

Lawrence Berkeley National Laboratory

LBL Publications

Title

Macroecological distributions of gene variants highlight the functional organization of soil microbial systems

Permalink

<https://escholarship.org/uc/item/85q794zm>

Journal

The ISME Journal: Multidisciplinary Journal of Microbial Ecology, 16(3)

ISSN

1751-7362

Authors

Escalas, Arthur

Paula, Fabiana S

Guilhaumon, François

et al.

Publication Date

2022-03-01

DOI

10.1038/s41396-021-01120-8

Copyright Information

This work is made available under the terms of a Creative Commons Attribution-NonCommercial License, available at <https://creativecommons.org/licenses/by-nc/4.0/>

Peer reviewed

Macroecological distributions of gene variants highlight the functional organization of soil microbial systems

Escalas Arthur ^{1,2,*}, Paula Fabiana S. ³, Guilhaumon François ^{1,4}, Yuan Mengting, Yang Yunfeng ^{5,6},
Wu Linwei ², Liu Feifei ^{2,7,8}, Feng Jiaye ², Zhang Yuguang ⁹, Zhou Jizhong ^{2,6,10,11}

¹ MARBEC, Montpellier University-CNRS-IRD-IFREMER, Place Eugène Bataillon, Cedex 5, 34095, Montpellier, France

² Institute for Environmental Genomics and Department of Microbiology and Plant Biology, University of Oklahoma, Norman, OK, 73019, USA

³ Oceanographic Institute, University of São Paulo, São Paulo, 05508-120, Brazil

⁴ IRD, Saint-Denis de la Réunion, France

⁵ Department of Environmental Science, Policy, and Management, University of California, Berkeley, CA, USA, 94704

⁶ State Key Joint Laboratory of Environment Simulation and Pollution Control, School of Environment, Tsinghua University, 100084, Beijing, China

⁷ Guangdong Provincial Key Laboratory of Microbial Culture Collection and Application, Guangdong Institute of Microbiology, Guangdong Academy of Sciences, 510070, Guangzhou, China

⁸ State Key Laboratory of Applied Microbiology Southern China, 510070, Guangzhou, China

⁹ Research Institute of Forest Ecology, Environment and Protection, Chinese Academy of Forestry, and the Key Laboratory of Biological Conservation of National Forestry and Grassland Administration, 100091, Beijing, China

¹⁰ School of Civil Engineering and Environmental Sciences, University of Oklahoma, Norman, OK, 73019, USA

¹¹ Earth and Environmental Sciences, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720, USA

* Corresponding author : Arthur Escalas, email address : arthur.escalas@gmail.com

Abstract :

The recent application of macroecological tools and concepts has made it possible to identify consistent patterns in the distribution of microbial biodiversity, which greatly improved our understanding of the microbial world at large scales. However, the distribution of microbial functions remains largely uncharted from the macroecological point of view. Here, we used macroecological models to examine how the genes encoding the functional capabilities of microorganisms are distributed within and across soil systems. Models built using functional gene array data from 818 soil microbial communities showed that the occupancy-frequency distributions of genes were bimodal in every studied site, and that their rank-abundance distributions were best described by a lognormal model. In addition, the relationships between gene occupancy and abundance were positive in all sites. This allowed us to identify genes with high abundance and ubiquitous distribution (core) and genes with low abundance and limited spatial distribution (satellites), and to show that they encode different sets of microbial traits. Common genes

encode microbial traits related to the main biogeochemical cycles (C, N, P and S) while rare genes encode traits related to adaptation to environmental stresses, such as nutrient limitation, resistance to heavy metals and degradation of xenobiotics. Overall, this study characterized for the first time the distribution of microbial functional genes within soil systems, and highlight the interest of macroecological models for understanding the functional organization of microbial systems across spatial scales.

85 INTRODUCTION

86 The functional potential of microbes relies on the collection of metabolic capabilities encoded by the
87 genes contained in their genomes, and that, once expressed, define the traits of the microorganism carrying these
88 genes. While several functional genes are specific to certain taxa [1], many genes are common to most
89 microorganisms and compose the “core genome” [2, 3]. This results in high levels of functional redundancy
90 among microbial taxa [4–6]. In addition, the wide occurrence of mobile accessory genes exchanged through
91 horizontal transfer [7, 8] plays significant contribution in these systems and adds to their complexity. Further,
92 some genes exist with different sequences despite encoding similar products, which correspond to the functional
93 redundancy among variants of the same gene. For all these reasons, the insights provided by taxonomic
94 approaches into the role of microbial communities in ecosystem functioning are limited. As an alternative, the
95 use of functional approaches has been widely advocated, notably through the direct study of the gene content of
96 microbial communities and regarding genes as potential microbial functional traits [6, 9, 10].

97 In the recent years, researchers have used macroecological models to disentangle the complexity of
98 microbiomes [11, 12]. Such approaches have been notably used to explore commonness and rarity patterns in
99 microbial systems and successfully uncovered regularities in the distribution of microorganisms across various

100 spatial scales (species, communities, or ecosystems) and revealed similarities or idiosyncrasies in the processes
101 underlying these distributions [12–14]. Such macroecology-based frameworks do not rely on arbitrarily defined
102 thresholds that oppose rare *versus* abundant biological units. Instead, full distributions are used to classify units
103 along gradients ranging from rarity to commonness. To investigate patterns in local abundance, rank-abundance
104 distributions (RAD) place biological units from a given area or community along a gradient from low to high
105 abundance. RAD have been used to describe the distribution of taxa within microbial taxa, revealing the
106 presence of a long tail of rare organisms composing the so-called “rare biosphere” [15–17]. In spatial occupancy
107 studies, occupancy-frequency distribution (OFD) describes how biological units are spatially distributed across a
108 set of communities [18], and classifies these units along a distribution gradient, from spatially restricted to
109 ubiquitous. OFD models describing both macro and microorganisms were found to be either unimodal or
110 bimodal, and to exhibit a higher left mode [19], that is, high proportion of taxa represents small fractions of
111 communities. Finally, the relationship between local abundance and spatial occupancy (occupancy-abundance
112 relationships - OAR) is one of the most reported trends in macroecology and has been shown to be positive for a
113 wide range of macro- [20–25] and microorganisms [12, 26–34]. The positive OAR for biological units in natural
114 systems predicts that some units have a restricted spatial distribution with low abundance (*i.e.* “satellites”) while
115 others are ubiquitous and found in high abundance (*i.e.* “core”) [35–38]. In microbial ecology, observations of
116 these distribution patterns contributed to unveil community dynamics, which later led to the concept of
117 “conditionally rare taxa”, *i.e.* satellite organisms having the potential to bloom and temporarily influence
118 community dynamics [39–41]. Such life strategy is known to be related to the metabolic capabilities of certain
119 microbes [42, 43]. Therefore, if we aim to understand the mechanisms underpinning the macroecological
120 patterns of microbial communities, applying these concepts in the context of functional genes could provide
121 valuable information.

122 In this study, we aimed to address a simple, and yet unanswered, question: how are the functions carried
123 by microorganisms distributed within soil systems ? To tackle this question, we aimed to assess the distribution
124 of microbial functional genes at different scales in soils using macroecological models, to identify abundant and
125 rare functions across these systems. For that, we constructed a database with functional gene array (FGA) [44]

126 data from 818 topsoil microbial communities sampled from ten sites located around the globe and representing
127 various ecosystem types (*i.e.* tundra, grassland, forest, shrubland and pasture; Figure 1). The FGA was used to
128 hybridize microbial community DNA to a set of 39,681 probes that correspond to variants of 194 functional
129 genes encoding various microbial functions involved in biogeochemical cycles, pollutant breakdown, virulence
130 and resistance to various types of physical and chemical stress. By considering genes instead of taxa as the unit
131 of our study, we can make the following predictions about their distribution patterns: the presence of a set of core
132 genes shared among most microorganisms should lead to (i) OFD displaying either unimodal with a right mode
133 or bimodal with a stronger right mode; and (ii) RAD characterized by few dominant genes and a long tail of rare
134 genes. Consequently, as observed for most of the biological units from the smallest to the largest, the two
135 previous patterns should result in positive OAR. By using this approach, we classified genes along a continuum
136 from low abundance and limited occurrence to high abundance and ubiquity, with the two ends of this gradient
137 representing satellite and core genes, respectively. Then, by investigating the functions carried by these genes we
138 show that rare and common microbial genes encode different functions in soil ecosystems.

139

140 **MATERIAL AND METHODS**

141 **Composition of the database**

142 The database consisted of ten datasets collected in the frame of previous projects and comprised a total
143 of 818 surface topsoil samples from three continents, representing a wide gradient of environmental conditions
144 (Figure S1 and Table S1). The spatial scale covered by each dataset range from hundreds of meters in some
145 experimental sites to dozens of kilometers across natural landscapes. We did not investigate within site
146 differences among samples as our goal was to look for repeatable patterns across geographically distant sites,
147 considered as separate entities from the physico-chemical, climatic and pedoclimatic standpoint. Here, we opted
148 for a large spatial scale macroecological approach that did not consider local contingencies and focused on
149 comparing the distribution of functional gene variants within and across communities from isolated ecosystems
150 [45].

151 Five sites were located in the United States, including three grassland [46–48] and two Alaskan tundra
152 ecosystems [49, 50]. Climate change experiments were conducted in four of these sites (*i.e.* variation in
153 temperature, CO₂ concentration, etc.). Four sites were located in China, comprising two grassland ecosystems
154 from the Qinghai province [51, 52], in addition to forests and shrubland sites from the Hubei Province [53, 54].
155 The last site corresponded to pasture areas located in the Brazilian Amazon basin [55, 56].

156 All the samples were analyzed using a functional gene array (FGA)[44] composed of 39 681 DNA
157 probes targeting protein-coding genes. Probe design was done as described elsewhere [44, 57, 58], by searching
158 keywords against the NCBI nr database. Candidate sequences were validated with HMM models and 50-mer
159 oligonucleotide probes were designed using CommOligo 2.0 [59]. These probes (hereafter termed genes variants
160 or variants) served as the unit of our study to characterize the macroecological distribution of microbial
161 functions. The potential role of these gene variants in the microbial communities were defined according to a
162 functional classification performed using information available in databases such as NCBI, UniProt, or EXPasy
163 and were also based on extensive literature reviews [10]. The 39 681 variants correspond to 194 genes (*e.g.* nirB,
164 ureC, exochitinase, arsB), defined as collections of variants encoding a similar product but with slightly different
165 DNA sequence and originating from different organisms [60]. Genes were further classified into 56 gene
166 families, defined as collections of genes that, together, represent a coherent set of microbial functions (*e.g.*
167 resistance to oxygen or heat stress, C fixation, denitrification). Finally, these families were grouped into 9 broad
168 categories of microbial functions (*e.g.* C, N, P or S cycling, antibiotic resistance and virulence; *cf.* Table S2 for a
169 full description of the distribution of variants in this different levels of functional resolution). This classification
170 allowed the linkage of genes with the function they carry.

171

172 **Functional Gene Array analyses**

173 FGA hybridizations were performed according to standardized laboratory procedure from the Institute
174 for Environmental Genomics (IEG, OK, USA), as described in [44]. Total community DNA was quantified using
175 picogreen and, for each sample, 800 ng were labeled with Cy-5 (GE Healthcare), dried in a Speedvac at 45°C for

176 45 min and stored at -20°C before hybridization. The pellet was re-hydrated in 2.68 µl of tracking control
177 completed with 7.12 µl of hybridization solution (Formamide, SSC, SDS, oligo Cy-3, oligo Cy-5 and universal
178 standard). Labeled DNA was incubated at 95°C for 5 minutes before loading onto the array. The hybridization
179 was done at 42°C in the presence of 40% formamide for 16 hours. After washing and drying, arrays were
180 scanned and gridded before signal intensity quantification using ImaGene 6.0 (Biodiscovery Inc., El Segundo,
181 CA, USA). Original raw hybridization signal intensity data were retrieved from the IEG microarray data
182 repository (<http://ieg2.ou.edu/NimbleGen/analysis.cgi>). To estimate the abundance of functional genes, noise
183 data were removed using a hybridization signal cutoff of 2000 intensity [57, 61].

184

185 **Distribution patterns of functional gene variants**

186 The abundance of the 39,681 genes variants within a given sample was estimated as the logged
187 hybridization signal intensity on the FGA. The shape of variants rank-abundance distributions (RAD) within
188 each of the 818 samples was assessed using four widely used rank-abundance models (Logseries, Poisson
189 lognormal, Negative binomial and Zipf). Models were fitted using maximum likelihood estimation of parameters
190 and their goodness of fit compared using the AIC (Akaike Information Criterion). For each sample, the model
191 with the lowest AIC value was considered to be the best fitting model. Models were fitted and compared using
192 the python package *macroecotools* [62].

193 Gene variants occupancy was estimated within each site by counting the number of samples in which the
194 variant was detected and dividing it by the total number of samples in the site (ranged between 0 and 1). The
195 shape of occupancy-frequency distribution (OFD) of variants across multiple soil communities was analyzed
196 using the Mitchell-Olds & Shaw test as implemented in the *MOSTest* function of the R package *vegan* [63]. This
197 approach fits a quadratic generalized linear models of the type $\mu = b_0 + b_1x + b_2x^2$ to the OFD, where b_0
198 corresponds to the intercept, b_1 to rate of change and b_2 determines whether the model is convex or concave and
199 was used to estimate the model shape: if $b_2 < 0$ the model is unimodal and if $b_2 > 0$ the model is bimodal.

200 To test occupancy-abundance relationships (OAR) within each site, variants abundance was estimated as
201 the average abundance across samples from that site, occupancy in each site was estimated as described above
202 and the relationships were analyzed using linear models (*lm()* function in R) relating variants occupancy with
203 their average abundance across samples.

204

205 **Translation of gene variants distribution into microbial functions distribution**

206 To determine how the distribution of genes variants translates into the distribution of microbial functions
207 in soil systems, we associated the variants to the functions they encode (Figure 1) using the classification
208 provided by the FGA (genes, gene families and broad categories, *cf.* Table S2). First, variant distributions were
209 used to rank the 39 681 variants and to group them into bins that represented gradients of abundance within a
210 community (RAD), occupancy across communities (OFD) or commonness within sites (OAR). Bins were
211 defined by splitting variants into 6 sets of equal number based on their rank (see supplementary information for
212 an explanation of the choice of bin number). Then, for each bin (ranked 1 for the lowest end of the gradient to 6
213 for the highest end), the importance of a function at a given level of functional classification was estimated as
214 the proportion of the summed hybridization signal of all the variants from that bin. However, different functions
215 (*e.g.* gene families) differed in numbers of variants on the FGA design (Table S2), and this must be accounted for
216 when estimating function importance. If the importance of a function in a bin is simply estimated by counting
217 the number of variants from this function, or their summed signal intensity, then functions represented by many
218 variants are more likely to be considered important than those represented by few variants. To avoid this bias,
219 and thus take into account unequal sampling effort across function on the FGA, function importance within a bin
220 was estimated by dividing the observed proportion of the total signal intensity in the bin represented by variants
221 from this function by the proportion of the total number of variants represented by this function on the FGA
222 design (Figure 1). The obtained ratio, here termed weight of the function, describes how much the proportion of
223 the signal represented by a function in the bin departs from a null expectation, which corresponded to the
224 proportion of the signal represented by this function if variants were randomly sampled on the FGA. Functions

225 with weight values > 1 were considered over-represented in a given bin, *i.e.* more abundant than expected by
226 chance, while functions with a weight < 1 were considered under-represented, *i.e.* less abundant than expected
227 by chance. This provided matrices describing the composition of each bin (column) in terms of microbial
228 functions (rows), with each function being associated with a weight. As bins (B_1 to B_6) represented gradients of
229 increasing abundance within communities, increasing occupancy across communities and increasing
230 commonness within sites, we were able to identify the functions encoded by variants along these gradients.

231

232 **Analysis of the distribution of microbial functions in soil systems**

233 To characterize the distribution of microbial functions in soil systems, we analyzed the weighted
234 matrices described above (*i.e.* functions x bins). The dissimilarity between bins was estimated with the Bray-
235 Curtis index and visualized using Detrended Correspondence Analysis (DCA). We tested for differences in the
236 composition of bins using permutational multivariate analysis of variance (PERMANOVA)[64], implemented as
237 the *adonis* function in the R package *vegan* [63]. We tested the differences between bins of different ranks (1 to
238 6) and originating from different sites. This was done after associating genes to function weight at the three
239 levels of functional classification (genes, gene families and broad categories).

240 For each of the 194 genes, we fitted linear models describing the relationship between the weight of
241 genes in each bin and the rank of occupancy-abundance bins (1 to 6). By looking at the slopes of these models,
242 we identified the genes, and the corresponding gene families, that were under- or over-represented along the
243 occupancy-abundance gradient. Negative relationships (significant negative slopes) corresponded to genes over-
244 represented in rare variants, whereas positive ones (significant positive slopes) corresponded to genes over-
245 represented in abundant variants. When the slope of the linear model was not significant the case was classified
246 as “no relation”. Finally, we characterized the composition in terms of function weight of in rare (B_1) and
247 common (B_6) occupancy-abundance bins.

248

249 **RESULTS**

250 **Macro-ecological distribution patterns of functional gene variants in soil ecosystems**

251 The rank-abundance distributions (RAD) of gene variants within communities (*i.e.* samples) was
252 described using four different RAD models (Logseries, Poisson lognormal, Negative binomial and Zipf). Poisson
253 lognormal was found to be the best model to describe variants RAD in 100% of the samples ($n = 818$, Figure 2).
254 The occupancy-frequency distributions (OFD) of variants across communities were significantly bimodal in all
255 ten sites, with a maximum at low and high occupancy (Figure 3, MOS test, p value < 0.001). In all but one site
256 (CiPEHR), we observed increased variant frequencies at high occupancy, in comparison to low occupancy (*i.e.*
257 the right mode of the OFD was stronger). This was supported by the observation of higher F-values when testing
258 the presence of a frequency maximum at high occupancy ($F = 103 \pm 74$) compared with low occupancy ($F =$
259 59 ± 57 , Table S3). F-values at high occupancy were 1.7 ± 0.7 times higher than at low occupancy. F-values of the
260 left mode was less pronounced for the two datasets with the lowest number of samples (KAEFS, $n = 12$ and
261 Fazenda nova vida, $n = 24$), suggesting that the sampling effort was not high enough to capture variants with
262 spatially restricted distribution. The relationships between average variant abundance and occupancy (OAR)
263 were linear, positive and highly significant in all ten sites (Figure 4). This linear trend represents a gradient
264 ranging from rarity, *i.e.* low abundance and restricted spatial distribution (bottom left), to commonness, *i.e.*
265 ubiquitous distribution across communities and high abundance (top right). At the two ends of this gradients lie
266 satellite (B_1) and core gene variants (B_1), respectively.

267

268 **Distribution of microbial functions in soil ecosystems**

269 We found that abundance bins from different samples but with similar rank had a more similar
270 distribution of function weights than bins from the same sample but with different ranks. This was validated for
271 each site, as bins with similar rank clustered together on the DCA, based on the Bray-Curtis dissimilarity
272 estimated on function weights (Figure S2). Additionally, bins located at the two ends of the abundance gradient
273 within samples were the most dissimilar (B_1 and B_6). This result was also validated when comparing the weight
274 of functions in abundance bins across sites ($n = 818$ samples, times 6 abundance bins). The weight of genes in

275 abundance bins was better predicted by the rank of the bin along abundance gradient (B_1 to B_6 , PERMANOVA,
276 $p < 0.01$) than by its site of origin (Table 1). Bin rank explained between 65 and 84% of the variation in gene
277 weight while the site explained between 8 and 16%. This trend was confirmed when higher levels of variants
278 classification were used (*e.g.* gene families and broad categories), as suggested by higher F values of bin rank
279 compared with the site effect (*i.e.* 18.6, 8.3 and 8.1 times higher for broad categories, gene families and genes,
280 respectively).

281 We performed a similar analysis using occupancy bins (B_1 to B_6) and found that variants with similar
282 occupancy within sites exhibited similar function weights. Bin rank explained from 31 to 57% of the variation in
283 gene weight distribution between bins from the ten sites, while the factor site explained only 8 to 15% (Table 1).
284 According to the F values, the effect of occupancy rank was 5.4 to 13.3 times higher than the site effect. As
285 observed for abundance bins, the greatest differences in distribution of function weight among occupancy bins
286 were observed between the two extremes of the gradient, B_1 and B_6 (Table S4).

287

288 **Functions of satellites and core genes in soil ecosystems**

289 We observed clear trends in the distribution of the 194 gene families along the occupancy-abundance
290 gradient in soil systems, and we identified the gene families, and the corresponding broad ecological categories
291 that were systematically over-represented at one end of this gradient (Figures 5 and S3-S4-S5, Table S5). Among
292 the 194 linear models fitted between genes weight and bin rank, only 22 (11%) were not significant (p value $>$
293 0.05, Figure 5-A and Figure S3). This corresponded to genes that were not associated with rare or abundant
294 variants. We observed 91 (47%) negative relationships (p value < 0.05 and slope < 0 , Figure 5-A and Figure S4),
295 corresponding to genes that were over-represented in rare (*i.e.* satellites) variants and under-represented in
296 abundant (*i.e.* core) variants. Among these, 38.5% of the genes were related to stress responses (*e.g.* osmotic,
297 oxygen or radiation stress, cold or heat shocks, sigma factors, N or P limitations), 18.7% to metal resistance,
298 14.3% to C cycling, 11% to virulence, and the remaining 6% comprised three categories (antibiotic resistance, N
299 and S cycling). The 20 genes with the strongest negative slope were related to various forms of stress responses,

300 virulence proteins (toxin, adhesin, aerobactin), metal resistance (cadmium, cobalt, aluminum), broad biological
301 functions (blue copper protein, thioredoxin), C cycling (acetogenesis) and energy processes (hydrogenase).
302 Significant positive relationships were found for 81 (42%) genes (p value < 0.05 and slope > 0, Figure 5-A and
303 Figure S5). These genes were under-represented in rare and over-represented in abundant variants. From those,
304 39.5% were related to C cycling, 18.5% to metal resistance, 12.3% to N cycling, 9.9% to stress responses, 4.9%
305 to antibiotic resistance and S cycling, 3.7% to energy processes and P cycling and 2.5% to virulence. The 20
306 genes with the strongest positive slope were related to the degradation of C-based substrates, the N cycle
307 (denitrification, assimilatory-N-reduction and ammonification), metal resistance (lead, silver and mercury), C
308 fixation (pcc and CODH genes), S oxidation (sox gene) and energy processes (hydrogenase).

309 We also looked at the weight of each gene family across the ten sites in the first (B_1 , satellite variants)
310 and in the sixth (B_6 , core variants) occupancy-abundance bins (Figure 5-C-D, figure S6). Core variants were
311 enriched in functions related to the C, N, P and S biogeochemical cycles, but were depleted in functions related
312 to stress response, virulence, heavy metal and antibiotics resistance. Satellite variants were more evenly
313 distributed across the categories, despite notable depletion in functions related to the N cycle and enrichment in
314 stress response and virulence related functions. Twelve processes were clearly enriched in core variants: S
315 oxidation, denitrification, C fixation, ammonification, assimilatory-N-reduction, C degradation, P utilization and,
316 surprisingly, resistance to mercury, lead and silver contamination. On the contrary, the processes enriched in
317 satellite variants included stress response (*e.g.* oxygen limitation, heat shocks, radiation, osmotic and protein
318 stresses, P and N limitation), antibiotics resistance (*e.g.* membrane transporters), resistance to heavy metal (*e.g.*
319 Cr, Cu, As, Te and Al) and virulence (*e.g.* hemolysin, capsule formation, pilin, aerobactin and pilin).
320 Interestingly, two C-related processes (methane metabolism and acetogenesis) were also enriched in satellite
321 variants.

322

323 **DISCUSSION**

324 **Macro-ecological distribution patterns of microbial gene variants in soil systems**

325 In this study, we applied an analytical framework derived from macroecological concepts to describe the
326 distribution of microbial gene variants at two scales, *i.e.* within and across communities, in ten different soil
327 ecosystems. We showed that rank-abundance distributions (RAD) of gene variants within soil communities can
328 be adequately described using classic macroecological models that were designed to capture the intrinsically
329 uneven distribution of species within natural assemblages. Here the Poisson lognormal model was the best one to
330 describe variants RAD. While many RAD models have been developed over time to describe these data, the
331 Poisson lognormal model is often considered as the most widely applicable due to its “*positive range, right*
332 *skewness, heavy right tail, and easily computed parameter estimates*” [65]. In microbes, it was identified as the
333 best model to characterize bacterial RAD at the global scale [14], in the marine environment [66] and in
334 wastewater treatment plants [67], and it was used to predict the total number of microbial OTUs at a global
335 scale [68, 69]. Overall, our results highlight that the wide applicability of the lognormal model to describe RAD
336 of biological units can be extended to microbial gene variants. Furthermore, we found that RAD of microbial
337 taxa and gene variants are very similar, which demonstrates the usefulness of macroecological tools beyond the
338 dichotomy micro- vs. macro-organisms [12], and toward a wider range of biological units (*e.g.* genes,
339 interactions, viruses).

340 When looking at the spatial distribution of gene variants across samples, the observed occupancy-
341 frequency distributions (OFD) differed from what is generally reported for taxa in communities of both macro
342 and microorganisms (*i.e.* the “hollow” distribution), with a higher number of taxa being found in a few sites and
343 only a small number of ubiquitous taxa [70]. In their review, McGeoch & Gaston (2002), analyzed OFD models
344 describing the distribution of macro-organisms (*e.g.* plants, insects, birds, fishes) from small (< 1 km²) to
345 continental scales. Among the 68 reported models, some were unimodal (57%), other bimodal (31%), but the
346 large majority exhibited a higher left mode (68%), that is a higher proportion of taxa observed in a small
347 proportion of communities than widely distributed. Similar right-skewed OFD have been reported for microbial
348 taxa, from the microscale [33] to hundreds of km in both marine [31, 32, 71, 72] and soil environments [27, 73].
349 Here, we found that the OFD of microbial functional gene variants contrast with these general trends reported
350 for taxa, as they exhibit a stronger right mode with a much higher proportion of variants that were ubiquitously

351 distributed within a site than unique to a single community. This pattern was expected and due to the functional
352 redundancy among microbes [4–6, 74], *i.e.* the fact that most microbes share a common set of functional genes
353 that can be detected in any soil sample collected within a given site. Despite this stronger right mode, the observed
354 OFD were bimodal in the ten studied ecosystems, a pattern known as the Raunkiaer's law of distribution of
355 frequencies. It has been suggested that this pattern can emerge from random sampling of biological units from a
356 lognormal rank-abundance distribution [75], which seems to be the case in our study.

357 The combination of within-community abundance distribution (RAD) and across-communities
358 occupancy distribution (OFD) corresponds to occupancy-abundance relationships (OAR), and these OAR have
359 been reported to be positive for a wide range of macro- and microbial taxa [12, 20–26, 28–34]. Several theories
360 have been proposed to explain the existence of positive OAR in taxa, including stochastic processes resulting
361 from neutral dynamics [19] or differences between species in terms of ecological niche [76]. However, there is
362 currently no consensus on the underlying mechanisms of OAR for taxa and no studies for functional genes.
363 Positive OAR can be seen as a gradient of commonness (or rarity) across a set of communities, with biological
364 units that are both spatially restricted and locally scarce at one end (*i.e.* the satellites), and the biological units
365 that are widespread and very abundant at the other end (*i.e.* the core). Here, we observed that some gene variants
366 were present in a small number of communities within each site and exhibited low abundance in these
367 communities, and that other gene variants exhibited high abundance in all the communities from all the sites.
368 This resulted in positive OAR of microbial functional gene variants within each studied site, and allowed the
369 identification of rare and common microbial gene variants in soil systems, along with the function they encode.
370 However, despite the fact that OAR of microbial taxa and gene variants are both positive, they differ greatly
371 regarding the distribution of biological units along the rarity to commonness gradient. As mentioned, while the
372 RAD are quite similar the OFD are very different. These differences resulted in taxonomic OAR with many
373 satellite and few core taxa while functional OAR had only few satellite and many core gene variants (Figure 6).
374 Interestingly, ecological theories suggest that core taxa are more likely to be generalists while satellite ones are
375 more likely to be specialists [77, 78], which leads to the question whether core and satellite gene variants encode
376 general and specialized functions, respectively.

377

378 **From gene variants distribution to the functional organization of soil microbial systems**

379 Overall, we found strong differences in the functions encoded by satellite and core gene variants. There
380 are several possibilities for gene variants to be identified as core. It could be present in the genome of a single
381 widespread and abundant taxa (generalist), in the genomes of several widespread and low abundance taxa or in
382 the genomes of many spatially restricted and low abundance taxa (specialists). Unfortunately, we could not
383 quantify the contributions of these different scenarii, as the FGA did not allow to link functional genes with the
384 identity of the taxa carrying them. Our results show that core functional gene variants correspond mostly to
385 genes related to the main biogeochemical cycles (C, N, P and S) and support the hypothesis that a wide range of
386 microorganisms have the abilities to carry out fundamental ecological processes such as degradation of C-based
387 substrates, denitrification or assimilatory-N-reduction [79, 80]. This is not surprising from a functional
388 perspective, as microbial systems are well known for their redundancy in the metabolic capabilities between
389 organisms [6] and, at larger scale, across communities [4, 5, 81–85].

390 By contrast, there are fewer possibilities for a gene variant to be identified as a satellite as it must be
391 present in only spatially restricted and low abundance taxa. Consequently, core variants are expected to represent
392 the functions that are shared by microorganisms with many different macroecological distributions, while
393 satellite variants represent the function that are found only in spatially restricted and low abundance
394 microorganisms. We found that satellite gene variants encoded the capabilities of microorganisms to cope with
395 environmental stresses (*e.g.* osmotic, oxygen or radiation, cold or heat shocks, sigma factors), withstand nutrient
396 limitations (N and P) and resist to pollutants or potentially toxic compounds (*i.e.* heavy metals, antibiotics). This
397 result could explain why rare microbes that likely carry these variants appears less affected by disturbances and
398 abiotic changes compared with dominant ones, which tend to respond to a higher number of disturbances and
399 oscillate in abundance when facing them [86]. A step further, these results support previous observations that the
400 ecological strategy of some microorganisms is to maintain a low abundance and a slow growth, while
401 prioritizing the expression of maintenance and survival functions [87, 88]. Our results also support recent

402 findings showing that rare microorganisms are particularly important for the adaptation of microbial
403 communities to environmental variation and their ability to withstand perturbations and maintain ecosystem
404 functions across spatio-temporal scales [41, 89]. In fact, rare and dominant microorganisms are thought to carry
405 redundant metabolic potential regarding major functions (C, N, P cycles), but the rare ones harbor distinct
406 abilities to cope with environmental changes. Hence, they may temporarily thrive and support the functioning at
407 the community level by replacing dominant taxa that were affected by these changes [90–92].

408 It is worth mentioning that we characterized the functional content of microbial communities using
409 FGA, which was constrained by the array design and represented only a fraction of the gene diversity that can be
410 assessed using deep shotgun sequencing [93]. This could have resulted in an underestimation of the functional
411 potential represented by rare genes (and thus the rare biosphere). With the advance in sequencing techniques and
412 expansion of databases, our ability to detect rare genes is rapidly increasing. It is likely that the differences
413 observed here between the functional potential encoded by core and satellite genes would have been even greater
414 using deep shotgun sequencing. However, the FGA approach also offered several advantages over sequencing
415 that are particularly relevant for our study, as it provided a level of reproducibility and standardization that could
416 not be matched by sequencing approaches. Such a standardize microbial data system was recently termed as
417 “highly needed” for pursuing questions related to ,microbial macroecology [45]. In addition, microarrays are
418 often more accurate for genes quantification and are more sensitive to rare genes than sequencing approaches
419 [94, 95], making them particularly well suited for analyzing occurrence and abundance patterns of functional
420 genes. To conclude, we foresee that the conceptual approach proposed here could be adapted to the analysis of
421 publicly available metagenomic datasets in order to characterize the distribution of microbial functions across a
422 wide range of environments.

423 In this study, we showed that the distribution of microbial gene variants can be adequately described
424 using concepts and tools derived from the field of macroecology. This approach allowed us to classify gene
425 variants along a gradient from rarity to commonness, showing that variants with low abundance and limited
426 spatial distribution encode functions that are distinct from those encoded by variants with high abundance and

427 ubiquitous distribution. Common variants encode microbial traits involved in the major biogeochemical cycles
428 (C, N, P and S) while rare ones encode traits allowing microorganisms to withstand environmental stresses and
429 nutrient limitation, along with their resistance to heavy metals and xenobiotics. Our results support the
430 hypothesis that the rare biosphere carries different functional capabilities compared with more prevalent
431 microbes and that these capabilities may determine the essential role of rare microbes in the resilience of
432 microbial communities and their ability to sustain ecological processes across temporal and spatial scales.

433

434 **ACKNOWLEDGEMENTS**

435 The authors would like to thank all the persons that contributed to sample collection and laboratory analyses.
436 This synthesis was primarily funded by the U.S. Department of Energy (DOE), Office of Science, Office of Bio-
437 logical and Environmental Research's (OBER) Systems Biology Research to Advance Sustainable Bioenergy
438 Crop Development (DE-SC0014079), Biological Systems Research on the Role of Microbial Communities in
439 Carbon Cycling program (DE-SC0004730, DE-SC001057, DE-SC0004601 and DE-SC0010715), by the U.S.
440 National Science Foundation MacroSystems Biology program under the contract (NSF EF-1065844), and by the
441 Office of the Vice President for Research at the University of Oklahoma, all to J.Z.. This work was also sup-
442 ported by the National Natural Science Foundation of China (No.31670614) to Y.Y.

443

444 **REFERENCES**

445

- 446 1. Gupta A, Sharma VK. Using the taxon-specific genes for the taxonomic classification of bacterial
447 genomes. *BMC Genomics* 2015; **16**.
- 448 2. Gil R, Silva FJ, Pereto J, Moya A. Determination of the Core of a Minimal Bacterial Gene Set. *Microbiol*
449 *Mol Biol Rev* 2004; **68**: 518–537.
- 450 3. Mira A, Martín-Cuadrado AB, D'Auria G, Rodríguez-Valera F. The bacterial pan-genome: A new
451 paradigm in microbiology. *Int Microbiol* 2010; **13**: 45–57.
- 452 4. Escalas A, Troussellier M, Yuan T, Bouvier T, Bouvier C, Mouchet MA, et al. Functional diversity and
453 redundancy across fish gut, sediment and water bacterial communities. *Environ Microbiol* 2017; **19**:
454 3268–3282.
- 455 5. Jurburg SD, Salles JF. Functional Redundancy and Ecosystem Function — The Soil Microbiota as a Case
456 Study. In: Lo Y-H, Blanco JA, Shovonlal R (eds). *Biodiversity in Ecosystems - Linking Structure and*
457 *Function*. 2015. pp 29–49.

- 458 6. Louca S, Polz MF, Mazel F, Albright MBN, Huber JA, O'Connor MI, et al. Function and functional
459 redundancy in microbial systems. *Nat Ecol Evol* 2018; **2**: 936–943.
- 460 7. Polz MF, Hunt DE, Preheim SP, Weinreich DM. Patterns and mechanisms of genetic and phenotypic
461 differentiation in marine microbes. *Philos Trans R Soc Lond B Biol Sci* 2006; **361**: 2009–2021.
- 462 8. Young JPW. Bacteria Are Smartphones and Mobile Genes Are Apps. *Trends Microbiol* 2016; **24**: 931–
463 932.
- 464 9. Boon E, Meehan CJ, Whidden C, Wong DHJ, Langille MGI, Beiko RG. Interactions in the microbiome:
465 Communities of organisms and communities of genes. *FEMS Microbiol Rev* 2014; **38**: 90–118.
- 466 10. Escalas A, Hale L, Voordeckers JW, Yang Y, Firestone MK, Alvarez-Cohen L, et al. Microbial Functional
467 Diversity: From Concepts to Applications. *Ecol Evol* 2019.
- 468 11. Barberán A, Casamayor EO, Fierer N. The microbial contribution to macroecology. *Front Microbiol* .
469 2014. , **5**: 1–8
- 470 12. Shade A, Dunn RR, Blowes SA, Keil P, Bohannan BJM, Herrmann M, et al. Macroecology to Unite All
471 Life, Large and Small. *Trends Ecol Evol* 2018; **33**: 731–744.
- 472 13. Chase AB, Martiny JB. The importance of resolving biogeographic patterns of microbial microdiversity.
473 *Microbiol Aust* 2018; 5–8.
- 474 14. Shoemaker WR, Locey KJ, Lennon JT. A macroecological theory of microbial biodiversity. *Nat Ecol*
475 *Evol* 2017; **1**: e1450v4.
- 476 15. Bachy C, Worden AZ. Microbial ecology: Finding structure in the rare biosphere. *Curr Biol* . 2014.
477 Elsevier. , **24**: R315–R317
- 478 16. Lynch MDJ, Neufeld JD. Ecology and exploration of the rare biosphere. *Nat Rev Microbiol* 2015; **13**:
479 217–229.
- 480 17. Pedrós-Alió C. The Rare Bacterial Biosphere. *Ann Rev Mar Sci* 2012; **4**: 449–466.
- 481 18. Rabinowitz D. Seven forms of rarity and their frequency in the flora of the British Isles. In: Soulé ME
482 (ed). *Conservation Biology: The Science of Scarcity and Diversity*. 1986. Sinauer Associates.
- 483 19. McGeoch MA, Gaston KJ. Occupancy frequency distributions: Patterns, artefacts and mechanisms. *Biol*
484 *Rev Camb Philos Soc* 2002; **77**: 311–331.
- 485 20. Blackburn TM, Cassey P, Gaston KJ. Variations on a theme: Sources of heterogeneity in the form of the
486 interspecific relationship between abundance and distribution. *J Anim Ecol* 2006; **75**: 1426–1439.
- 487 21. Buckley HL, Freckleton RP. Understanding the role of species dynamics in abundance-occupancy
488 relationships. *J Ecol* 2010; **98**: 645–658.
- 489 22. Gaston KJ, Blackburn TM, Greenwood JJD, Gregory RD, Quinn RM, Lawton JH. Abundance-occupancy
490 relationships. *J Appl Ecol* 2000; **37**: 39–59.
- 491 23. Miranda LE, Killgore KJ. Abundance–occupancy patterns in a riverine fish assemblage. *Freshw Biol*
492 2019; **64**: 2221–2233.

- 493 24. Suhonen J, Jokimäki J. Temporally stable species occupancy frequency distribution and abundance-
494 occupancy relationship patterns in urban wintering bird assemblages. *Front Ecol Evol* 2019; **7**.
- 495 25. Webb TJ, Barry JP, McClain CR. Abundance–occupancy relationships in deep sea wood fall
496 communities. *Ecography* 2017; **40**: 1339–1347.
- 497 26. Amend AS, Oliver TA, Amaral-Zettler LA, Boetius A, Fuhrman JA, Horner-Devine MC, et al.
498 Macroecological patterns of marine bacteria on a global scale. *J Biogeogr* 2013; **40**: 800–811.
- 499 27. Barberán A, Bates ST, Casamayor EO, Fierer N. Using network analysis to explore co-occurrence
500 patterns in soil microbial communities. *ISME J* 2012; **6**: 343–351.
- 501 28. Barnes CJ, Burns CA, van der Gast CJ, McNamara NP, Bending GD. Spatio-temporal variation of core
502 and satellite arbuscular mycorrhizal fungus communities in *Miscanthus giganteus*. *Front Microbiol* 2016;
503 **7**: 1–12.
- 504 29. Fillol M, Auguet JC, Casamayor EO, Borrego CM. Insights in the ecology and evolutionary history of the
505 Miscellaneous Crenarchaeotic Group lineage. *ISME J* 2016; **10**: 665–677.
- 506 30. Jeanbille M, Gury J, Duran R, Tronczynski J, Agogué H, Saïd O Ben, et al. Response of core microbial
507 consortia to chronic hydrocarbon contaminations in coastal sediment habitats. *Front Microbiol* 2016; **7**:
508 1–13.
- 509 31. Lindh M V., Sjöstedt J, Ekstam B, Casini M, Lundin D, Hugerth LW, et al. Metapopulation theory
510 identifies biogeographical patterns among core and satellite marine bacteria scaling from tens to
511 thousands of kilometers. *Environ Microbiol* 2017; **19**: 1222–1236.
- 512 32. Logares R, Audic SS, Bass D, Bittner L, Boutte C, Christen R, et al. Patterns of Rare and Abundant
513 Marine Microbial Eukaryotes. *Curr Biol* 2014; **24**: 813–821.
- 514 33. Michelland R, Thioulouse J, Kyselková M, Grundmann GL. Bacterial Community Structure at the
515 Microscale in Two Different Soils. *Microb Ecol* 2016; **72**: 717–724.
- 516 34. Unterseher M, Jumpponen A, Öpik M, Tedersoo L, Moora M, Dormann CF, et al. Species abundance
517 distributions and richness estimations in fungal metagenomics - Lessons learned from community
518 ecology. *Mol Ecol* 2011; **20**: 275–285.
- 519 35. Grime JP. Benefits of plant diversity to ecosystems: Immediate, filter and founder effects. *J Ecol* 1998;
520 **86**: 902–910.
- 521 36. Grime JP. Dominant and subordinate components of plant communities: implications for succession, sta-
522 bility and diversity. In: Gray AJ, Crawley MJ (eds). *Colonization, Succession and Stability*. 1984.
523 Blackwell Scientific Publications, Oxford, pp 413–428.
- 524 37. Hanski I. Dynamics of Regional Distribution: The Core and Satellite Species Hypothesis. *Oikos* 1982;
525 **38**: 210.
- 526 38. Magurran AE, Henderson PA. Explaining the excess of rare species in natural species abundance
527 distributions. *Nature* 2003; **422**: 714–716.

- 528 39. Newton R, Shade A. Lifestyles of rarity: understanding heterotrophic strategies to inform the ecology of
529 the microbial rare biosphere. *Aquat Microb Ecol* 2016; **78**: 51–63.
- 530 40. Shade A, Jones SE, Caporaso JG, Handelsman J, Knight R, Fierer N, et al. Conditionally rare taxa
531 disproportionately contribute to temporal changes in microbial diversity. *MBio* 2014; **5**: e01371-14.
- 532 41. Shade A, Gilbert JA. Temporal patterns of rarity provide a more complete view of microbial diversity.
533 *Trends Microbiol* 2015; **23**: 335–340.
- 534 42. Koch AL. Oligotrophs versus copiotrophs. *BioEssays* 2001; **23**: 657–661.
- 535 43. Cobo-Simón M, Tamames J. Relating genomic characteristics to environmental preferences and ubiquity
536 in different microbial taxa. *BMC Genomics* 2017; **18**: 1–11.
- 537 44. Tu Q, Yu H, He Z, Deng Y, Wu L, Van Nostrand JD, et al. GeoChip 4: A functional gene-array-based
538 high-throughput environmental technology for microbial community analysis. *Mol Ecol Resour* 2014; **14**:
539 914–928.
- 540 45. Xu X, Wang N, Lipson D, Sinsabaugh R, Schimel J, He L, et al. Microbial macroecology: In search of
541 mechanisms governing microbial biogeographic patterns. *Glob Ecol Biogeogr* 2020; **29**: 1870–1886.
- 542 46. Reich PB, Knops J, Tilman D, Craine J, Ellsworth D, Tjoelker M, et al. Plant diversity enhances
543 ecosystem responses to elevated CO₂ and nitrogen deposition. *Nature* 2001; **410**: 809–812.
- 544 47. Field CB, Chapin FS, Chiariello NR, Holland EA, Mooney HA. The Jasper Ridge CO₂ Experiment:
545 Design and Motivation. *Carbon Dioxide and Terrestrial Ecosystems*. 1996. pp 121–145.
- 546 48. Luo C, Rodriguez-R LM, Johnston ER, Wu L, Cheng L, Xue K, et al. Soil microbial community
547 responses to a decade of warming as revealed by comparative metagenomics. *Appl Environ Microbiol*
548 2014; **80**: 1777–1786.
- 549 49. Mauritz M, Bracho R, Celis G, Hutchings J, Natali SM, Pegoraro E, et al. Nonlinear CO₂ flux response
550 to 7 years of experimentally induced permafrost thaw. *Glob Chang Biol* 2017; **23**: 3646–3666.
- 551 50. Natali SM, Schuur EAG, Mauritz M, Schade JD, Celis G, Crummer KG, et al. Permafrost thaw and soil
552 moisture driving CO₂ and CH₄ release from upland tundra. *J Geophys Res Biogeosciences* 2015; **120**:
553 525–537.
- 554 51. Yang Y, Gao Y, Wang S, Xu D, Yu H, Wu L, et al. The microbial gene diversity along an elevation
555 gradient of the Tibetan grassland. *ISME J* 2014; **8**: 430–440.
- 556 52. Yang Y, Wu L, Lin Q, Yuan M, Xu D, Yu H, et al. Responses of the functional structure of soil microbial
557 community to livestock grazing in the Tibetan alpine grassland. *Glob Chang Biol* 2013; **19**: 637–648.
- 558 53. Zhang Y, Cong J, Lu H, Li G, Xue Y, Deng Y, et al. Soil bacterial diversity patterns and drivers along an
559 elevational gradient on Shennongjia Mountain, China. *Microb Biotechnol* 2015; **8**: 739–746.
- 560 54. Zhang Y, Cong J, Lu H, Deng Y, Liu X, Zhou J, et al. Soil bacterial endemism and potential functional
561 redundancy in natural broadleaf forest along a latitudinal gradient. *Sci Rep* 2016; **6**.

- 562 55. Paula FS, Rodrigues JLM, Zhou J, Wu L, Mueller RC, Mirza BS, et al. Land use change alters functional
563 gene diversity, composition and abundance in Amazon forest soil microbial communities. *Mol Ecol* 2014;
564 **23**: 2988–2999.
- 565 56. Rodrigues JLM, Pellizari VH, Mueller R, Baek K, Jesus EDC, Paula FS, et al. Conversion of the Amazon
566 rainforest to agriculture results in biotic homogenization of soil bacterial communities. *Proc Natl Acad*
567 *Sci U S A* 2013; **110**: 988–93.
- 568 57. He Z, Deng Y, Van Nostrand JD, Tu QC, Xu MY, Hemme CL, et al. GeoChip 3.0 as a high-throughput
569 tool for analyzing microbial community composition, structure and functional activity. *Isme J* 2010; **4**:
570 1167–1179.
- 571 58. He Z, Gentry TJ, Schadt CW, Wu L, Liebich J, Chong SC, et al. GeoChip: A comprehensive microarray
572 for investigating biogeochemical, ecological and environmental processes. *ISME J* 2007; **1**: 67–77.
- 573 59. Li X, He Z, Zhou J. Selection of optimal oligonucleotide probes for microarrays using multiple criteria,
574 global alignment and parameter estimation. *Nucleic Acids Res* 2005; **33**: 6114–6123.
- 575 60. Tu Q, He Z, Deng Y, Zhou J. Strain/species-specific probe design for microbial identification
576 microarrays. *Appl Environ Microbiol* 2013; **79**: 5085–5088.
- 577 61. Wu L, Liu X, Schadt CW, Zhou J. Microarray-based analysis of subnanogram quantities of microbial
578 community DNAs by using whole-community genome amplification. *Appl Environ Microbiol* 2006; **72**:
579 4931–4941.
- 580 62. Xiao X, Thibault K, J. Harris D, Baldrige E, White E. macroecotools: v0.3 (Version v0.3). 2016.
- 581 63. Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O’Hara R. Vegan: community ecology
582 package. *R Package*. 2016.
- 583 64. Anderson MJ, Bueno AS. A new method for non-parametric multivariate analysis of variance. *Austral*
584 *Ecol* 2001; **26**: 32–46.
- 585 65. Crow EL, Patil GP. Applications in Ecology. In: Cros E, Shimizu K (eds). *Lognormal Distributions*.
586 1988. Marcel Dekker, New York and Basel, pp 303–330.
- 587 66. Ser-Giacomi E, Zinger L, Malviya S, De Vargas C, Karsenti E, Bowler C, et al. Ubiquitous abundance
588 distribution of non-dominant plankton across the global ocean. *Nat Ecol Evol* 2018; **2**: 1243–1249.
- 589 67. Wu L, Ning D, Zhang B, Li Y, Zhang P, Shan X, et al. Global diversity and biogeography of bacterial
590 communities in wastewater treatment plants. *Nat Microbiol* 2019; **4**: 1183–1195.
- 591 68. Locey KJ, Lennon JT. Scaling laws predict global microbial diversity. *Proc Natl Acad Sci* 2016; **113**:
592 5970–5975.
- 593 69. Louca S, Mazel F, Doebeli M, Parfrey LW. A census-based estimate of Earth’s bacterial and archaeal
594 diversity. *PLoS Biol* 2019; 1–30.
- 595 70. Tokeshi M. Dynamics of distribution in animal communities: Theory and analysis. *Res Popul Ecol*
596 (*Kyoto*) 1992; **34**: 249–273.

- 597 71. Logares R, Deutschmann IM, Junger PC, Giner CR, Krabberød AK, Schmidt TSB, et al. Disentangling
598 the mechanisms shaping the surface ocean microbiota. *Microbiome* 2020; **8**: 55.
- 599 72. Azovsky A, Mazei Y. Do microbes have macroecology? Large-scale patterns in the diversity and
600 distribution of marine benthic ciliates. *Glob Ecol Biogeogr* 2013; **22**: 163–172.
- 601 73. Noguez AM, Arita HT, Escalante AE, Forney LJ, García-Oliva F, Souza V. Microbial macroecology:
602 Highly structured prokaryotic soil assemblages in a tropical deciduous forest. *Glob Ecol Biogeogr* 2005;
603 **14**: 241–248.
- 604 74. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, et al. A communal catalogue
605 reveals Earth’s multiscale microbial diversity. *Nature* 2017; **551**: 457–463.
- 606 75. Papp L, Izsák J, Papp L, Izsak J. Bimodality in Occurrence Classes: A Direct Consequence of Lognormal
607 or Logarithmic Series Distribution of Abundances- A Numerical Experimentation. *Oikos* 1997; **79**: 191.
- 608 76. Verberk WCEP, van der Velde G, Esselink H. Explaining abundance-occupancy relationships in
609 specialists and generalists: A case study on aquatic macroinvertebrates in standing waters. *J Anim Ecol*
610 2010; **79**: 589–601.
- 611 77. Liao J, Cao X, Zhao L, Wang J, Gao Z, Wang MC, et al. The importance of neutral and niche processes
612 for bacterial community assembly differs between habitat generalists and specialists. *FEMS Microbiol*
613 *Ecol* 2016; **92**: fiw174.
- 614 78. Slatyer RA, Hirst M, Sexton JP. Niche breadth predicts geographical range size: A general ecological
615 pattern. *Ecol Lett* 2013; **16**: 1104–1114.
- 616 79. Fierer N, Barberán A, Laughlin DC. Seeing the forest for the genes: Using metagenomics to infer the
617 aggregated traits of microbial communities. *Front Microbiol* 2014; **5**: 1–6.
- 618 80. Rivett DW, Bell T. Abundance determines the functional role of bacterial phylotypes in complex
619 communities. *Nat Microbiol* 2018; **3**: 767–772.
- 620 81. Wertz S, Degrange V, Prosser JI, Poly F, Commeaux C, Guillaumaud N, et al. Decline of soil microbial
621 diversity does not influence the resistance and resilience of key soil microbial functional groups
622 following a model disturbance. *Environ Microbiol* 2007; **9**: 2211–2219.
- 623 82. Wertz S, Degrange V, Prosser JI, Poly F, Commeaux C, Freitag T, et al. Maintenance of soil functioning
624 following erosion of microbial diversity. *Environ Microbiol* 2006; **8**: 2162–2169.
- 625 83. Mendes LW, Tsai SM, Navarrete AA, de Hollander M, van Veen JA, Kuramae EE. Soil-Borne
626 Microbiome: Linking Diversity to Function. *Microb Ecol* 2015; **70**: 255–265.
- 627 84. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, et al. Structure and function of
628 the global ocean microbiome - SM. *Science* 2015; **348**: 1261359–1261359.
- 629 85. Wohl DL, Arora S, Gladstone JR. Functional redundancy supports biodiversity and ecosystem function in
630 a closed and constant environment. *Ecology* 2008; **85**: 1534–1540.
- 631 86. Kurm V, Geisen S, Gera Hol WH. A low proportion of rare bacterial taxa responds to abiotic changes
632 compared with dominant taxa. *Environ Microbiol* 2019; **21**: 750–758.

633 87. Bergkessel M, Basta DW, Newman DK. The physiology of growth arrest: Uniting molecular and
634 environmental microbiology. *Nat Rev Microbiol* . 2016. , **14**: 549–562

635 88. Hofer U. Life in the slow lane. *Nat Rev Microbiol* 2019.

636 89. Baho DL, Peter H, Tranvik LJ. Resistance and resilience of microbial communities - Temporal and spatial
637 insurance against perturbations. *Environmental Microbiology* . 2012.

638 90. Jousset A, Bienhold C, Chatzinotas A, Gallien L, Gobet A, Kurm V, et al. Where less may be more: How
639 the rare biosphere pulls ecosystems strings. *ISME J* . 2017. , **11**: 853–862

640 91. Aanderud ZT, Jones SE, Fierer N, Lennon JT. Resuscitation of the rare biosphere contributes to pulses of
641 ecosystem activity. *Front Microbiol* 2015; **6**: 1–11.

642 92. Lawson CE, Strachan BJ, Hanson NW, Hahn AS, Hall ER, Rabinowitz B, et al. Rare taxa have potential
643 to make metabolic contributions in enhanced biological phosphorus removal ecosystems. *Environ*
644 *Microbiol* 2015; **17**: 4979–4993.

645 93. Zhou J, He Z, Yang Y, Deng Y, Tringe SG, Alvarez-Cohen L. High-throughput metagenomic technologies
646 for complex microbial community analysis: Open and closed formats. *MBio* 2015; **6**: e02288-14.

647 94. Zhou J, Wu L, Deng Y, Zhi X, Jiang YH, Tu Q, et al. Reproducibility and quantitation of amplicon
648 sequencing-based detection. *ISME J* 2011; **5**: 1303–1313.

649 95. Shi Z, Yin H, Van Nostrand JD, Voordeckers JW, Tu Q, Deng Y, et al. Functional Gene Array-Based
650 Ultrasensitive and Quantitative Detection of Microbial Populations in Complex Communities. *mSystems* 2019;
651 **4**: 99–117.
652

653

654

655

656

657

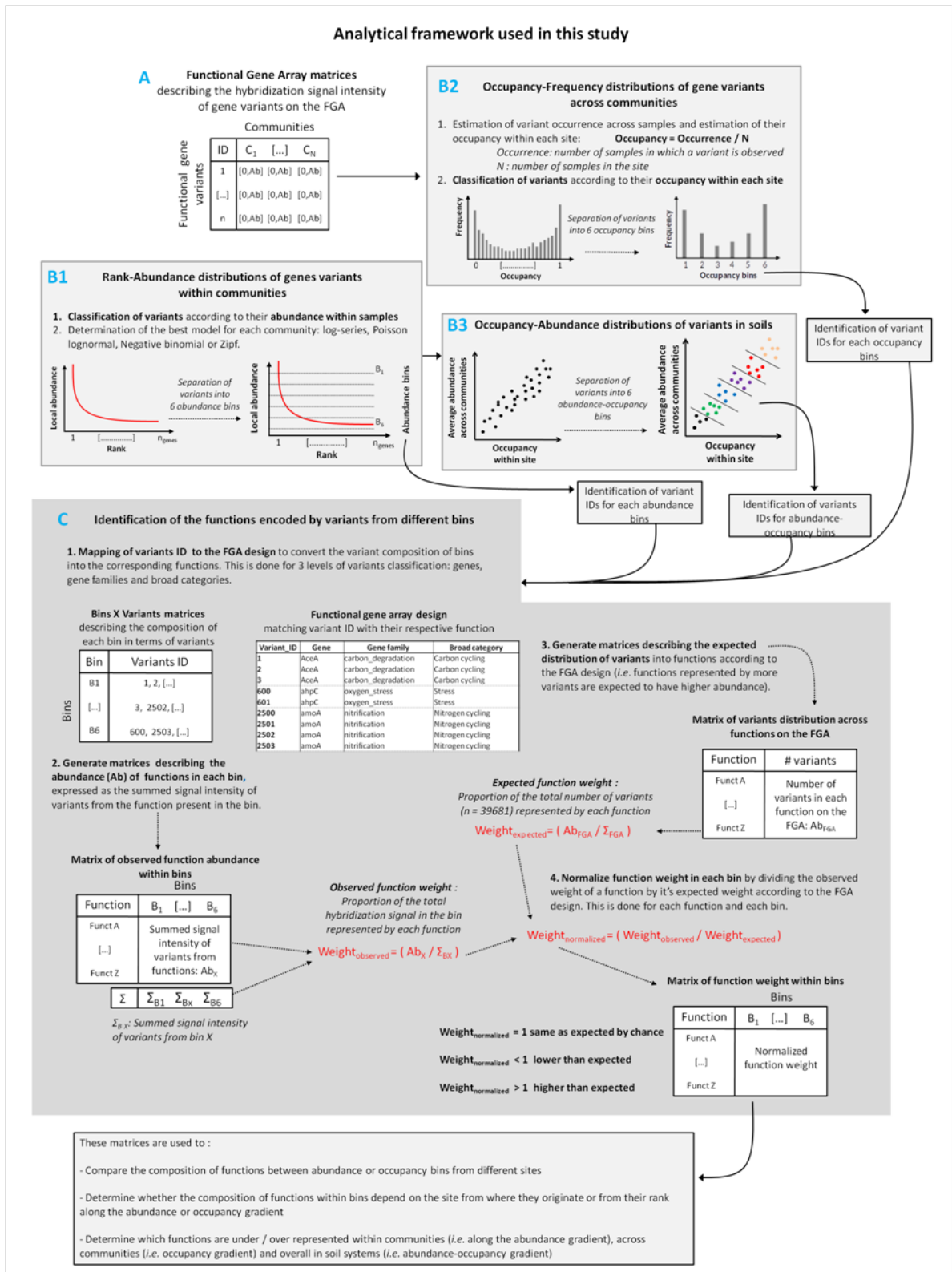
658

659

660

661

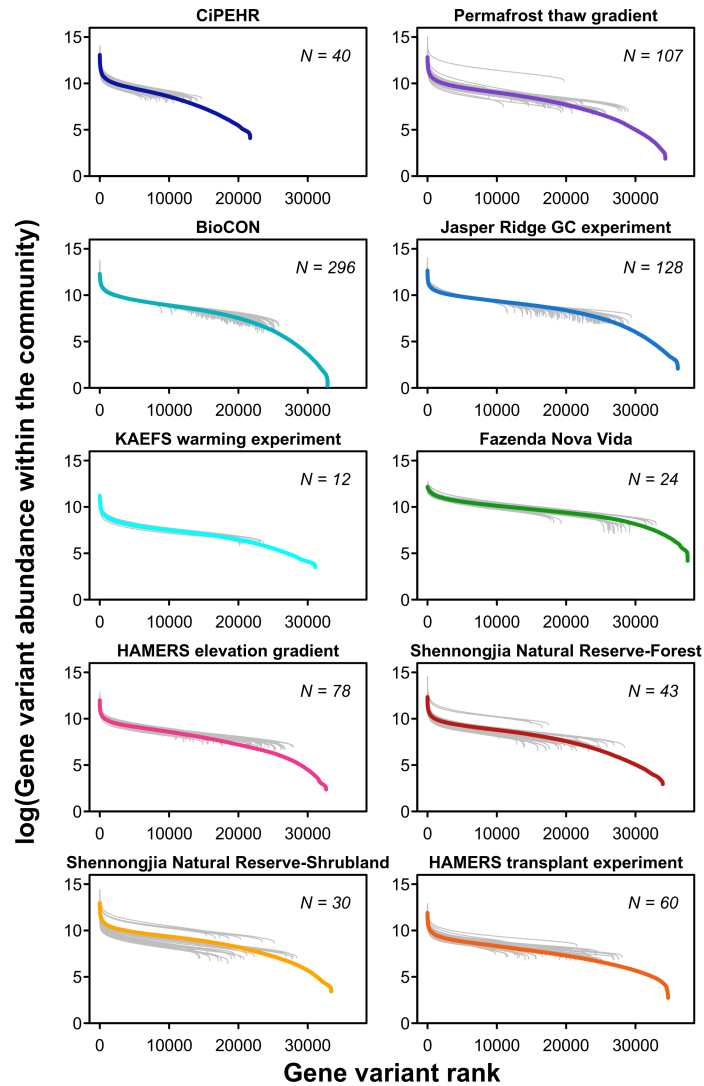
662



666

667 **Figure 2. Rank-abundance relationship of functional gene variants within each studied community.**

668 For each community (*i.e.* sample), we fitted four rank-abundance models (Logseries, Poisson lognormal,
669 Negative binomial and Zipf) using maximum likelihood estimation (MLE). Each subplot corresponds to one site
670 and each gray line represents the RAD of variants within a sample (logged hybridization signal intensity). The
671 thick lines correspond to the average model for each site.



673

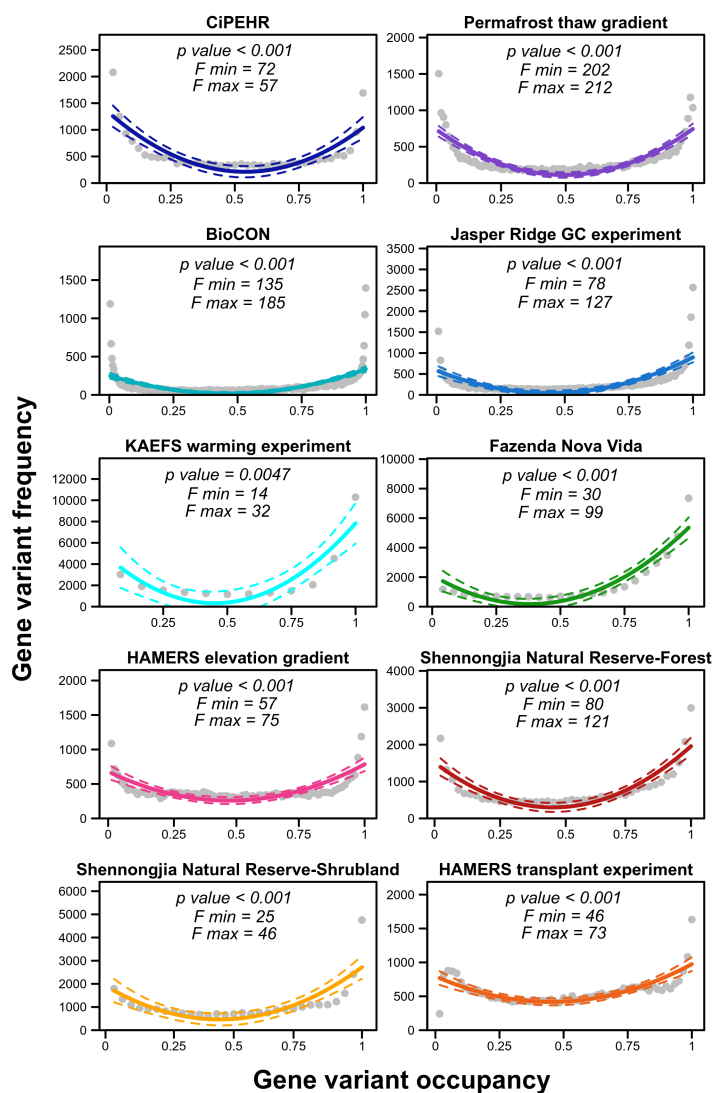
674

675

676 **Figure 3. Occupancy-frequency relationship of microbial gene variants within each studied site**

677 Colored lines correspond to the best model describing the relationship. The color of data points corresponds to
678 the colors used in Figure 1. The *p* values of the MOS test of bimodality along with the F values associated with
679 the test of the presence of local maxima at low and high occupancy are depicted.

680



682

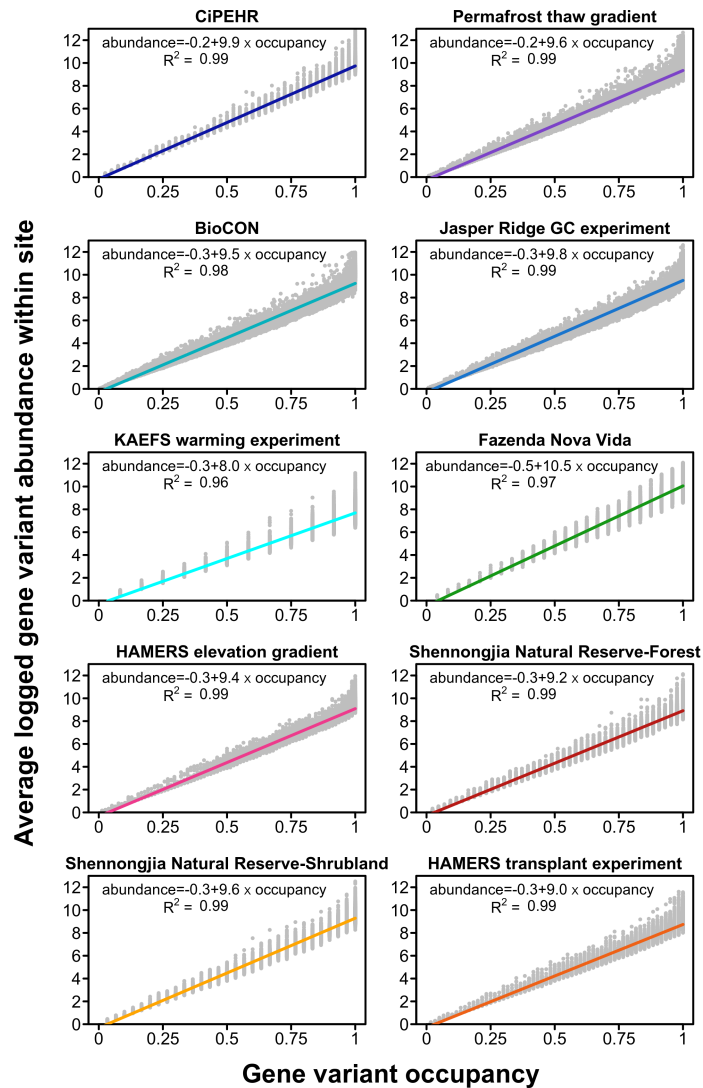
683

684

685 **Figure 4. Occupancy-abundance relationships of microbial gene variants within soil ecosystems**

686 In each site, the occupancy of the 39,681 gene variants present on the FGA was estimated as the proportion of
687 samples in which it was detected. Their abundance was estimated as the average abundance across all the
688 samples from the site. Black lines represent the best linear models describing the occupancy-abundance
689 relationship.

690



692

693

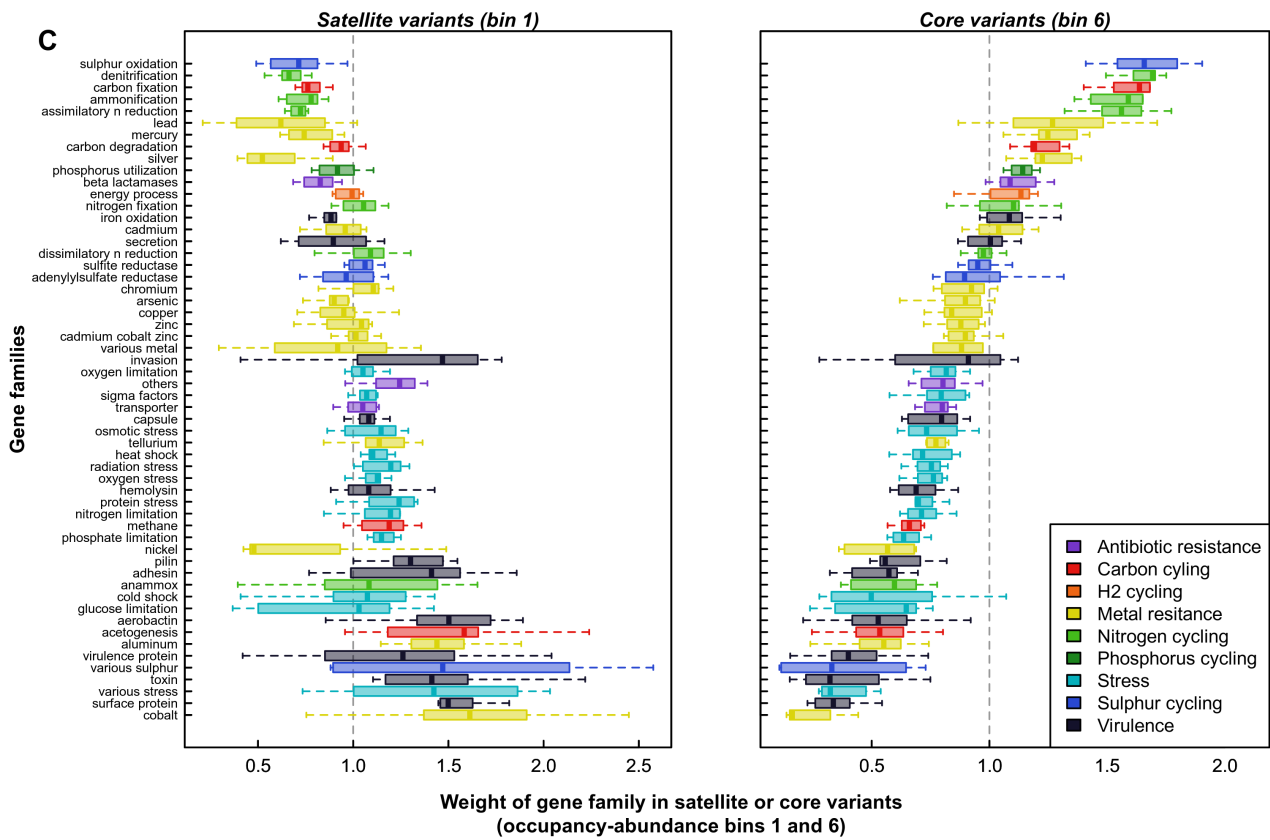
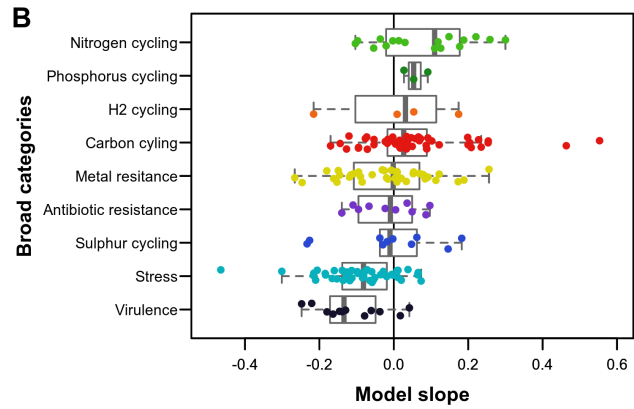
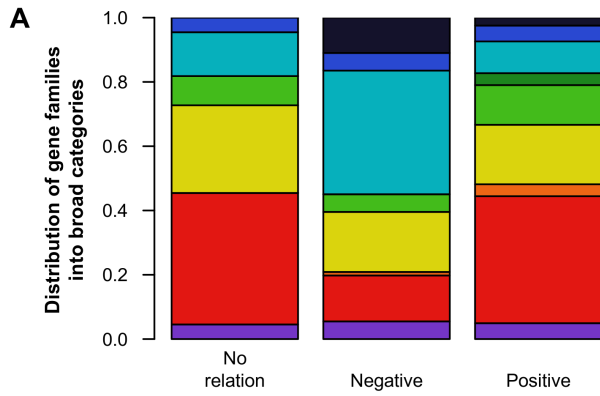
694

695 **Figure 5. Distribution of genes weight across rank of occupancy-abundance bins**

696 For each of the 194 genes, we fitted linear models describing the relationship between the weight of genes in
697 each occupancy-abundance bin and the rank of the bin (1 to 6). Negative relationships (significant negative
698 slopes) corresponded to genes over-represented in rare gene variants whereas positive ones (significant positive
699 slopes) corresponded to genes over-represented in abundant variants, when the slope of the linear model was not
700 significant the case was classified as “no relation”. (A) Relative proportions of genes across broad categories for
701 the models with non-significant (n = 22), negative (n = 91) and positive (n = 81) relationships. (B) Slopes of the
702 models classified by broad categories. (C-D) In each of the ten studied sites, satellite and core variants were
703 defined as those from the extreme occupancy-abundance bins (B_1 and B_6 , respectively). Boxplots represent the
704 weight of ecological processes in bins from each of the ten sites for satellite (C) and core (D) variants. Gene
705 families were ranked according to their average weight in core variants across the ten sites. Colors represent
706 different broad categories of functions as depicted in the legend in (D).

707

708



711

712

713

714

715

716 **Table 1 – Comparison of the functional composition of abundance and occupancy bins across sites**

717 We tested the effects of sites and rank on the composition of abundance and occupancy bins. This was performed
 718 at three levels of functional classification (broad categories, gene families and genes) using permutational
 719 multivariate analysis of variance (PERMANOVA; adonis function in the R package vegan) on Bray-Curtis
 720 dissimilarity.

721

722

723

Functional level	Factor	Df	Abundance bins				Occupancy bins					
			<i>F value</i>	$F_{\text{rank}} / F_{\text{site}}$	R^2	<i>p value</i>	<i>F value</i>	$F_{\text{rank}} / F_{\text{site}}$	R^2	<i>p value</i>		
<i>Broad categories</i>	<i>Rank</i>	5	92.90	18.6	0.84	0.001	***	14.35	13.3	0.57	0.001	***
	<i>Sites</i>	9	5.01		0.08	0.001	***	1.08		0.08	0.352	
<i>Gene families</i>	<i>Rank</i>	5	70.62	8.3	0.74	0.001	***	8.81	4.9	0.42	0.001	***
	<i>Sites</i>	9	8.53		0.16	0.001	**	1.81		0.15	0.003	**
<i>Genes</i>	<i>Rank</i>	5	28.85	8.1	0.65	0.001	***	4.77	5.4	0.31	0.001	***
	<i>Sites</i>	9	3.57		0.15	0.001	***	0.88		0.10	0.828	

725

726

727

728