not a probe on the chip. Second, that the sampling effort is not uniformly distributed across microbial functions, in other terms certain genes or genes families are represented by a higher/lower number of probes on the FGA. These two aspects will influence the outcome of data analysis. For instance, if we sample randomly a thousand variants (which correspond to probes on the FGA), count the number of variant for each gene they encode and divide by the total number of variants (i.e. 1000), then the obtained genes proportions will be determined by the number of variant from each genes on the FGA. Doing so we will conclude that the gene with the highest number of variants was the more important in the community. Hence, the “importance” of each gene in this random sample corresponds to the proportions of variants it represent on the FGA design and can be considered as a null model. In our analytical framework (Figure 1), variant bins were defined according to variants occurrence, abundance or a combination of both. Nonetheless, we can expect that genes with a higher number of variants are more likely to be found in any bin. Consequently, we divided the observed proportion of the summed signal intensity of variants within a bin corresponding to each gene ($\text{weight}_{\text{observed}}$) by the proportion expected according to the FGA design ($\text{weight}_{\text{expected}}$), which provided us with the normalized weight.

Justification of the choice of bins number

The choice of the number of bins resulted from sensibility analyses performed prior to data analyses and has a concrete justification that is related with the way we estimated the “importance” of functions (*i.e.* weight in Fig. 1) along the abundance and occupancy gradients (i.e. within the different bins). Indeed, and as described above, the weight of a function within a bin is based on the estimation of the proportion of the total hybridization signal in the bin that corresponds to gene variants from this particular function. If we used a higher number of bins (e.g. 10), then it is more likely that some bins will not contain variants from some functions and it would then be impossible to estimate the weight of such functions within these bins. On the contrary, if we used few bins (e.g. 3) we would be able to estimate function weight in all the bins but the rarity to commonness gradient represented by these bins would not be very informative. Hence, the choice of 6 bins appeared as an intermediate compromise between our capacity to describe variants distribution along a gradient of abundance/occupancy and our capacity to successfully estimate function importance along this gradient.

Table S3. MOS test of bimodality for the occupancy-frequency distribution within each site

To test the bimodality of the FOD we used the MOS test which determines (i) whether there is a local maximum frequency at low occupancy (*i.e.* 0), (ii) whether there is a local maximum frequency at high occupancy (*i.e.* 1) and (iii) whether the relationship is bimodal.

Our results indicated that the occupancy-frequency distribution (OFD) is bimodal in all the sites, with two maxima, one at low and one at high occupancy; Additionally, we observed a stronger right mode in all but one site (higher F values for the right mode, "Max at 1").

Site name	**Test**	***F value***	***p value***	
CiPEHR	Max at 0	72	0.000	***
	Max at 1	57	0.000	***
	Bimodality	NA	0.000	***
Permafrost thaw gradient	Max at 0	202	0.000	***
	Max at 1	212	0.000	***
	Bimodality	NA	0.000	***
BioCON	Max at 0	135	0.000	***
	Max at 1	185	0.000	***
	Bimodality	NA	0.000	***
Jasper Ridge GC experiment	Max at 0	78	0.000	***
	Max at 1	127	0.000	***
	Bimodality	NA	0.000	***
KAEFS warming experiment	Max at 0	14	0.004	***
	Max at 1	32	0.000	***
	Bimodality	NA	0.005	***
Fazenda Nova Vida	Max at 0	30	0.000	***
	Max at 1	99	0.000	***
	Bimodality	NA	0.000	***
HAMERS elevation gradient	Max at 0	57	0.000	***
	Max at 1	75	0.000	***
	Bimodality	NA	0.000	***
Shennongjia Natural Reserve-Forest	Max at 0	80	0.000	***
	Max at 1	121	0.000	***
	Bimodality	NA	0.000	***
Shennongjia Natural Reserve-Shrubland	Max at 0	25	0.000	***
	Max at 1	46	0.000	***
	Bimodality	NA	0.000	***
HAMERS transplant experiment	Max at 0	46	0.000	***
	Max at 1	73	0.000	***
	Bimodality	NA	0.000	***

Figure S2. Dissimilarity in the composition of abundance bins within each site.

This figure represents a detrended correspondence analysis (DCA) of abundance bins within each studied site. Differences in the composition between bins were estimated using Bray-Curtis dissimilarity on the matrices depicting the weight of gene families in each bins from each sample. The size of the bubbles corresponds to the rank of the bins (small for B_1 corresponding to genes with low abundance and big for B_6 corresponding to genes with high abundance). The centroid of each bin level is depicted with black letters (1 to 6).

Our results indicated that bins of the same rank from different samples contain variants that correspond to the same gene families.

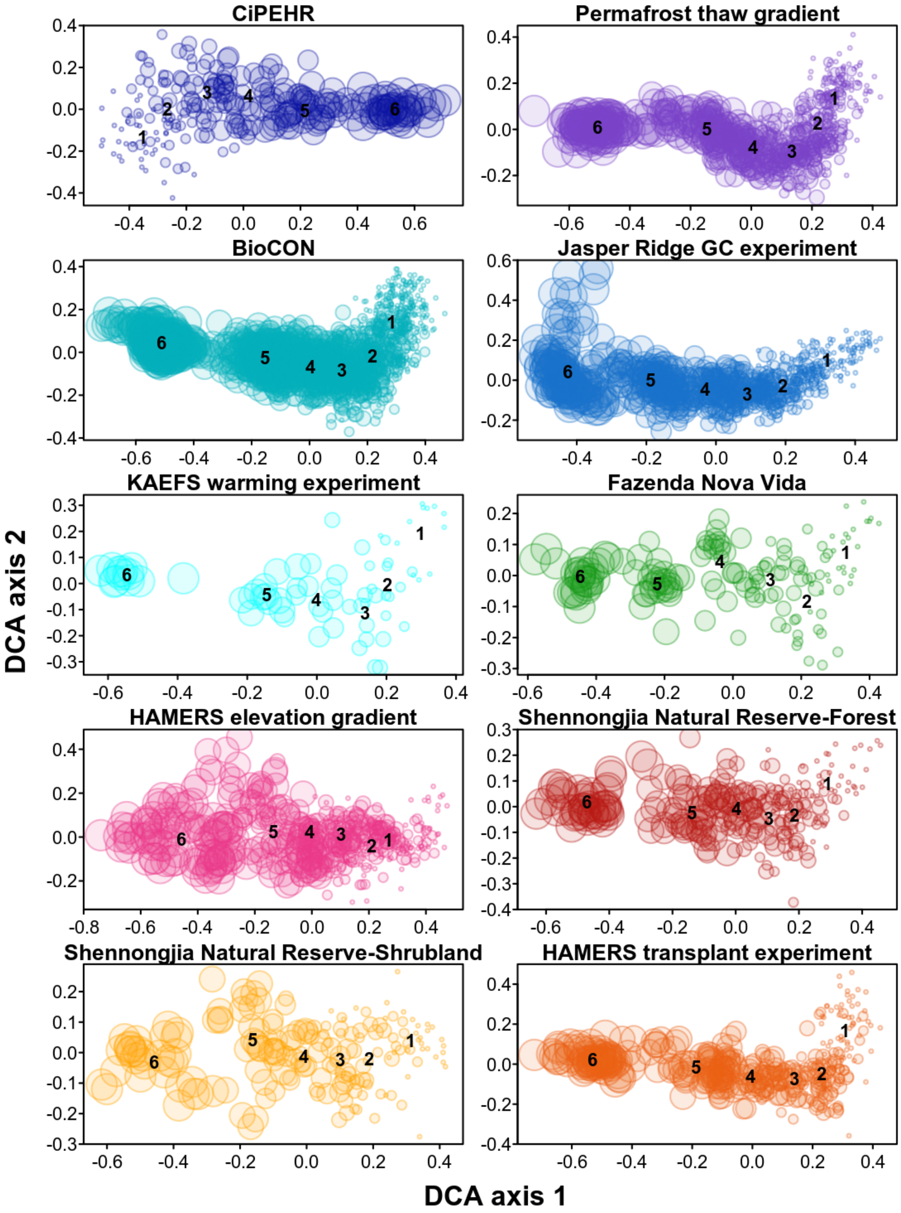


Table S4: Test of composition dissimilarity between occupancy bins across soil ecosystems.

Pairwise differences between bins were tested using pairwise PERMANOVA.

Our results showed that (i) the sixth bins (B6) has the most different composition among all occupancy bins. In other terms ubiquitous variants carry a significantly different set of traits than variants with a narrower spatial distribution; (ii) this was confirmed when associating variants to different levels of resolution (genes, gene families, and broad categories).

	Broad categories				**Gene families**				**Genes**			
pairs	**F value**	**R^2**	***p value***		**F value**	**R^2**	***p value***		**F value**	**R^2**	***p value***	
B1 vs B2	1.3	0.1	0.312		2.7	0.1	0.016	*	3.0	0.1	0.001	***
B1 vs B3	4.8	0.2	0.011	*	6.0	0.3	0.002	**	4.0	0.2	0.001	***
B1 vs B4	8.0	0.3	0.002	**	7.7	0.3	0.002	**	5.0	0.2	0.001	***
B1 vs B5	6.9	0.3	0.002	**	12.7	0.4	0.002	**	7.1	0.3	0.001	***
B1 vs B6	28.4	0.6	0.002	**	35.1	0.7	0.002	**	22.1	0.6	0.001	***
B2 vs B3	1.4	0.1	0.285		1.9	0.1	0.040	*	1.6	0.1	0.009	**
B2 vs B4	4.6	0.2	0.005	**	2.8	0.1	0.005	**	2.0	0.1	0.003	**
B2 vs B5	5.5	0.2	0.004	**	5.5	0.2	0.002	**	3.5	0.2	0.001	***
B2 vs B6	35.0	0.7	0.002	**	20.6	0.5	0.002	**	12.5	0.4	0.001	***
B3 vs B4	1.0	0.1	0.495		1.4	0.1	0.128		1.4	0.1	0.035	*
B3 vs B5	2.0	0.1	0.134		2.5	0.1	0.005	**	2.0	0.1	0.001	***
B3 vs B6	31.9	0.6	0.002	**	13.2	0.4	0.002	**	9.1	0.3	0.001	***
B4 vs B5	0.9	0.0	0.495		2.0	0.1	0.018	*	1.6	0.1	0.006	**
B4 vs B6	34.3	0.7	0.002	**	11.8	0.4	0.002	**	7.3	0.3	0.001	***
B5 vs B6	24.1	0.6	0.002	**	8.2	0.3	0.002	**	6.1	0.3	0.001	***

Figure S3-S4-S5. Linear relationships between occupancy-abundance bins and genes weight.

For each of the 194 genes, we estimated their weight within occupancy-abundance bins. The first panel plot corresponds to gene families for which the relationship was not significant. The second panel corresponds to genes with negative relationships and the third panel corresponds to genes with positive relationships. Genes are ranked according to the absolute value of the slope of the relationship. The name of the gene is depicted on top of each plot, with a color code that relates to the broad category it belongs to (purple: antibiotic resistance, red: carbon cycling, yellow: metal resistance, cyan: stress, dark blue: sulphur cycle, light green: nitrogen cycle, dark green: phosphorus cycle and black: virulence).

Figure S3. Non significant relationships between occupancy-abundance bins and genes weight

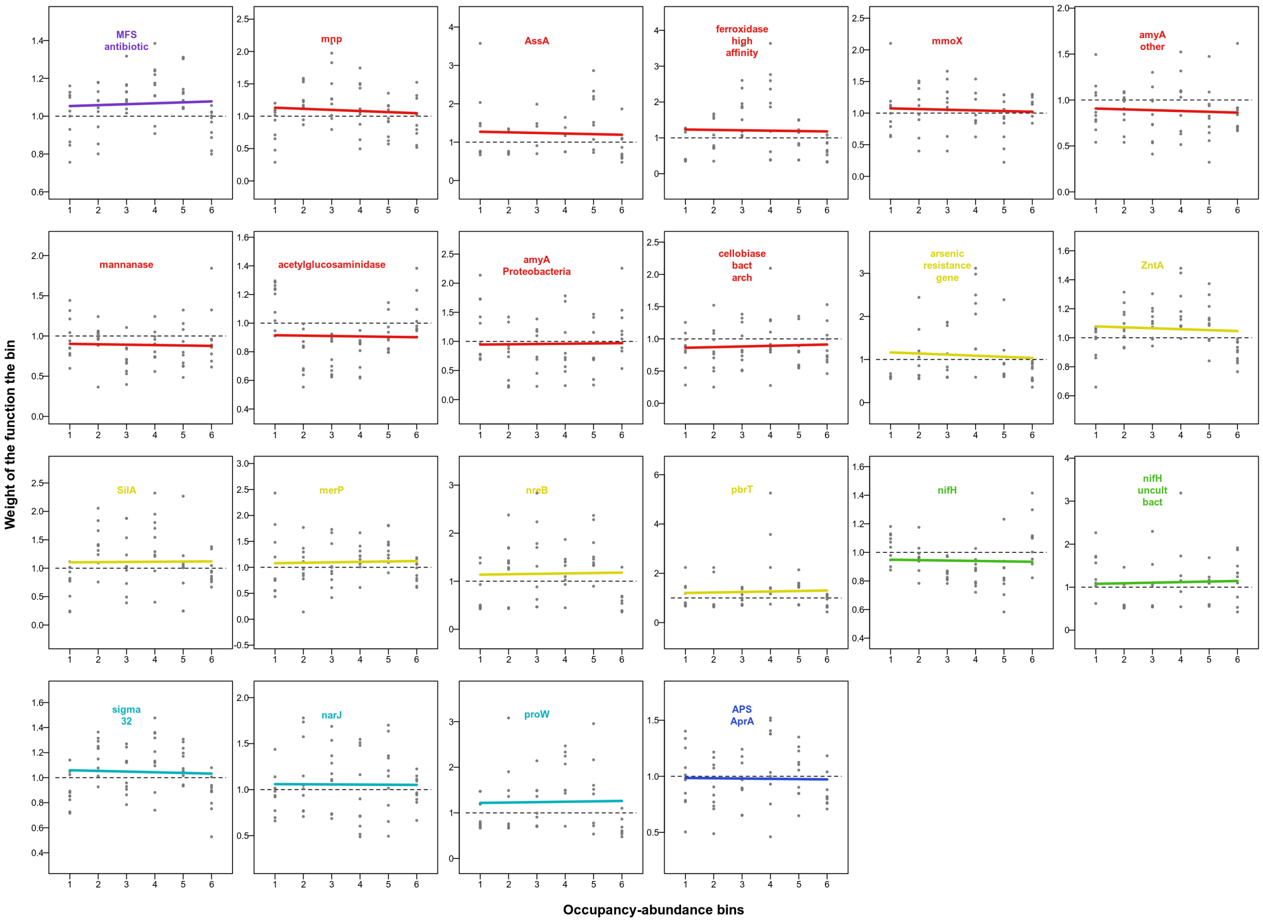


Figure S4. Negative relationships between occupancy-abundance bins and genes weight

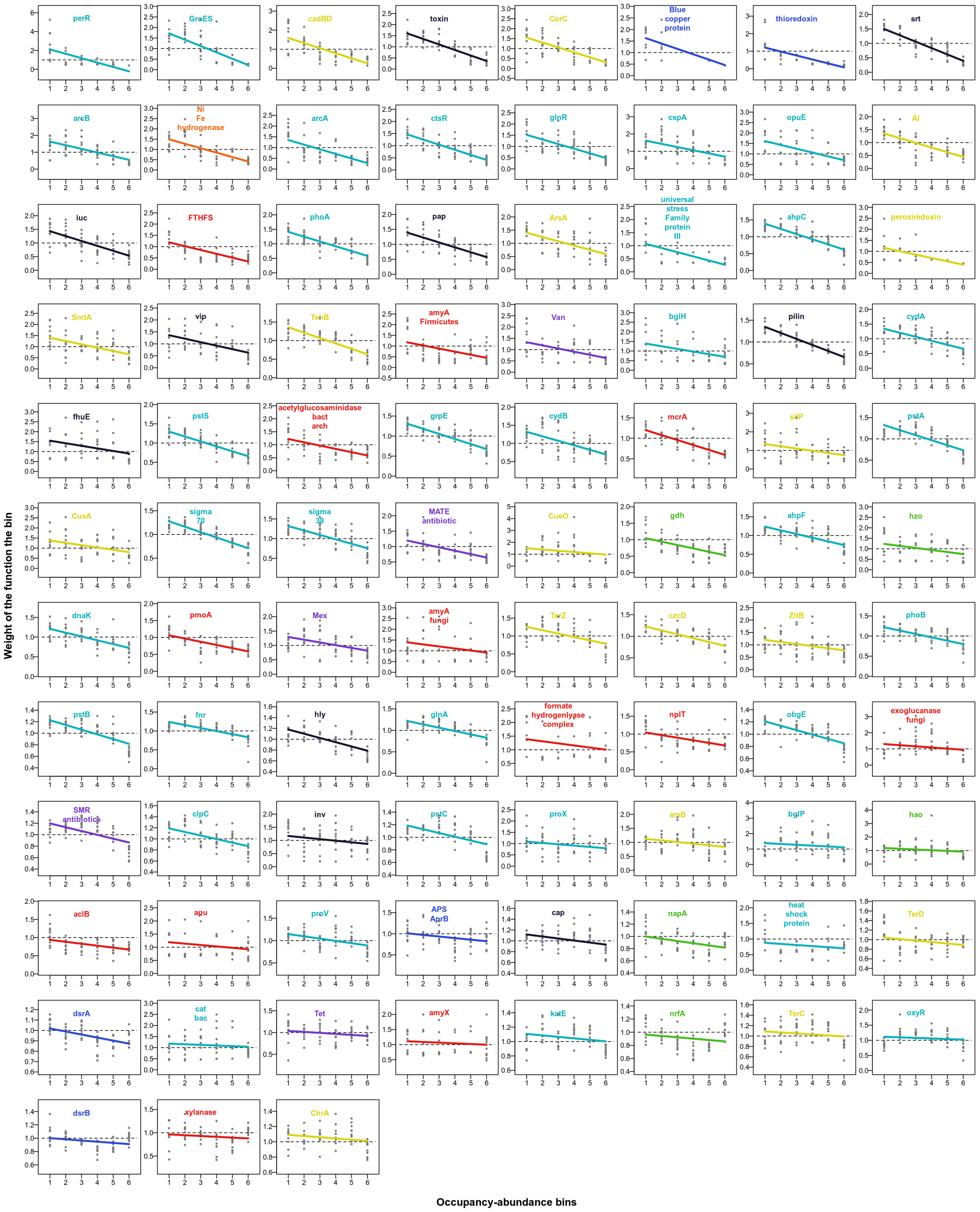


Figure S5. Positive relationships between occupancy-abundance bins and genes weight

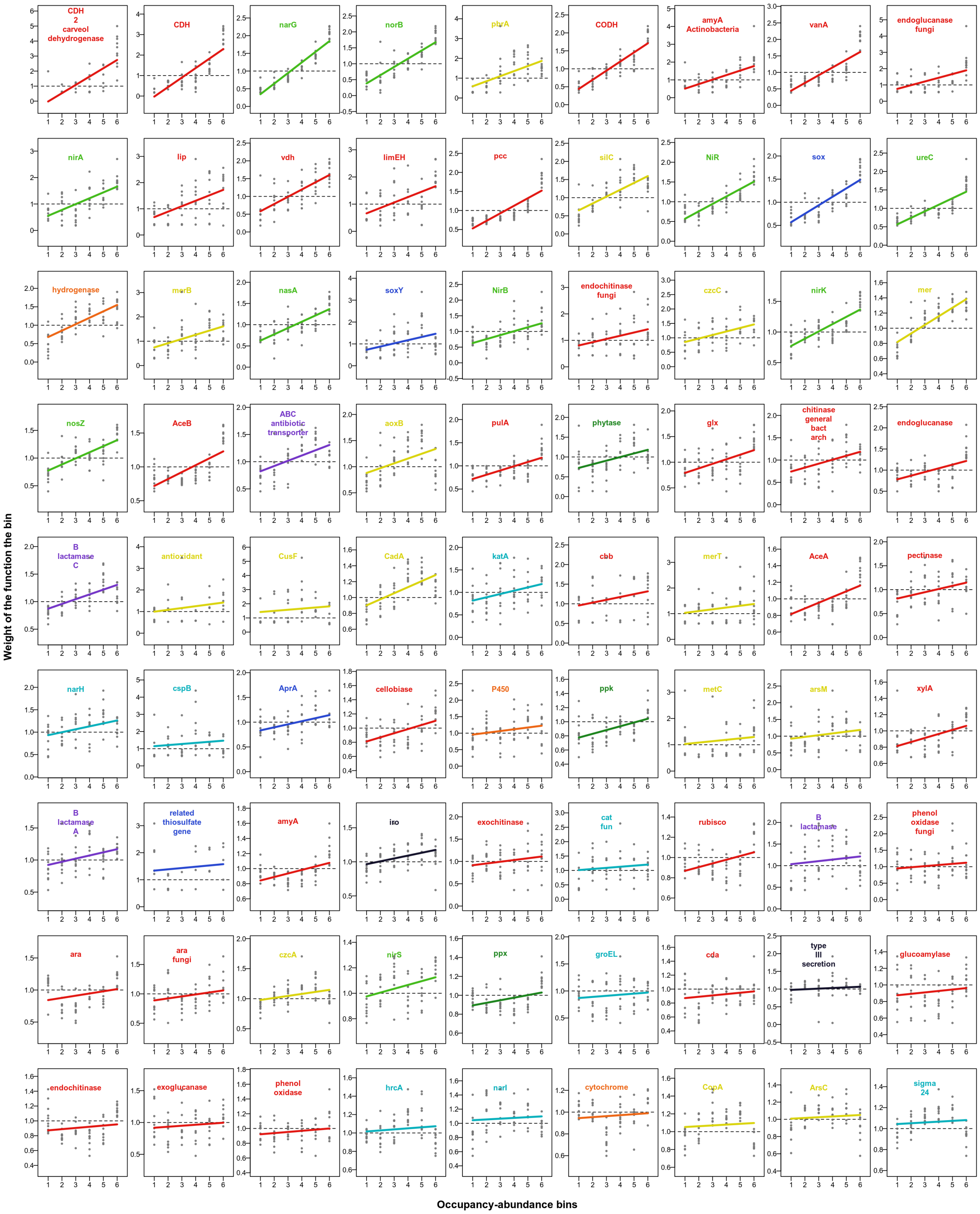


Table S5. Result of the linear models describing the relationship between the rank of occupancy-abundance bins (1 to 6) and the weight of gene within the bins

Rows colored in red correspond to the 20 models with the lowest and higher slopes.

Broad category	Gene family	Gene	Model slope	p-value of model	R2	Type of relation
Stress	*oxygen_stress*	*perR*	*-0.465*	*0*	*0.402*	*Negative*
Stress	*heat_shock*	*GroES*	*-0.301*	*0*	*0.506*	*Negative*
Metal resistance	*cadmium*	*cadBD*	*-0.266*	*0*	*0.543*	*Negative*
Virulence	*toxin*	*toxin*	*-0.248*	*0*	*0.675*	*Negative*
Metal resistance	*cobalt*	*CorC*	*-0.247*	*0*	*0.535*	*Negative*
Sulphur cycling	*various_sulphur*	*Blue_copper_protein*	*-0.233*	*0*	*0.38*	*Negative*
Sulphur cycling	*various_sulphur*	*thioredoxin*	*-0.226*	*0*	*0.435*	*Negative*
Virulence	*surface_protein*	*srt*	*-0.221*	*0*	*0.768*	*Negative*
Stress	*oxygen_limitation*	*arcB*	*-0.216*	*0*	*0.343*	*Negative*
Energy process	*energy_process*	*Ni_Fe_hydrogenase*	*-0.215*	*0*	*0.48*	*Negative*
Stress	*oxygen_limitation*	*arcA*	*-0.215*	*0*	*0.436*	*Negative*
Stress	*protein_stress*	*ctsR*	*-0.21*	*0*	*0.633*	*Negative*
Stress	*nitrogen_limitation*	*glnR*	*-0.206*	*0*	*0.477*	*Negative*
Stress	*cold_shock*	*cspA*	*-0.184*	*0*	*0.242*	*Negative*
Stress	*osmotic_stress*	*opuE*	*-0.183*	*0*	*0.296*	*Negative*
Metal resistance	*aluminum*	*Al*	*-0.18*	*0*	*0.522*	*Negative*
Virulence	*aerobactin*	*iuc*	*-0.18*	*0*	*0.507*	*Negative*
Carbon cycling	*acetogenesis*	*FTHFS*	*-0.17*	*0*	*0.458*	*Negative*
Stress	*phosphate_limitation*	*phoA*	*-0.166*	*0*	*0.622*	*Negative*
Virulence	*adhesin*	*pap*	*-0.163*	*0*	*0.525*	*Negative*
Metal resistance	arsenic	ArsA	-0.162	0	0.369	Negative
Stress	various_stress	universal_stress_Family_protein_III	-0.16	0	0.347	Negative
Stress	oxygen_stress	ahpC	-0.153	0	0.734	Negative
Metal resistance	various_metal	peroxiredoxin	-0.153	0	0.206	Negative
Metal resistance	others	SmtA	-0.149	0	0.261	Negative
Virulence	virulence_protein	vip	-0.146	0	0.251	Negative
Metal resistance	tellurium	TehB	-0.145	0	0.606	Negative
Carbon cycling	carbon_degradation	amyA_Firmicutes	-0.144	0	0.195	Negative
Antibiotic resistance	others	Van	-0.14	0	0.211	Negative
Stress	glucose_limitation	bglH	-0.139	0	0.168	Negative
Virulence	pilin	pilin	-0.138	0	0.805	Negative
Stress	oxygen_limitation	cydA	-0.136	0	0.42	Negative
Virulence	iron_oxidation	fhuE	-0.13	0	0.108	Negative
Stress	phosphate_limitation	pstS	-0.128	0	0.662	Negative
Carbon cycling	carbon_degradation	acetylglucosaminidase_bact_arch	-0.127	0	0.301	Negative
Stress	heat_shock	grpE	-0.127	0	0.69	Negative
Stress	oxygen_limitation	cydB	-0.124	0	0.368	Negative
Carbon cycling	methane	mcrA	-0.12	0	0.605	Negative
Metal resistance	silver	silP	-0.119	0	0.141	Negative
Stress	phosphate_limitation	pstA	-0.118	0	0.564	Negative
Metal resistance	copper	CusA	-0.115	0	0.153	Negative
Stress	sigma_factors	sigma_70	-0.113	0	0.761	Negative
Stress	sigma_factors	sigma_38	-0.113	0	0.489	Negative
Antibiotic resistance	transporter	MATE_antibiotic	-0.11	0	0.348	Negative
Metal resistance	copper	CueO	-0.107	0	0.045	Negative
Nitrogen cycling	ammonification	gdh	-0.103	0	0.356	Negative
Stress	oxygen_stress	ahpF	-0.098	0	0.431	Negative
Nitrogen cycling	anammox	hzo	-0.097	0	0.111	Negative
Stress	heat_shock	dnaK	-0.097	0	0.404	Negative
Carbon cycling	methane	pmoA	-0.095	0	0.399	Negative
Antibiotic resistance	transporter	Mex	-0.095	0	0.179	Negative

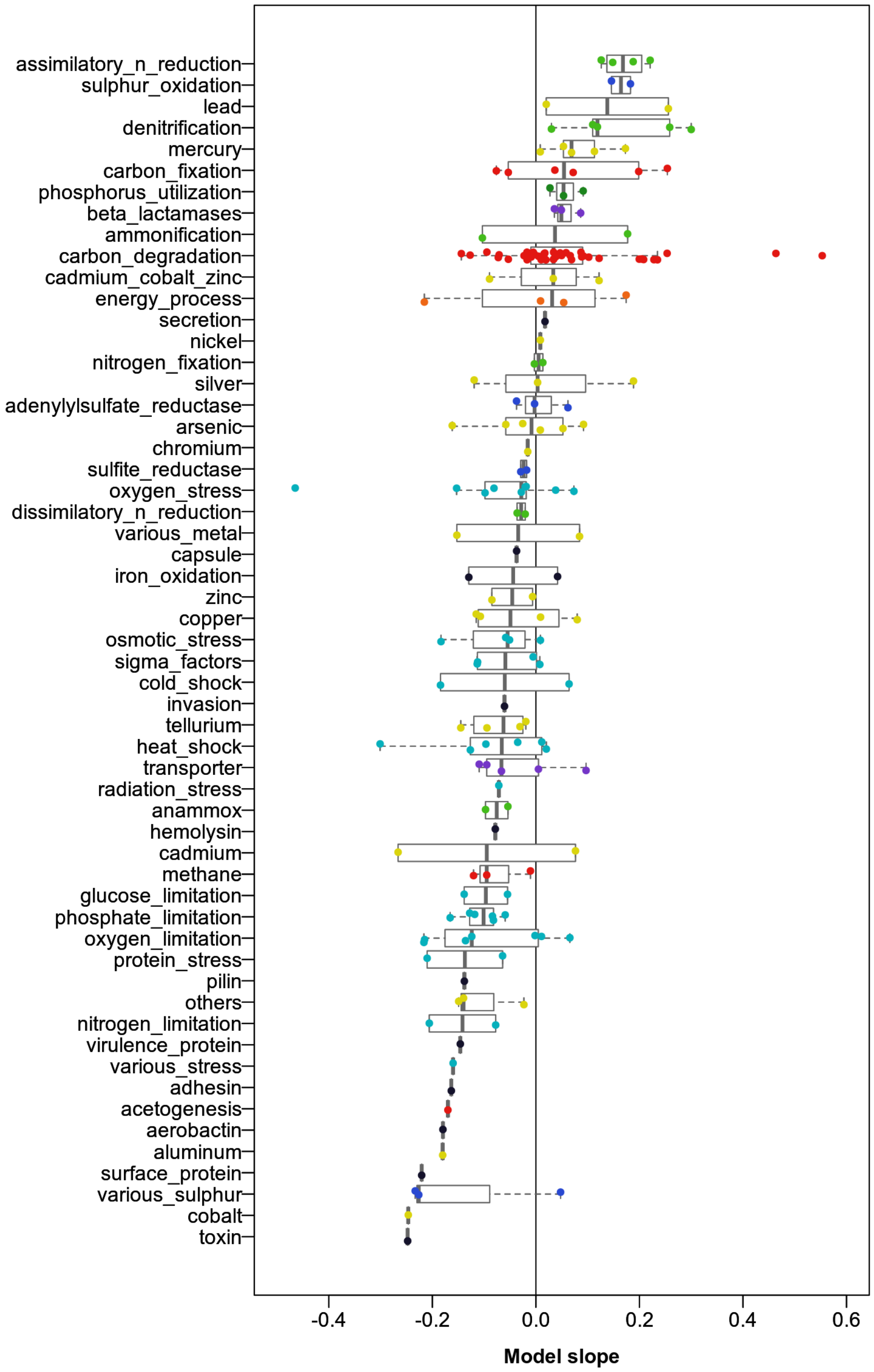
Broad category	Gene family	Gene	Model slope	p-value of model	R2	Type of relation
Carbon cycling	carbon_degradation	amyA_fungi	-0.095	0	0.064	Negative
Metal resistance	tellurium	TerZ	-0.095	0	0.301	Negative
Metal resistance	cadmium_cobalt_zinc	czcD	-0.09	0	0.529	Negative
Metal resistance	zinc	ZitB	-0.085	0	0.121	Negative
Stress	phosphate_limitation	phoB	-0.084	0	0.445	Negative
Stress	phosphate_limitation	pstB	-0.082	0	0.504	Negative
Stress	oxygen_stress	fnr	-0.081	0	0.47	Negative
Virulence	hemolysin	hly	-0.078	0	0.452	Negative
Stress	nitrogen_limitation	glnA	-0.078	0	0.459	Negative
Carbon cycling	carbon_fixation	formate_hydrogenlyase_complex	-0.076	0	0.055	Negative
Carbon cycling	carbon_degradation	nplT	-0.073	0	0.206	Negative
Stress	radiation_stress	obgE	-0.072	0	0.529	Negative
Carbon cycling	carbon_degradation	exoglucanase_fungi	-0.071	0	0.038	Negative
Antibiotic resistance	transporter	SMR_antibiotics	-0.067	0	0.35	Negative
Stress	protein_stress	clpC	-0.064	0	0.317	Negative
Virulence	invasion	inv	-0.061	0	0.051	Negative
Stress	phosphate_limitation	pstC	-0.059	0	0.297	Negative
Stress	osmotic_stress	proX	-0.058	0	0.056	Negative
Metal resistance	arsenic	arsB	-0.058	0	0.081	Negative
Stress	glucose_limitation	bglP	-0.055	0.001	0.018	Negative
Nitrogen cycling	anammox	hao	-0.054	0	0.032	Negative
Carbon cycling	carbon_fixation	aclB	-0.053	0	0.172	Negative
Carbon cycling	carbon_degradation	apu	-0.053	0	0.039	Negative
Stress	osmotic_stress	proV	-0.051	0	0.159	Negative
Sulphur cycling	adenylylsulfate_reductase	APS_AprB	-0.037	0	0.071	Negative
Virulence	capsule	cap	-0.037	0	0.12	Negative
Nitrogen cycling	dissimilatory_n_reduction	napA	-0.036	0	0.142	Negative
Stress	heat_shock	heat_shock_protein	-0.035	0	0.032	Negative
Metal resistance	tellurium	TerD	-0.031	0	0.059	Negative
Sulphur cycling	sulfite_reductase	dsrA	-0.029	0	0.269	Negative
Stress	oxygen_stress	cat_bac	-0.028	0.021	0.009	Negative
Metal resistance	arsenic	arsenic_resistance_gene	-0.025	0.184	0.004	No relation
Antibiotic resistance	others	Tet	-0.023	0	0.042	Negative
Carbon cycling	carbon_degradation	amyX	-0.023	0.028	0.011	Negative
Stress	oxygen_stress	katE	-0.021	0	0.058	Negative
Nitrogen cycling	dissimilatory_n_reduction	nrfA	-0.021	0	0.04	Negative
Metal resistance	tellurium	TerC	-0.019	0	0.03	Negative
Stress	oxygen_stress	oxyR	-0.019	0.001	0.017	Negative
Sulphur cycling	sulfite_reductase	dsrB	-0.018	0	0.074	Negative
Carbon cycling	carbon_degradation	xylanase	-0.017	0.001	0.018	Negative
Carbon cycling	carbon_degradation	mnp	-0.017	0.057	0.006	No relation
Carbon cycling	carbon_degradation	AssA	-0.016	0.351	0.002	No relation
Metal resistance	chromium	ChrA	-0.016	0	0.044	Negative
Carbon cycling	carbon_degradation	ferroxidase_high_affinity	-0.011	0.601	0	No relation
Carbon cycling	methane	mmoX	-0.011	0.197	0.003	No relation
Carbon cycling	carbon_degradation	amyA_other	-0.009	0.174	0.003	No relation
Metal resistance	zinc	ZntA	-0.006	0.09	0.005	No relation
Stress	sigma_factors	sigma_32	-0.005	0.258	0.002	No relation
Carbon cycling	carbon_degradation	mannanase	-0.005	0.427	0.001	No relation
Nitrogen cycling	nitrogen_fixation	nifH	-0.003	0.422	0.001	No relation

Broad category	Gene family	Gene	Model slope	p-value of model	R2	Type of relation
Carbon cycling	carbon_degradation	acetylglucosaminidase	-0.003	0.542	0.001	No relation
Sulphur cycling	adenylylsulfate_reductase	APS_AprA	-0.003	0.671	0	No relation
Stress	oxygen_limitation	narJ	-0.002	0.833	0	No relation
Metal resistance	silver	SilA	0.003	0.777	0	No relation
Carbon cycling	carbon_degradation	amyA_Proteobacteria	0.005	0.674	0	No relation
Antibiotic resistance	transporter	MFS_antibiotic	0.005	0.136	0.004	No relation
Stress	sigma_factors	sigma_24	0.008	0.008	0.012	Positive
Metal resistance	mercury	merP	0.008	0.399	0.001	No relation
Metal resistance	arsenic	ArsC	0.009	0.018	0.009	Positive
Metal resistance	nickel	nreB	0.009	0.585	0.001	No relation
Stress	osmotic_stress	proW	0.009	0.617	0.001	No relation
Metal resistance	copper	CopA	0.009	0.027	0.008	Positive
Energy process	energy_process	cytochrome	0.009	0.007	0.012	Positive
Carbon cycling	carbon_degradation	cellobiase_bact_arch	0.01	0.22	0.003	No relation
Stress	oxygen_limitation	narI	0.011	0.037	0.007	Positive
Stress	heat_shock	hrcA	0.011	0.003	0.015	Positive
Nitrogen cycling	nitrogen_fixation	nifH_uncult_bact	0.013	0.366	0.002	No relation
Carbon cycling	carbon_degradation	phenol_oxidase	0.016	0	0.03	Positive
Carbon cycling	carbon_degradation	exoglucanase	0.016	0.007	0.012	Positive
Carbon cycling	carbon_degradation	endochitinase	0.016	0	0.023	Positive
Carbon cycling	carbon_degradation	glucoamylase	0.017	0	0.021	Positive
Virulence	secretion	type_III_secretion	0.018	0.02	0.011	Positive
Carbon cycling	carbon_degradation	cda	0.019	0	0.029	Positive
Metal resistance	lead	pbrT	0.02	0.331	0.002	No relation
Stress	heat_shock	groEL	0.02	0.002	0.017	Positive
Phosphorus cycling	phosphorus_utilization	ppx	0.027	0	0.123	Positive
Nitrogen cycling	denitrification	nirS	0.03	0	0.129	Positive
Metal resistance	cadmium_cobalt_zinc	czcA	0.034	0	0.096	Positive
Carbon cycling	carbon_degradation	ara_fungi	0.034	0	0.064	Positive
Carbon cycling	carbon_degradation	ara	0.034	0	0.072	Positive
Carbon cycling	carbon_degradation	phenol_oxidase_fungi	0.035	0	0.023	Positive
Antibiotic resistance	beta_lactamases	B_lactamase	0.036	0.001	0.018	Positive
Carbon cycling	carbon_fixation	rubisco	0.037	0	0.154	Positive
Stress	oxygen_stress	cat_fun	0.038	0.001	0.02	Positive
Carbon cycling	carbon_degradation	exochitinase	0.039	0	0.048	Positive
Virulence	iron_oxidation	iro	0.042	0	0.154	Positive
Carbon cycling	carbon_degradation	amyA	0.047	0	0.228	Positive
Sulphur cycling	various_sulphur	related_thiosulfate_gene	0.048	0.027	0.018	Positive
Antibiotic resistance	beta_lactamases	B_lactamase_A	0.049	0	0.133	Positive
Carbon cycling	carbon_degradation	xylA	0.05	0	0.217	Positive
Metal resistance	arsenic	arsM	0.052	0	0.05	Positive
Metal resistance	mercury	metC	0.053	0.001	0.023	Positive
Phosphorus cycling	phosphorus_utilization	ppk	0.053	0	0.235	Positive
Energy process	energy_process	P450	0.054	0	0.045	Positive
Carbon cycling	carbon_degradation	cellobiase	0.058	0	0.238	Positive
Sulphur cycling	adenylylsulfate_reductase	AprA	0.062	0	0.17	Positive
Stress	cold_shock	cspB	0.064	0.003	0.017	Positive
Stress	oxygen_limitation	narH	0.066	0	0.119	Positive
Carbon cycling	carbon_degradation	pectinase	0.066	0	0.125	Positive
Carbon cycling	carbon_degradation	AceA	0.069	0	0.363	Positive

Broad category	Gene family	Gene	Model slope	p-value of model	R2	Type of relation
Metal resistance	mercury	merT	0.069	0	0.039	Positive
Carbon cycling	carbon_fixation	cbb	0.072	0	0.08	Positive
Stress	oxygen_stress	katA	0.073	0	0.12	Positive
Metal resistance	cadmium	CadA	0.076	0	0.386	Positive
Metal resistance	copper	CusF	0.08	0.009	0.016	Positive
Metal resistance	various_metal	antioxidant	0.084	0	0.051	Positive
Antibiotic resistance	beta_lactamases	B_lactamase_C	0.086	0	0.341	Positive
Carbon cycling	carbon_degradation	endoglucanase	0.087	0	0.24	Positive
Carbon cycling	carbon_degradation	chitinase_general_bact_arch	0.089	0	0.203	Positive
Carbon cycling	carbon_degradation	glx	0.089	0	0.242	Positive
Phosphorus cycling	phosphorus_utilization	phytase	0.091	0	0.166	Positive
Carbon cycling	carbon_degradation	pulA	0.092	0	0.377	Positive
Metal resistance	arsenic	aoxB	0.092	0	0.194	Positive
Antibiotic resistance	transporter	ABC_antibiotic_transporter	0.097	0	0.295	Positive
Carbon cycling	carbon_degradation	AceB	0.102	0	0.483	Positive
Nitrogen cycling	denitrification	nosZ	0.11	0	0.467	Positive
Metal resistance	mercury	mer	0.113	0	0.678	Positive
Nitrogen cycling	denitrification	nirK	0.119	0	0.599	Positive
Metal resistance	cadmium_cobalt_zinc	czcC	0.122	0	0.207	Positive
Carbon cycling	carbon_degradation	endochitinase_fungi	0.122	0	0.132	Positive
Nitrogen cycling	assimilatory_n_reduction	NirB	0.126	0	0.295	Positive
Sulphur cycling	sulphur_oxidation	soxY	0.146	0	0.141	Positive
Nitrogen cycling	assimilatory_n_reduction	nasA	0.148	0	0.585	Positive
Metal resistance	*mercury*	*merB*	*0.173*	*0*	*0.253*	*Positive*
Energy process	*energy_process*	*hydrogenase*	*0.174*	*0*	*0.475*	*Positive*
Nitrogen cycling	*ammonification*	*ureC*	*0.177*	*0*	*0.637*	*Positive*
Sulphur cycling	*sulphur_oxidation*	*sox*	*0.183*	*0*	*0.668*	*Positive*
Nitrogen cycling	*assimilatory_n_reduction*	*NiR*	*0.188*	*0*	*0.689*	*Positive*
Metal resistance	*silver*	*silC*	*0.189*	*0*	*0.513*	*Positive*
Carbon cycling	*carbon_fixation*	*pcc*	*0.199*	*0*	*0.661*	*Positive*
Carbon cycling	*carbon_degradation*	*limEH*	*0.2*	*0*	*0.273*	*Positive*
Carbon cycling	*carbon_degradation*	*vdh*	*0.204*	*0*	*0.548*	*Positive*
Carbon cycling	*carbon_degradation*	*lip*	*0.208*	*0*	*0.261*	*Positive*
Nitrogen cycling	*assimilatory_n_reduction*	*nirA*	*0.221*	*0*	*0.438*	*Positive*
Carbon cycling	*carbon_degradation*	*endoglucanase_fungi*	*0.228*	*0*	*0.257*	*Positive*
Carbon cycling	*carbon_degradation*	*vanA*	*0.235*	*0*	*0.632*	*Positive*
Carbon cycling	*carbon_degradation*	*amyA_Actinobacteria*	*0.253*	*0*	*0.376*	*Positive*
Carbon cycling	*carbon_fixation*	*CODH*	*0.254*	*0*	*0.782*	*Positive*
Metal resistance	*lead*	*pbrA*	*0.256*	*0*	*0.351*	*Positive*
Nitrogen cycling	*denitrification*	*norB*	*0.259*	*0*	*0.694*	*Positive*
Nitrogen cycling	*denitrification*	*narG*	*0.3*	*0*	*0.873*	*Positive*
Carbon cycling	*carbon_degradation*	*CDH*	*0.464*	*0*	*0.676*	*Positive*
Carbon cycling	*carbon_degradation*	*CDH_2_carveol_dehydrogenase*	*0.553*	*0*	*0.487*	*Positive*

Figure S6. Slopes of the relationships between the weight of genes in a bin and the rank of the occupancy-abundance bin.

For each of the 194 gene, we fitted linear models explaining the weight of gene in eachoccupancy-abundance bin in function of the bin rank. Negative relationships (significant negative slopes) corresponded to gene over-represented in rare variants whereas positive ones (significant positive slopes) corresponded to gene over-represented in abundant variants. This figure is similar to the Figure 5B of the main text except that model slopes are classified by gene families and not broad categories.



Alaskan datasets: Permafrost thaw gradient and CiPEHR

The two Alaskan datasets represented tundra ecosystems and originated from the Eight Mile Lake study area in Alaska. They corresponded to the Carbon in Permafrost Experimental Heating Project (CiPEHR) and the Permafrost Thaw Gradient experiment.

Eight Mile Lake

https://www2.nau.edu/schuurlab-p/EightMileLake.html

The Eight Mile Lake study area is upland tundra located in the northern foothills of the Alaska range about 14 km west of Healy, Alaska (63º 52' 42"N, 149º15' 12"W). The site is situated on moist acidic tundra on a relatively well-drained gentle northeast-facing slope. The active layer (ground which thaws annually during the growing season) is ~ 50–80 cm thick and is situated above a perennially frozen permafrost layer. Mean monthly temperatures range from -16°C in December to +15°C in July, with a mean annual temperature (1976-2009) of -1.0°C. These soils hold substantial amounts of organic carbon in the top meter, ranging from 55 to 69 kg C m-2. Permafrost temperatures in this region are currently around -1°C and therefore susceptible to thaw. Vegetation at the site is dominated by the tussock-forming sedge, *Eriophorum vaginatum,* and deciduous shrub, *Vaccinium uliginosum*.

The Carbon in Permafrost Experimental Heating Project (CiPEHR)

https://www2.nau.edu/schuurlab-p/CiPEHR.html

The Carbon in Permafrost Experimental Heating Research (CiPEHR) project is an ecosystem warming experiment that was established in 2008 to test hypotheses about changes in the carbon cycle that are expected as a result of warming temperatures and permafrost thaw. The CiPEHR project uses snow fences coupled with spring snow removal to increase soil and permafrost temperatures and open-top chambers to increase growing season air temperatures.

The soil warming treatment, hereafter called winter warming, was achieved using six replicate snow fences (1.5 m tall × 8 m long) that trap insulating layers of snow. Each winter warming treatment and winter warming control area contains two summer warming plots and two summer warming control plots. Summer warming is achieved using 0.36 m2 × 0.5 m tall open-top chambers, which are set out during the snow-free period, between the first week in May and the last week of September.

The Permafrost Thaw Gradient

https://www2.nau.edu/schuurlab-p/Gradient.html

The thaw gradient contains three sites named *minimal*, *moderate*, and *extensive* thaw for the amount of vegetation change, active layer thickening, and thermokarst formation they have undergone due to different durations of permafrost thaw. At the extensive thaw site, permafrost thaw has been documented for the past two decades but likely began earlier. Extensive thaw has more shrubs than moderate and minimal thaw and has an undulating terrain with high, dry areas next to low, wet areas as a result of permafrost thaw. Next to moderate thaw is a 30 m deep borehole that has been used to measure permafrost temperatures since 1985. The goal of research at the permafrost thaw gradient has been to investigate the source and strength of feedbacks between permafrost warming and climate change. As permafrost soils thaw, their large carbon pools become vulnerable to microbial degradation, releasing CO2 into the atmosphere, which is a positive feedback to climate change.

References from these projects:

Mauritz, M, Bracho, R, Celis, G, Hutchings, J, Natali, SM, Pegoraro, E, Salmon, VG, Schädel, C, Webb, EE and Schuur, EAG (2017), Non-linear CO2 flux response to seven years of experimentally induced permafrost thaw. Glob Change Biol.doi:10.1111/gcb.13661

Xue K, M. Yuan M, J. Shi Z, Qin Y, Deng Y, Cheng L, Wu L, He Z, Van Nostrand JD, Bracho R, Natali S, Schuur EAG, Luo C, Konstantinidis KT, Wang Q, Cole JR, Tiedje JM, Luo Y, Zhou J (2016) Tundra soil carbon is vulnerable to rapid microbial decomposition under climate warming. Nature Clim. Change.

Bracho R, Natali S, Pegoraro E, Crummer KG, Schädel C, Celis G, Hale L, Wu L, Yin H, Tiedje JM, Konstantinidis KT, Luo Y, Zhou J, Schuur EAG (2016) Temperature sensitivity of organic matter decomposition of permafrost-region soils during laboratory incubations. Soil Biology and Biochemistry, 97, 1-14. doi:10.1016/j.soilbio.2016.02.008

Natali SM, Schuur EAG, Mauritz M, Schade JD, Celis G, Crummer KG, Johnston C, Krapek J, Pegoraro E, Salmon VG, Webb EE (2015) Permafrost thaw and soil moisture driving CO2 and CH4 release from upland tundra. Journal of Geophysical Research: Biogeosciences, 2014JG002872, doi:10.1002/2014JG002872

Grassland datasets

The five datasets representing grassland ecosystems originated from North America and China.

BioCON: Biodiversity, CO2, and Nitrogen: BC

https://cbs-cedarcreek.oit.umn.edu/research/experiments/e141
http://www.biocon.umn.edu/experiments/

BioCON is located at the Cedar Creek Ecoscience Reserve in east central Minnesota, USA about 50 km north of Minneapolis/St. Paul (Lat. 45N, Long. 93W). The site is located on a glacial outwash sandplain and production is nitrogen limited. The experiment was set up in a secondary successional old field after the existing vegetation was cleared. Plots were planted in 1997. BioCON consists of 371 2-meter x 2-meter plots, arranged into 6 circular areas or "rings" (20 meter diameter), each containing 61, 62, or 63 plots. Sixteen species of herbaceous perennial prairie species, native or naturalized to the Cedar Creek area, were planted in the experiment.

BioCON is a split-plot arrangement of treatments in a completely randomized design. CO2 treatment is the whole-plot factor and is replicated three times among the six rings. The subplot factors of species nmber and N treatment were assigned randomly and replicated in individual plots among the six rings. For each of the four combinations of CO2 and N levels, pooled across all rings, there were 32 randomly assigned replicates for the plots plant to 1 species (2 replicates per species), 15 for those planted to 4 species, 15 for 9 species, and 12 for 16 species (Reich et al., 2001). This arrangement applies to the "main" experiment which utilizes 296 plots.

References from this project:

Reich P. B., J. Knops, D. Tilman, J. Craine, D. Ellsworth, M. Tjoelker, T. Lee, D. Wedin, S. Naeem, D. Bahauddin, G. Hendrey, S. Jose, K. Wrage, J. Goth, and W. Bengston. 2001. Plant diversity enhances ecosystem responses to elevated CO2 and nitrogen deposition. (.pdf) Nature 410:809-812.
Reich P. B., D. Tilman, S. Naeem, D. S. Ellsworth, J. Knops, J. Craine, D. Wedin, and J. Trost. 2004. Species and functional group diversity independently influence biomass accumulation and its response to CO2 and N. (.pdf) Proceedings of the National Academy of Sciences of the United States of America 101:10101-10106.

The Jasper Ridge Global Change Experiment

http://globalecology.stanford.edu/DGE/Dukes/JRGCE/home.html

The Jasper Ridge Global Change Project examines the response of California grassland to four components of global change: elevated atmospheric carbon dioxide, elevated temperature, increased precipitation, and increased nitrogen deposition. Initially funded by a grant from the National Science Foundation to Professors Christopher Field and Harold Mooney, this is the first experimental study to address this broad suite of interacting global changes in the field.

The experimental design includes 8 replicate quarter-circle plots for all possible combinations of the four treatments (128 total) and an additional 8 sampling sites that control for the effects of project infrastructure. Studies focus on four integrated components of ecosystem response to the treatments: plant primary production, soil carbon storage, soil nutrient availability, and species or functional-group composition. The combination of a complete factorial design for the treatments together with measurements on multiple, related response variables, provides a test of the experiment's null hypothesis that responses to warming and elevated CO2 in combination are essentially additive, i.e. the sum of the responses to the individual factors.

The JRGCE is currently funded by grants from the David and Lucile Packard Foundation and from NSF.

References from this project:

Field, CB, Chapin, FS, III, Chiariello, NR, Holland, EA, and Mooney, HA (1996) The Jasper Ridge CO2 Experiment: Design and Motivation. In: Carbon Dioxide and Terrestrial Ecosystems, pp. 121-145. Academic Press, San Diego.

Dukes, JS, and Field, CB (2000) Diverse mechanisms for CO2 effects on grassland litter decomposition. Global Change Biology 6, pp. 145-154.
Shaw, MB, Zavaleta, ES, Chiariello, NR, Cleland, EE, Mooney, HA, and Field, CB (2002) Grassland responses to global environmental changes suppressed by elevated CO2. In: Science 298, pp. 1987-1990.
Zavaleta, ES, Thomas, BD, Chiariello, NR, Asner, GP, Shaw, MR, and Field, CB (2002) Plants reverse warming effect on ecosystem water balance.

The KAEFS warming experiment

http://www.ou.edu/content/ieg/facilities/field-sites.html

The Kessler Atmospheric and Ecological Field Station (KEAFS) is a 360 acre (146 ha) environmental research and education facility located approximately 28 km southwest of the University of Oklahoma campus. It is home to a number of long-term meteorological and biological experiments. The mixed grass prairie ecosystem at KEAFS includes a diverse landscape with mixed and tall grass prairie, woodlands, and riparian communities. This site is an example of the predominant land use in the southern Great Plains and has a land use legacy commonly seen in this area.

A long-term global warming experiment under the direction of Dr. Yiqi Luo has taken place at KAEFS since November, 1999. This project consists of 6 paired plots, one warmed continuously by a quartz heater and the other with a dummy heater. Within each plot are nested subplots, either clipped annually or left unclipped to mimic one of the dominant land use practices in the area, mowing for hay. Warming. Work on the project has included studies on plant and microbial responses, phenology, ecosystem fluxes, mycorrhizal fungi and soil structure.

References from this project:

Guo, X., Zhou, X., Hale, L., Yuan, M., Ning, D., Feng, J., … Zhou, J. (2019). Climate warming accelerates temporal scaling of grassland soil microbial biodiversity. *Nature Ecology and Evolution*, *3*(4), 612–619.
Luo, C., Rodriguez-R, L., Johnston, E., Wu, L., Cheng, L., Xue, K., Tu, Q., Deng, Y., He, Z., Shi, J., Yuan, M., Sherry, R., Li, D., Luo, Y., Schuur, E., Chain, P., Tiedje, J., Zhou, J. and Konstantinidis, K. (2014), Soil Microbial Community Responses to a Decade of Warming as Revealed by Comparative Metagenomics. Applied Environmental Microbiology, 80(5): 1777-1786

The elevation experiment and the Tibetan plateau datasets

https://data.lter-europe.net/deims/site/lter-eap-cn-28

The national field observation station of Haibei Alpine Meadow Ecosystem Research Station in Qinghai province (Haibei Station) was founded in 1976 by Northwest Plateau Institute of Biology, the Chinese Academy of Sciences (CAS). It was located in the northeast of Tibet in a large valley surrounded by the Qilian Mountains at latitude 37°29' - 37°45'N and longitude 101°12' - 101°23'E. The average altitude of the mountain area is 4000 m, and 2900 – 3500 m for the valley area. It is 160 km from Xining City. It belongs to a typical plateau continental climate which is dominated by the southeast monsoon in summer and high pressure from Siberia in winter. There is no obvious seasonal changes, except only a short cool summer and a long severe cold winter. The annual average air temperature is -1.7℃ with extremes of maximum at 27.6℃ and minimum at -37.1℃. The annual precipitation ranges from 426 mm to 860 mm, 80% of which falls in the growing season from May to September. Haibei Research Station has become one of the open stations of Chinese Ecosystem Research Network (CERN) since 1989, and one of the key stations of CERN since 1992. Moreover, it has become one of the field observation and testing stations of State Department of Science and Technology in China since 2001 and become a formal member in 2006. It is now a national and international important research base on alpine terrestrial ecosystem.

The CE dataset originates from a grassland transplant experiment across various elevations (3200, 3400, 3600 and 3800 m). The TB dataset originates from a grazing experiment on the Tibetan plateau.

References from these projects:

Y Yang, L Wu, Q Lin, M Yuan, D Xu, H Yu, Y Hu, J Duan, X Li, Z He, K Xue. Responses of the functional structure of soil microbial community to livestock grazing in the Tibetan alpine grassland. Global change biology 19 (2), 637-648
Y Yang, Y Gao, S Wang, D Xu, H Yu, L Wu, Q Lin, Y Hu, X Li, Z He. The microbial gene diversity along an elevation gradient of the Tibetan grassland. The ISME journal 8 (2), 430-440
Wu, L., Yang, Y., Wang, S., Yue, H., Lin, Q., Hu, Y., ... & Gilbert, J. A. (2017). Alpine soil carbon is vulnerable to rapid microbial decomposition under climate cooling. The ISME journal.

Yue, H., Wang, M., Wang, S., Gilbert, J. A., Sun, X., Wu, L., ... & Zhou, J. (2015). The microbe-mediated mechanisms affecting topsoil carbon stock in Tibetan grasslands. The ISME journal, 9(9), 2012-2020.

Pasture dataset

The two datasets representing forest ecosystems originated from South America and China.

Amazon Rainforest Microbial Observatory (ARMO)

The study was performed at the Amazon Rainforest Microbial Observatory (ARMO; Rodrigues et al. 2013), located at the Fazenda Nova Vida (10o10′5″S and 62o49′27″W), a 22 000-ha cattle ranch in the State of Rondônia, Brazil. The ARMO was established in 2009 to study the effect of land use change on the biodiversity of microorganisms. In particular, ARMO is focused on understanding how conversion of the Amazon Rainforest to agricultural uses impacts the biodiversity of microorganisms. ARMO has three overarching goals: to search for biological novelty, to describe the microbial communities of the Amazon Rainforest, and to determine the effect of agricultural conversion on these communities. Sampling occurred at the end of the rainy season, March 2010 for the following sites: a primary forest (F), two pastures that had been continuously managed for 6 (P6) and 38 (P38) years and a 13-year-old secondary forest (S), which was established in 1997 after pasture abandonment. All pastures were established following the same procedures (details of pasture establishment and management practices are described in the Supporting Information). A 100-m2 transect was placed at each site, and nested transects of 10 m, 1 m and 0.1 m were made sharing the same point of origin, for a total of 12 sampling points, as described in Rodrigues et al. (2013). After the removal of the litter layer, a 5-cm diameter soil core was collected from 0 to 10 cm depth, homogenized and passed through a 2-mm mesh sieve. Samples for total DNA extraction were kept at −80°C, while samples for physicochemical analysis were stored at 4°C.Total carbon and nitrogen were determined with auto analyzer LECO Truspec CN (St. Joseph, MI, USA) at the Centro de Energia Nuclear na Agricultura, University of Sao Paulo, Brazil. Elemental concentrations and soil fertility parameters were analysed according to the methods described by Van Raij et al. (2001), and soil granulometry was determined according to Camargo et al. (1986).

References from this project:

Rodrigues, J. L. M., V. Pellizari, R. Mueller, K. Baek, E. da C Jesus, F. Paula, B. Mirza, G. S. Hamaoui Jr., S. M. Tsai, B. Feigl, J. M. Tiedje, B. J.M. Bohannan, and K. Nusslein. 2013. Conversion of the Amazon Rainforest to agriculture results in biotic homogenization of soil bacterial communities. Proceedings of the National Academy of Sciences USA 110(3): 988-993.

Paula, F. S., Rodrigues, J. L., Zhou, J., Wu, L., Mueller, R. C., Mirza, B. S., ... & Pellizari, V. H. (2014). Land use change alters functional gene diversity, composition and abundance in Amazon forest soil microbial communities. Molecular ecology, 23(12), 2988-2999.

Forest dataset

Shennongjia National Nature Reserve Forest

The study sites, located in Shennongjia Mountain, had a mean annual air temperature of 7.2 °C and annual precipitation of about 1,500 mm, most of which falls during summer. The unique vertical distribution of vegetation on Shennongjia Mountain transforms from evergreen broadleaved forest elevations below 1,300 m, deciduous broadleaved forest between 1,500 and 2,200 m, coniferous forest between 2,200 and 2,600 m, and sub-alpine shrubs above 2,600 m; the plant communities here are generally undisturbed by man. In this study, the plant survey and soil collected were permitted by the administrative bureau of Shennongjia National Nature Reserve. Three typical plant community types along the elevation gradient from 1000 to 2800 m were selected, including evergreen broadleaved forest (EBF1050), deciduous broadleaved forest (DBF1750) and coniferous forest (CF2550). The dominant plant communities are *Cyclobalanopsis oxyodon* (Miq.) Oerst*, Cyclobalanopsis myrsinaefolia* (Blume) in EBF1050, *Carpinus viminea*, *Quercus aliena var. acuteserrata*, *Fagus engleriana* in DBF1750, *Abies fargesii* Franch in CF2550. Samples of the mountain yellow brown soil were collected in September, 2011. At each site, eight 20 × 20 m plots were established with about 20 meters between adjacent plots. In each plot, fifteen 0 - 10 cm deep soil cores were collected and composited to obtain about 400 g soil in total; these were mixed thoroughly and plant roots and stones were removed. Soil samples were preserved at - 80 °C until being thawed for DNA extraction.

References from this project:
Zhang, Y., Cong, J., Lu, H., Deng, Y., Liu, X., Zhou, J., & Li, D. (2016). Soil bacterial endemism and potential functional redundancy in natural broadleaf forest along a latitudinal gradient. Scientific reports, 6, 28819.
Zhang, Y., Cong, J., Lu, H., Li, G., Xue, Y., Deng, Y., ... & Li, D. (2015). Soil bacterial diversity patterns and drivers along an elevational gradient on Shennongjia Mountain, China. Microbial biotechnology, 8(4), 739-746.

Shrubland dataset

The dataset representing shrubland ecosystems originated from China and was collected in the frame of the same project as the SNNRf dataset. Samples were collected in a subalpine shrub dominated by *Rhododendron oreodoxa*.

References from this project:
Zhang, Y., Cong, J., Lu, H., Li, G., Xue, Y., Deng, Y., ... & Li, D. (2015). Soil bacterial diversity patterns and drivers along an elevational gradient on Shennongjia Mountain, China. Microbial biotechnology, 8(4), 739-746.