esa

ECOSPHERE

METHODS, TOOLS, AND TECHNOLOGIES

Toward a unified diversity–area relationship (DAR) of species and gene diversity illustrated with the human gut metagenome

ZHANSHAN (SAM) MA [1,2,†] AND AARON M. ELLISON [3,4]

[1]*Computational Biology and Medical Ecology Lab, State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223 China*
[2]*Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223 China*
[3]*Harvard University, Harvard Forest, 324 North Main Street, Petersham, Massachusetts 01366 USA*
[4]*Sound Solutions for Sustainable Science, Boston, Massachusetts 02135 USA*

Citation: Ma, Z. S., and A. M. Ellison. 2021. Toward a unified diversity–area relationship (DAR) of species and gene diversity illustrated with the human gut metagenome. Ecosphere 12(11):e03807. 10.1002/ecs2.3807

Abstract. The biogeographic diversity of the microbiome can be investigated from two perspectives: the spatiotemporal distribution of species (or any operational taxonomic unit) diversity and the spatiotemporal distribution of metagenomic gene diversity. Together, these provide a complementary understanding of taxonomic, ecological, evolutionary, and functional aspects of the microbiome. Here, we reformulate the species diversity–area relationship (DAR) to quantify and illustrate metagenomic diversity–area relationships (*m*-DAR) with the gut metagenome from the human microbiome project (HMP). Using the *m*-DAR, the estimated ranges of human gut metagenomic genes (MG) within and among individuals are 5.0–9.7 × 10^5 and 4.3–6.9 × 10^6, respectively; the among-individual standard errors of these estimates (6.0–7.5 × 10^5) are of the same order of magnitude as the within-individual ranges, suggesting high between-individual variability. We similarly estimated the number of metagenome functional gene clusters (MFCG) to be 222–245 (SE = 1–2). More detailed analysis of the *m*-DAR profile, pair-wise diversity overlap (PDO), maximal accrual diversity (MAD), and ratio of individual- to population-level diversity (RIP) of microbiomes of individuals with healthy guts and those with three microbiome-associated diseases (obesity, diabetes, and inflammatory bowel disease) identified differences in *m*-DAR parameters between healthy and diseased individuals. Methodologically, the *m*-DAR and its associated parameters offer a unified toolset with which to study and analyze microbiomes from both species and metagenomic perspectives and to explore spatial scaling of metagenomic diversity within and among individuals. To the best of our knowledge, our illustration of *m*-DAR with the human gut metagenome is the first statistically-based estimate of the richness of human metagenomic genes at population scale.

Key words: diversity–area relationship; metagenome biogeography; metagenome diversity–area relationship; metagenome functional gene cluster; metagenomic gene; species–area relationship.

Received 17 February 2021; revised 10 June 2021; accepted 29 June 2021. Corresponding Editor: Debra P. C. Peters.

† **E-mail:** ma@vandals.uidaho.edu

INTRODUCTION

Microbiomes (e.g., the human gut microbiome) can be described within- or among-individual "macrobes" (e.g., a single individual or a population of individuals). Microbial biogeography describes the spatiotemporal distribution of microbial species (or operational taxonomic units [OTUs]; Martiny et al. 2006, Costello et al. 2012, Hanson et al. 2012, van der Gast 2013), but it also

could represent the spatiotemporal distribution of genes or metagenomes (i.e., the total genomes of all member species in a microbial community or microbiota). In either case, microbial biogeography can be described within or among individuals (Costello et al. 2012). Assuming microbial OTUs, genes, or metagenomes can be distinguished (e.g., Qin et al. 2010, Zhu et al. 2010, Li et al. 2014), it is straightforward to identify factors associated with, or effects of, microbial species richness (e.g., Martiny et al. 2006, Hanson et al. 2012, Le Chatelier et al. 2013).

Our recent work has considered additional measures of diversity of microbial OTUs, genes, and metagenomes and their relationships to diversity–stability relationships and diseases associated with changes in structure and interactions within a single individual's microbiome (Ma 2018*b*, Ma and Li 2018, Ma et al. 2019, Ma and Ellison 2018, 2019). Here, we use our recent extension of the species–area relationship (SAR; Connor and McCoy 1979) to a more general diversity–area relationship (DAR; Ma 2018*a*, *b*, 2019) to explore how metagenomic diversity changes and scales from individuals to populations (i.e., groups of individuals).

We first outline how we extend the SAR to construct a DAR for metagenomes ("*m*-DAR") for arbitrary measures of diversity (expressed as Hill numbers; Hill 1973, Chao et al. 2014). We estimate the parameters of the *m*-DAR and associated measures—pair-wise diversity overlap (PDO), maximal accrual diversity (MAD), and ratio of individual- to population-level diversity (RIP)—to characterize spatial scaling of metagenomic diversity within and among individuals. We then illustrate our approach using data on the human gut microbiome collected by the human microbiome project (HMP Consortium 2012, iHMP Consortium 2019). The human gut microbiome is a good case study because there is large individual-to-individual variation and neither the distribution of microbial species diversity nor the distribution of metagenomic diversity within a human population is homogenous (Costello et al. 2012). Finally, detailed analysis of the *m*-DAR profile, PDO, MAD, and RIP of microbial metagenomes identified within- and among-individual variation of, and markers for, individuals with healthy guts and those with three microbiome-associated diseases (obesity, diabetes, and inflammatory bowel disease).

Materials and Methods

The process for constructing *m*-DAR models consists of three steps (Fig. 1): (1) bioinformatic analysis of the metagenomic sequencing raw reads (data) to obtain metagenomic gene (MG) abundance and metagenome functional gene cluster (MFGC) tables; (2) estimation of metagenome diversity of MGs and MFGCs; and (3) fitting *m*-DAR models and constructing *m*-DAR, PDO, MAD, and RIP profiles for MGs and MFGCs.

Bioinformatic analysis of metagenomic sequencing raw reads

Metagenomic gene (MG) abundance can be estimated using standard bioinformatic software pipelines applied to whole-genome shotgun sequencing reads (Qin et al. 2010, 2012, Zhu et al. 2010, Le Chatelier et al. 2013, Li et al. 2014, Xiao et al. 2015, 2016, Wang and Jia 2016, Sczyrba et al. 2017, Ma and Li 2018). Because there often are many MGs (order of 10^6) with similar functions in a metagenome, MGs are usually grouped into a smaller number (order of 10^2) metagenomic functional gene clusters (MFGCs; Ma and Li 2018). Two commonly used databases for identifying MFGCs are the eggNOG (protein functions) and KEGG (metabolic pathways) databases (Kanehisa and Goto 2000, Huerta-Cepas et al. 2016, Kanehisa et al. 2021).

Estimation of metagenome diversity

Analogous to the estimation of species diversity of macrobial assemblages, Ma and Li (2018) used Hill numbers (Hill 1973, Jost 2007, Chao et al. 2012, 2014) to define different measures of diversity of MGs and MFGCs:

$$ {}^{q}D = \left(\sum_{i=1}^{G} p_i^q \right)^{1/(1-q)} \quad (1) $$

In Eq. 1, G is the number of MGs or MFGCs, p_i is the relative abundance of the i-th MG or MFGC, and q is the order number of diversity. When $q = 0$, ${}^{0}D = G$, the number ("richness") of MGs or MFGCs. When $q = 1$, ${}^{1}D$ is MG or MFGC diversity weighted proportionately by gene or functional gene cluster frequency.

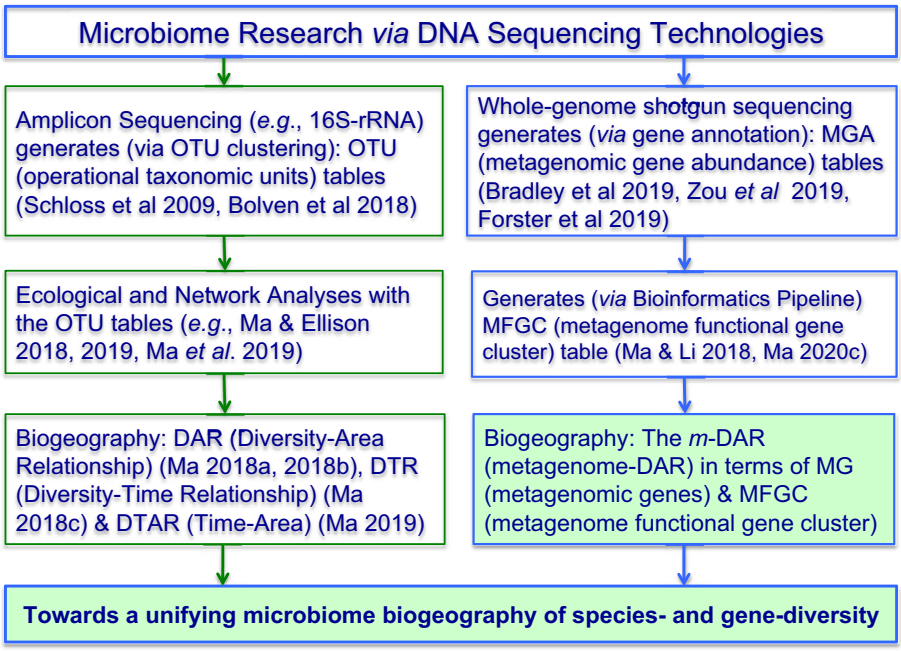

Fig. 1. Diagram showing the bioinformatic pipelines and ecological tools to establish a unifying biogeography for microbiome species and gene diversity.

When $q = 2$, ${}^2D$ is MG or MFGC diversity weighted by dominant (more abundant) genes or functional gene clusters. When $q = 3$, ${}^3D$ is weighted more heavily than ${}^2D$ by dominant genes or functional gene clusters (Ma and Li 2018).

Measures of metagenome diversity based on MGs use single genes, but metagenome diversity of MFGCs can differ depending on whether clusters are based on metabolic functions (KEGG) or protein functions (eggNOG). Ma and Li (2018) also distinguished two types of MFGC diversity, depending on how individual gene abundance information is used in defining the functional clusters. Type I MFGCs ignore the abundance of individual genes and only count the number of genes in a cluster (analogous to incidence data in macrobial species diversity calculations; e.g., Broms et al. 2015). In contrast, type II MFGCs depend on both the number of genes and their relative abundances. However, we only use ${}^0D$ in our *m*-DAR models ($q = 0$) since ${}^0D$ of type I and type II MFGCs are equivalent. This is because when $q = 0$, the abundance of gene or MFGC is not weighed in the computation of ${}^0D$.

Fitting m-DAR models and constructing m-DAR diversity profiles

Following Ma (2018*a*), who extended the classic species–area relationship (SAR) to a diversity–area relationship (DAR), we use a basic power law (PL) model to define the metagenomic diversity–area relationship (*m*-DAR):

$$ {}^qD = cA^z \quad (2)$$

In Eq. 2, ${}^qD$ is metagenome diversity (Eq. 1) of order q; A ("area") is the number of subjects whose metagenome diversity are sampled; and c and z are fitted parameters. As applied to microbial metagenomes, c can be thought of as the estimated diversity of any single individual, whereas the diversity scaling parameter z is the rate of increase in metagenome diversity with increasing number of individuals sampled.

We follow Plotkin et al. (2000) and Ulrich and Buszko (2003) in modifying Eq. 2 to include a third parameter, d:

$$ {}^qD = cA^z \exp(dA) \quad (3)$$

In this "power law with exponential cutoff" (PLEC) model, $d < 0$ and $\exp(dA)$ eventually

overwhelms the exponential function at very large values of A, leading to an asymptotic value of ${}^qD$. Using this exponential decay term makes sense because there are a finite number of people and thus a finite diversity of metagenomes.

We use log-transformed versions of Eqs. 2, 3

$$\ln(D) = \ln(c) + z\ln(A) \quad (4)$$

$$\ln(D) = \ln(c) + z\ln(A) + dA \quad (5)$$

to estimate the parameters of the models because their computation is simpler; z is scale-invariant in Eq. 4; and the ecological interpretation of z as a scaling parameter is preserved in Eq. 5. On a log–log plot, z is the slope of the linearized functions. Fitting of Eqs. 4, 5 to the data was evaluated using the linear correlation coefficients (r) and associated P values.

There is a notable difference between the human microbiome and assemblages of macrobes or microbes in "natural" biomes. In the latter, we can usually identify a natural spatial order or environmental gradient among plants, animals, soil strata, etc. But there is not a similar natural ordering among humans (the hosts of human microbiomes). To deal with this absence of spatial ordering among sampled human subjects, we first enumerated all possible permutations of the ordering of sample subjects. We then randomly selected 50 (for MG) or 100 (for MFGC) orderings and fit the m-DAR models (Eqs. 4, 5) for each of permutations. We next eliminated those few poorly fitting models ($P > 0.05$) and, for PLEC models (Eq. 5), any model with $A_{\max} < 0$ (which is biologically infeasible). Finally, we used the averages of the estimated parameters from each of the remaining models from the permutated data sets as the estimated parameters in the overall m-DAR models.

The m-DAR profile

The relationship between diversity order q and the diversity scaling parameter z from Eq. 2 is defined as the m-DAR profile, similar to that for species diversity (Ma 2018*a*).

Metagenomic maximal accrual diversity of metagenome

Ma (2018*a*) derived the maximal accrual diversity (MAD) in a cohort or population based on the DAR-PLEC model (Eq. 3) as:

$${}^qD_{\max} = c\left(-\frac{z}{d}\right)^z \exp(-z) = cA_{\max}^z \exp(-z) \quad (6)$$

for which the number of individuals ($A_{\max}$) reaching the maximum diversity ($D_{\max}$) is estimated as:

$$A_{\max} = -z/d \quad (7)$$

We then define the m-MAD profile as the set of $D_{\max}$ values corresponding to different diversity orders q. ${}^qD_{\max}$ can be interpreted as a proxy for the so-called potential ("dark") diversity: species (OTUs) or metagenomes (MG or MFGC) that are absent locally but present in regional or global species pools (Partel et al. 2011, Real et al. 2017, Ma 2019).

Pair-wise diversity overlap

Because of the assumption that "areas" of different sampled individuals are approximately equal, the parameter z of the basic m-DAR model (Eqs. 2 or 4) can be used to estimate the pair-wise diversity overlap (PDO). The PDO, g, of two individuals (i.e., the proportion of the new diversity in the second area) is then:

$$g = 2 - 2^z \quad (8)$$

where z is the scaling parameter of basic m-DAR model (Eqs. 2, 4). g can take on values from 0 (no overlap; $z = 1$) to 1 (complete overlap; $z = 0$).

The m-PDO profile is the set of g values corresponding to different diversity orders (q). It approximates the similarity between a pair of human metagenomes.

The ratio of individual to population diversity

We define the ratio of individual to population accrual diversity as:

$${}^q\text{RIP} = {}^qc/{}^qD \quad (9)$$

where c and ${}^qD$ are estimated from Eq. 2 (or 4). We further define the set of ^{q}RIP values corresponding to different diversity orders q as the RIP profile.

The RIP profile can be derived from a population (i.e., sample cohort) of any size. In practice, using ${}^qD_{\max}$ in place of ${}^qD$ in Eq. 9 is more informative:

$${}^q\text{RIP} = {}^qc/{}^qD_{\max} \quad (10)$$

^{q}RIP in Eq. 10 represents the average diversity level an individual can represent in the sampled cohort. This representation is based on two assumptions. First, the sizes of sampled individuals are approximately equal. Second, the first subject sampled will not exert undue influence on the estimation of c. The first assumption is largely true for the human microbiome. However, the second assumption may not hold because human microbiomes are very variable in composition (Ma 2018*a*, 2019). To account for this compositional variability, we used the random permutation followed by averaging method described above for estimating parameters for Eqs. 4, 5. If the data sets do not allow for reliable estimation of the MAD (e.g., the PLEC model [Eq. 3, 5] fails to fit), we suggest using the predicted values from Eq. 2 in Eq. 9 to estimate ^{q}RIP.

Illustrative metagenome data sets

We used six metagenomic data sets of varying sample sizes (n = 47–168) to illustrate the application and interpretation of m-DAR models and associated profiles of MAD, PDO, and RIP as a function of q (Table 1). The six data sets consist of three pairs of healthy *vs*. diseased groups, including lean vs. overweight, healthy control vs. type II diabetes, and healthy vs. IBD (inflammatory bowel disease). These metagenomic data sets were collected using standard protocols of the human microbiome project (HMP) and have been successfully used to demonstrate the extension of ecological diversity in Hill numbers to metagenome diversity and heterogeneity (Ma and Li 2018, Ma 2020*c*). Hence, it is natural to demonstrate the scaling of metagenome diversity (m-DAR), with the same data sets in the present article. To maintain balanced sample sizes between the healthy and diseased treatments (groups), in some cases we randomly discarded certain number of samples so that the model parameters could be compared appropriately between the healthy and diseased treatments. All the metagenomic data sets are available in the public domain (sources in Table 1) and the computational codes used in this study consist of two parts, the code for computing metagenome diversity in Hill numbers was published in Ma and Li (2018), and the code for fitting m-DAR models is essentially the same as that used for fitting the DAR of OTUs (Ma 2018*a*, *b*), except for slight revision in the format of input data.

Results

Metagenomic diversity–area relationships

Both the basic m-DAR power law (PL) model (Eqs. 2, 4) and the PLEC model (Eqs. 3, 5) fit the gut metagenome data well ($P < 0.001$; Table 2). The scaling parameter (z) in Eq. 2 for metagenomic genes (MGs) ranged from 0.358 to 0.468 for diversity order $q = 0$; 0.264–0.363 for $q = 1$; 0.221–0.334 for $q = 2$; and 0.190–0.346 for $q = 3$ (Fig. 2). Estimates of z from the PLEC model

Table 1. Brief information on the metagenome data sets used to demonstrate m-DAR models.

Disease	Treatments	No. of samples	No. of samples selected for m-DAR	No. of genes†	No. of MFGC (eggNOG)†	No. of MFGC (KEGG)†	Reference
Obesity	Lean	96	96	973,838 (26,969)	245 (2)	174 (2)	Qin et al. (2010); Chatelier et al. (2013)
Obesity	Overweight	168	Randomly sampled 96	937,187 (24,350)	239 (2)	171 (2)	Qin et al. (2010); Chatelier et al. (2013)
Type II Diabetes	Healthy	74	74	523,793 (17,829)	223 (1)	144 (2)	Qin et al. (2012)
Type II Diabetes	Diseased	71	71	504,712 (23,219)	222 (2)	141 (3)	Qin et al. (2012)
IBD	Healthy	24	71	962,413 (30,600)	245 (2)	174 (2)	Nielsen et al. (2014)
IBD	Healthy Relative	47	71	962,413 (30,600)	245 (2)	174 (2)	Nielsen et al. (2014)
IBD	UC	127	Randomly sampled 71	838,009 (24,347)	236 (2)	167 (2)	Nielsen et al. (2014)

† Values are expressed as mean with SE in parentheses.

Table 2. Parameters of *m*-DAR (metagenome diversity–area relationship) models fitted for metagenomic gene (MG) diversity, averaged from 50 times of re-sampling.

Treatment by study case and diversity order	m-DAR PL (power law) Model						m-DAR PLEC (power law with exponential cutoff) Model							
	z	ln(c)	g	R	p	N^*	z	d	ln(c)	R	p	A_{max}	D_{max}	N^*
$q = 0$														
Obesity														
Lean	0.358	14.224	0.718	0.975	0.000	50	0.501	−0.005	13.969	0.994	0.000	94	6871100	50
Overweight	0.367	14.146	0.711	0.975	0.000	50	0.509	−0.005	13.890	0.993	0.000	96	6629101	50
Type II diabetes														
Healthy	0.435	13.503	0.648	0.982	0.000	50	0.577	−0.007	13.279	0.994	0.000	86	4305968	50
Disease	0.468	13.439	0.617	0.977	0.000	50	0.617	−0.007	13.209	0.990	0.000	85	4554045	50
IBD														
Healthy	0.376	14.110	0.702	0.976	0.000	50	0.525	−0.007	13.879	0.994	0.000	72	5951880	50
Disease	0.392	14.013	0.688	0.978	0.000	50	0.543	−0.007	13.781	0.995	0.000	74	5792650	50
Mean of HEA (healthy)	0.390	13.946	0.689	0.977	0.000	50	0.534	−0.006	13.709	0.994	0.000	84	5709650	50
SE of HEA	0.023	0.224	0.021	0.002	0.000	0	0.023	0.001	0.216	0.000	0.000	7	750330	0
Mean of DIS (diseased)	0.409	13.866	0.672	0.977	0.000	50	0.556	−0.007	13.627	0.993	0.000	85	5658598	50
SE of DIS	0.030	0.217	0.028	0.001	0.000	0	0.032	0.001	0.211	0.001	0.000	7	602755	0
$q = 1$														
Obesity														
Lean	0.325	12.432	0.748	0.954	0.000	50	0.474	−0.006	12.165	0.981	0.000	85	980004.2	50
Overweight	0.301	12.375	0.768	0.930	0.000	50	0.478	−0.007	12.057	0.972	0.000	73	831126.0	49
Type II diabetes														
Healthy	0.321	11.721	0.751	0.941	0.000	50	0.489	−0.008	11.461	0.973	0.000	61	435659.0	49
Disease	0.363	11.679	0.714	0.940	0.000	50	0.534	−0.008	11.418	0.966	0.000	64	491949.4	46
IBD														
Healthy	0.264	12.514	0.799	0.920	0.000	50	0.433	−0.008	12.251	0.966	0.000	53	757500.7	48
Disease	0.337	12.316	0.736	0.947	0.000	50	0.534	−0.009	12.007	0.982	0.000	57	831436.9	47
Mean of HEA (healthy)	0.303	12.222	0.766	0.938	0.000	50	0.465	−0.007	11.959	0.974	0.000	66	724388.0	49
SE of HEA	0.020	0.252	0.017	0.010	0.000	0	0.017	0.001	0.250	0.004	0.000	9	158008.7	1
Mean of DIS (diseased)	0.334	12.123	0.739	0.939	0.000	50	0.516	−0.008	11.827	0.973	0.000	65	718170.8	47
SE of DIS	0.018	0.223	0.016	0.005	0.000	0	0.019	0.001	0.205	0.005	0.000	5	113110.7	1
$q = 2$														
Obesity														
Lean	0.325	11.061	0.747	0.908	0.000	50	0.527	−0.007	10.699	0.956	0.000	71	247026.3	47
Overweight	0.265	11.063	0.798	0.811	0.000	50	0.501	−0.009	10.641	0.903	0.000	58	193637.8	45
Type II diabetes														
Healthy	0.244	10.549	0.816	0.856	0.000	50	0.437	−0.009	10.248	0.922	0.000	49	99610.3	46
Disease	0.317	10.408	0.755	0.871	0.000	50	0.528	−0.010	10.096	0.916	0.000	51	113517.0	44
IBD														
Healthy	0.221	11.266	0.835	0.820	0.000	50	0.437	−0.010	10.936	0.914	0.000	42	186203.9	44
Disease	0.334	10.950	0.740	0.890	0.000	50	0.579	−0.012	10.563	0.954	0.000	50	208940.3	45
Mean of HEA (healthy)	0.263	10.958	0.799	0.862	0.000	50	0.467	−0.009	10.628	0.931	0.000	54	177613.5	46
SE of HEA	0.032	0.213	0.027	0.026	0.000	0	0.030	0.001	0.202	0.013	0.000	9	42771.5	1
Mean of DIS (diseased)	0.305	10.807	0.764	0.858	0.000	50	0.536	−0.010	10.433	0.924	0.000	53	172031.7	45
SE of DIS	0.021	0.202	0.018	0.024	0.000	0	0.023	0.001	0.170	0.015	0.000	3	29588.9	0
$q = 3$														
Obesity														
Lean	0.346	9.945	0.729	0.879	0.000	50	0.561	−0.008	9.552	0.925	0.000	72	88601.1	46
Overweight	0.273	9.934	0.791	0.760	0.000	48	0.487	−0.008	9.578	0.847	0.000	59	64918.8	43

(Table 2. Continued.)

Treatment by study case and diversity order	m-DAR PL (power law) Model						m-DAR PLEC (power law with exponential cutoff) Model							
	z	ln(c)	g	R	p	N^*	z	d	ln(c)	R	p	A_{max}	D_{max}	N^*
Type II diabetes														
Healthy	0.220	9.669	0.835	0.804	0.000	50	0.404	−0.009	9.379	0.883	0.000	48	37648.9	46
Disease	0.319	9.350	0.752	0.809	0.001	49	0.505	−0.010	9.111	0.840	0.001	52	40409.1	39
IBD														
Healthy	0.190	10.088	0.859	0.729	0.001	46	0.382	−0.010	9.831	0.825	0.000	39	51321.6	41
Disease	0.294	9.676	0.774	0.760	0.001	46	0.523	−0.012	9.369	0.827	0.000	45	50731.4	41
Mean of HEA (healthy)	0.252	9.900	0.808	0.804	0.000	49	0.449	−0.009	9.587	0.878	0.000	53	59190.5	44
SE of HEA	0.048	0.123	0.040	0.043	0.000	1	0.056	0.001	0.132	0.029	0.000	10	15225.8	2
Mean of DIS (diseased)	0.295	9.653	0.773	0.776	0.001	48	0.505	−0.010	9.353	0.838	0.000	52	52019.8	41
SE of DIS	0.013	0.169	0.011	0.016	0.000	1	0.010	0.001	0.135	0.006	0.000	4	7104.6	1

Note: *N = The number of times (number of random re-samplings) that the m-DAR model was successfully fitted.

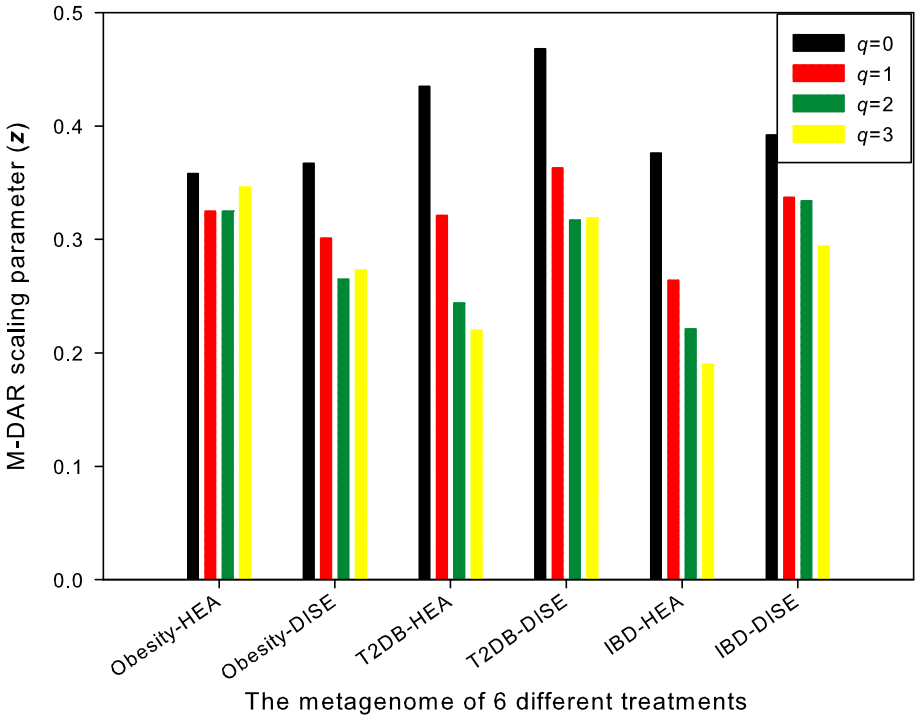

Fig. 2. Scaling parameter (z) of the m-DAR (metagenome diversity–area relationship) for the metagenomic genes (MGs) of the human gut metagenome.

(Eqs. 3, 5) were about 25% larger than those of the basic PL model (Table 2). Regardless of model, the differences in z between the healthy and diseased individuals were comparatively small, with the z-value being slightly, but not significantly higher ($P > 0.05$) for all values of q (Appendix S1: Table S1). Ma and Li (2018) provide additional estimates of metagenome diversity (^qD in Eq. 1).

The PLEC model did suggest an asymptote for MG diversity ($d < 0$ and D_{max} values in Table 2) and a corresponding number of

individual subjects (A_{max} in Table 2) needed to reach the asymptote. D_{max}, the maximal accrual diversity (MAD) of MGs in the gut metagenome, ranged from 4.3 to 6.8×10^6 (Table 2; Fig. 3). These estimates of gene richness at the population level are an order of magnitude greater than the average gene richness per individual (Table 1). The standard errors of the MAD of MGs in the gut metagenome (Table 2) are of the same order of magnitude as the average gene richness per individual (Table 1). That is, the amount of population-level variation of MAD is close to the total number of metagenomic genes within an individual and represents an exceptionally high inter-subject heterogeneity. To the best of our knowledge, no other approaches are available for estimating the aforementioned parameters. The parameter c ranged from $\approx 1 \times 10^4$ to 1.5×10^6 and is a rough estimate of the number of metagenomic genes (MG) in one individual ($A = 1$).

As would be expected, the pair-wise diversity overlap (PDO) profile exhibited the opposite pattern of the m-DAR profile (Table 2). This is because z in the m-DAR profile quantifies the dissimilarity of neighboring individuals whereas g in the PDO profile quantifies the overlap or similarity between individuals.

Metagenome functional gene cluster diversity–area relationships

The basic m-DAR PL model successfully fit all MFGC randomizations ($n = 100$) for $q = 0$, but not more than 80% of the randomizations for $q > 0$. Thus, we applied our m-DAR models for MFGCs only for MFGC richness (i.e., $q = 0$).

Overall, the z values for MFGCs were ≈80% smaller than those estimated for MGs (compare values in Table 3 with those in Table 2) and, again, were higher when estimated using PLEC models than when using PL models (Table 3). z values were higher for MFGCs derived from KEGG databases (Table 3); for MFGCs derived from either KEGG or eggNOG databases, the mean z was slightly but not significantly higher among diseased individuals (Appendix S1: Table S2). Similarly, D_{max} did not differ among individuals in the different diagnostic groups

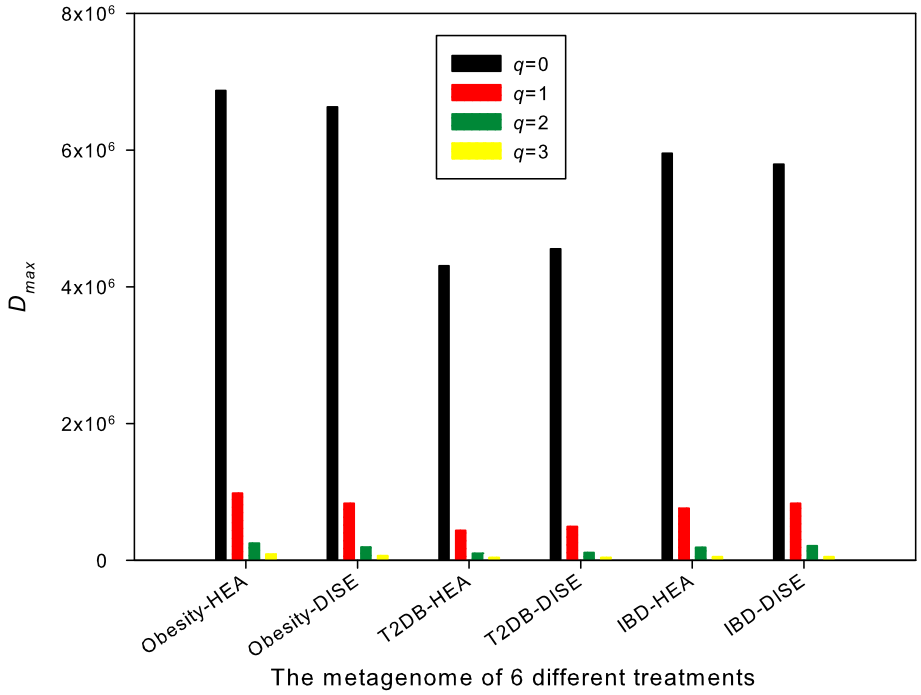

Fig. 3. MAD (maximum accrual diversity: D_{max}) of the m-DAR (metagenome diversity–area relationship) for the MG (metagenomic genes) of the human gut metagenome.

Table 3. Parameters of *m*-DAR (metagenome diversity–area relationship) models fitted for MFGC (metagenome functional gene cluster) diversity, averaged from 100 times of re-sampling.

Treatment by study case and diversity order	PL (power law) model						PLEC (power law with exponential cutoff) model							
	z	ln(c)	g	R	p	N	z	d	ln(c)	R	p	A_{max}	D_{max}	N
MFGC (eggNOG)														
$q = 0$														
Obesity														
Lean	0.060	5.615	0.957	0.929	0.000	100	0.100	−0.001	5.545	0.977	0.000	67	352	100
Overweight	0.068	5.577	0.951	0.951	0.000	100	0.103	−0.001	5.516	0.982	0.000	80	352	100
Type II diabetes														
Healthy	0.079	5.466	0.943	0.973	0.000	100	0.104	−0.001	5.427	0.987	0.000	89	327	100
Disease	0.075	5.462	0.947	0.972	0.000	100	0.100	−0.001	5.423	0.988	0.000	79	318	100
IBD														
Healthy	0.064	5.599	0.954	0.939	0.000	100	0.104	−0.002	5.538	0.981	0.000	53	346	100
Disease	0.079	5.540	0.944	0.958	0.000	100	0.115	−0.002	5.484	0.983	0.000	65	347	100
Mean of HEA (healthy)	0.068	5.560	0.952	0.947	0.000	100	0.103	−0.002	5.503	0.982	0.000	70	342	100
SE of HEA	0.006	0.047	0.004	0.013	0.000	0	0.001	0.000	0.038	0.003	0.000	10	8	0
Mean of DIS (diseased)	0.074	5.526	0.947	0.960	0.000	100	0.106	−0.001	5.475	0.984	0.000	75	339	100
SE of DIS	0.003	0.034	0.002	0.006	0.000	0	0.005	0.000	0.027	0.002	0.000	5	11	0
MFGC (KEGG)														
$q = 0$														
Obesity														
Lean	0.085	5.293	0.939	0.950	0.000	100	0.130	−0.002	5.213	0.983	0.000	77	284	100
Overweight	0.087	5.270	0.938	0.965	0.000	100	0.125	−0.001	5.201	0.989	0.000	88	280	100
Type II diabetes														
Healthy	0.108	5.097	0.922	0.966	0.000	100	0.148	−0.002	5.033	0.983	0.000	77	252	100
Disease	0.116	5.093	0.916	0.964	0.000	100	0.161	−0.002	5.025	0.982	0.000	73	258	100
IBD														
Healthy	0.094	5.252	0.933	0.967	0.000	100	0.129	−0.002	5.198	0.984	0.000	75	277	100
Disease	0.098	5.225	0.930	0.969	0.000	100	0.139	−0.002	5.163	0.990	0.000	69	273	100
Mean of HEA (healthy)	0.096	5.214	0.931	0.961	0.000	100	0.136	−0.002	5.148	0.983	0.000	76	271	100
SE of HEA	0.007	0.060	0.005	0.005	0.000	0	0.006	0.000	0.058	0.000	0.000	1	10	0
Mean of DIS (diseased)	0.100	5.196	0.928	0.966	0.000	100	0.142	−0.002	5.130	0.987	0.000	76	271	100
SE of DIS	0.008	0.053	0.006	0.001	0.000	0	0.010	0.000	0.054	0.002	0.000	6	7	0

($P > 0.05$; Appendix S1: Table S2, Fig. S2). Finally, The PDO for MFGCs was much larger than that of MGs (Appendix S1: Fig. S1, Table 2). The high PDO observed for MFGCs further illustrated the high functional similarity among individuals.

The ratio of individual diversity to population accrual diversity

The average individual has 20–30% of the MGs and ≈ 65% of the MFGCs of the whole population (Table 4). The RIP—the ratio of the average individual diversity (c in the m-DAR PL model) to the accrued diversity of the sampled population (or cohort) (D_{max})—was slightly but not significantly ($P > 0.05$) larger for the healthy individuals than the diseased ones. The RIP also increased with q, suggesting that for higher orders of diversity, any individual is more representative of the population to which they belong. Note that the parameter c in the m-DAR PL model (but not in the PLEC model) was used to define the RIP since, strictly speaking, the c from PL model is an estimation of the diversity of *one* area size. The parameter c estimated for the PLEC model co-varies with the third parameter d and is not an unbiased estimation of the diversity of one individual (Area = 1).

Table 4. RIP (the ratio of individual diversity to population MAD), averaged from the RIP of the three case studies on the human gut metagenome (converted to percentage).

	MG		MFGC	
Diversity order (q)	Average RIP (%) for healthy	Average RIP (%) for diseased	Average RIP (%) for healthy	Average RIP (%) for diseased
$q = 0$	20.0	18.6	67.8	66.6
$q = 1$	28.1	25.6	N/A	N/A
$q = 2$	32.3	28.7	N/A	N/A
$q = 3$	33.7	29.9	N/A	N/A

Discussion

Current assessments of biodiversity have generally ignored the diversity of genomes or metagenomes for at least two reasons. First, the tools for identifying (meta)genomes have been developed only recently and have been readily accessible only for the last two decades. Second, theories of biodiversity and biogeography developed for species have not yet been extended to use metagenomic data or included a metagenomic dimension of biogeography. Boon et al. (2013) identified a central challenge in microbial community ecology: "the delineation of appropriate units of biodiversity, which can be taxonomic, phylogenetic, or functional in nature." The *m*-DAR we have developed provides a first step toward the theoretical integration of metagenomic data into models of biodiversity.

In addition to developing *m*-DAR models (Eqs. 2–5) for describing how metagenomic genes (MGs) and functional gene clusters (MFGCs) scale with area (≡ number of individuals sampled), we defined four metagenome profiles for characterizing the biogeography of microbial metagenomes based on the parameters of the *m*-DAR models. We used these models and profiles to explore the diversity of metagenomic data sets of human gut microbiomes.

The estimates of z in the *m*-DAR power law model for these gut microbiome data sets (Table 2) for $q = 0$ are similar to estimates from models of the species-DAR for the gut microbiome reported in Ma (2018*a*). That is, the scaling rates (z) of gene richness and species richness (the carriers of those genes) are close to one another, but the scaling rates of diversity ($q > 0$) for MG are much larger than those of individual species. Indeed, the potential total number of MGs ($q = 0$) is nearly seven million. Among these, about 15% (≈1 million) are commonly recovered (i.e., for $q = 1$) and 3% are dominant. Despite known relationships between microbiome OTU diversity and disease (Ma et al. 2019, Ma 2020*a*, *b*), we found no significant influence of the disease on the scaling of MG diversity. However, the values of c reported here for MGs are several orders of magnitudes larger than those of microbial species (Ma 2018*a*, *b*). This is expected given that the gene richness and diversity are orders of magnitudes higher than species richness and diversity. Although we used pairs of healthy vs. diseased groups, the focus of this study was to demonstrate the feasibility of a unified DAR for species and gene diversity (or ecological and genetic diversity), rather than to determine the difference in DAR parameters between healthy and diseased individuals.

Although we could explore *m*-DAR models for MG diversity of different orders (up to $q = 3$), we successfully fit *m*-DAR models for MFGCs only for $q = 0$. Because an MFGC is a functional cluster of MGs, we would expect MFGCs to be much more homogenous than MGs among individuals. Thus, the number of MFGCs (i.e., MFGC richness) may be a sufficiently informative measure of metagenomic functional diversity. If this is true, *m*-DAR modeling of MFGC should be limited to the richness of the MFGC (i.e., $q = 0$).

If we limit the diversity order q for MFGC to $q = 0$, we might better describe its accrual as "maximum accrual richness" (MAR), not maximum accrual diversity. To the best of our knowledge, this has not been estimated previously. We observed that the MAR of MFGC (in a sampled population or cohort) is close to the average number of MFGCs of an individual metagenome. This supports other work (Ma and Li 2018, Ma 2020*c*) that found high similarity (homogeneity) among individuals in number of observed MFGCs.

Zhou et al. (2008) used the traditional species–area relationship (SAR) to estimate the number of metagenomic functional genes in a forest soil microbiome. Perhaps restricted by then available sequencing technology a decade ago, their study was limited to functional gene clusters without

measuring the metagenomic genes. Their estimate of z, ranging from 0.048 to 0.096, is similar to the estimate of z we found for the MFGC of the human gut microbiome (Table 3). Future work on other microbiomes are needed to determine the generality of this result.

Another area of future work would be to assess the diversity–time relationship (Ma 2018*c*) for metagenomes. Such an extension would require long time series of metagenomic data from single individuals or localities, which do not yet exist. As whole-genome sequencing costs continue to fall, however, such data should become available and it should become feasible to demonstrate the spatial–temporal changes in metagenomes.

Our work has certain limitations. First, successful estimation of the asymptote of the PLEC model depends on the sign of asymptotic parameter d (i.e., $d < 0$). Although this could be resolved by using non-linear optimization methods to fit the PLEC models (e.g., Ma 2020*d*) while manually imposing the constraint that $d < 0$, this constraint may change the fitting of the other parameters. In particular, it would be expected that the values of z in both the PL and PLEC models should have comparable ranges, but this cannot be guaranteed with non-linear optimization algorithms. Further, non-linear optimization can be far more computationally intensive than linear regression; given the sheer number of metagenomic genes, even linear regression is surprisingly computationally intensive.

Second, we still lack definitive guidance in selecting a proper DAR model. Indeed, there are more than a dozen alternative SAR models (Tjørve 2003, 2009), any of which could be used to model the m-DAR. We used the PLEC model primarily because of its asymptotic properties, which have been recognized previously (e.g., Plotkin et al. 2000, Ulrich and Buszko 2003) and derived explicitly in Ma (2018*a*).

Third, our approach did not consider the issue of sample completeness (Chao et al. 2020) to avoid the added complexity from accumulation of additional samples in DAR modeling. Additional research is needed on the influence of sample completeness on the accumulation of errors associated with the sample accumulation.

Nonparametric approaches, such as those used by Chao and Jost (2015) and Chao et al. (2020) to estimate macrobial biodiversity, may help resolve the three limitations we have identified so far. Finally, a recent advance in coupling the PLEC model to Taylor's power law of variance-mean could be used to estimate confidence intervals for estimates of metagenome diversity and potential diversity ($D_{\max}$) (Ma 2021).

Acknowledgments

This study was supported by a National Natural Science Foundation (NSFC) Grant (no. 31970116); the Cloud-Ridge Biotech Industry Leader award; and a China-US Collaborative Project on Genomics/Metagenomics Big Data. We thank two anonymous reviewers for their constructive comments on the initial submission. Z. S. Ma designed and conducted the analysis, interpreted the results, and wrote the manuscript. A. M. Ellison interpreted the results and revised the manuscript. Both authors approved the submission.

Literature Cited

Boon, E., C. J. Meehan, C. Whidden, D. H. J. Wong, M. G. I. Langille, and R. G. Beiko. 2013. Interactions in the microbiome: communities of organisms and communities of genes. FEMS Microbiology Reviews 38:90–118.

Broms, K. M., M. B. Hooten, and R. M. Fizpatrick. 2015. Accounting for imperfect detection in Hill numbers for biodiversity studies. Methods in Ecology and Evolution 6:99–108.

Chao, A., et al. 2020. Quantifying sample completeness and comparing diversities among assemblages. Ecological Research 35:292–314.

Chao, A., C. H. Chiu, and T. C. Hsieh. 2012. Proposing a resolution to debates on diversity partitioning. Ecology 93:2037–2051.

Chao, A., C. H. Chiu, and L. Jost. 2014. Unifying species diversity, phylogenetic diversity, functional diversity and related similarity and differentiation measures through Hill numbers. Annual Reviews of Ecology, Evolution, and Systematics 45:297–324.

Chao, A., and L. Jost. 2015. Estimating diversity and entropy profiles via discovery rates of new species. Methods in Ecology and Evolution 6:873–882.

Connor, E. F., and E. D. McCoy. 1979. The statistics and biology of the species–area relationship. American Naturalist 113:791–833.

Costello, E. K., K. Stagaman, L. Dethlefsen, B. J. M. Bohannan, and D. A. Relman. 2012. The application of ecological theory toward an understanding of the human microbiome. Science 336:1255–1262.

Hanson, C. A., J. A. Fuhrman, M. Claire Horner-Devine, and J. B. H. Martiny. 2012. Beyond biogeographic patterns: process shaping the microbial landscape. Nature Reviews Microbiology 10:497–506.

Hill, M. O. 1973. Diversity and evenness: a unifying notation and its consequences. Ecology 54:427–432.

HMP Consortium. 2012. Structure, function and diversity of the healthy human microbiome. Nature 486:207–214.

Huerta-Cepas, J., et al. 2016. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Research 44:D286–D293.

Jost, L. 2007. Partitioning diversity into independent alpha and beta components. Ecology 88:2427–2439.

Kanehisa, M., M. Furumichi, Y. Sato, M. Ishiguro-Watanabe, and M. Tanabe 2021. KEGG: integrating viruses and cellular organisms. Nucleic Acids Research 49:D545–D551.

Kanehisa, M., and S. Goto. 2000. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Research 28:27–30.

Le Chatelier, E., et al. 2013. Richness of human gut microbiome correlates with metabolic markers. Nature 500:541–546.

Li, J., et al. 2014. An integrated catalog of reference genes in the human gut microbiome. Nature Biotechnology 32:834–841.

Ma, Z. S. 2018*a*. Extending species-area relationships (SAR) to diversity-area relationships (DAR). Ecology and Evolution 8:10023–10038.

Ma, Z. S. 2018*b*. Sketching the human microbiome biogeography with DAR (diversity area relationship) profiles. Microbial Ecology 77:821–838.

Ma, Z. S. 2018*c*. Diversity time-period and diversity-time-area relationships exemplified by the human microbiome. Scientific Reports 8:7214.

Ma, Z. S. 2019. A new DTAR (diversity–time–area relationship) model demonstrated with the indoor microbiome. Journal of Biogeography 46:2024–2041.

Ma, Z. S. 2020*a*. Critical network structures and medical ecology mechanisms underlying human microbiome-associated diseases. iScience 23:101195.

Ma, Z. S. 2020*b*. Testing the Anna Karenina Principle in human microbiome associated diseases. iScience 23:101007.

Ma, Z. S. 2020*c*. Assessing and interpreting the metagenome heterogeneity with power law. Frontiers in Microbiology 11:648.

Ma, Z. S. 2020*d*. Predicting the outbreak risks and inflection points of COVID-19 pandemic with classic ecological theories. Advanced Science. https://doi.org/10.1002/advs.202001530

Ma, Z. S. 2021. Coupling power laws offers a powerful method for problems such as biodiversity and COVID-19 fatality predictions. arXiv:2105.11002 [cs.CY]

Ma, Z. S., and A. M. Ellison. 2018. A unified concept of dominance applicable at both community and species scale. Ecosphere 9:e02477.

Ma, Z. S., and A. M. Ellison. 2019. Dominance network analysis provides a new framework for studying the diversity-stability relationship. Ecological Monographs 89:e01358.

Ma, Z. S., and L. W. Li. 2018. Measuring metagenome diversity and similarity with Hill numbers. Molecular Ecology Resources 2018:1–17.

Ma, Z. S., L. W. Li, and N. J. Gotelli. 2019. Diversity-disease relationships and shared species analyses for human microbiome-associated diseases. ISME Journal 13:1911–1920.

Martiny, J. B. H., et al. 2006. Microbial biogeography: putting microorganisms on the map. Nature Reviews Microbiology 4:102–112.

Nielsen, H. B., et al. 2014. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nature Biotechnology 32:822–828.

Partel, M., R. Szava-Kovats, and M. Zobel. 2011. Dark diversity: shedding light on absent species. Trends in Ecology & Evolution 26:124–128.

Plotkin, J. B., et al. 2000. Predicting species diversity in tropical forests. Proceedings of the National Academy of Sciences of the United States of America 97:10850–10854.

Qin, J., et al. 2010. A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464:59–65.

Qin, J., et al. 2012. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature 490:55–60.

Real, R., A. M. Barbosa, and J. W. Bull. 2017. Species distributions, quantum theory, and the enhancement of biodiversity measures. Systematic Biology 66:453–462.

Sczyrba, A., et al. 2017. Critical assessment of metagenome interpretation-a benchmark of metagenomics software. Nature Methods 14:1063–1071.

The Integrative HMP (iHMP) Research Network Consortium. 2019. The integrative human microbiome project. Nature 569:641–648.

Tjørve, E. 2003. Shapes and functions of species-area curves: a review of possible models. Journal of Biogeography 30:827–835.

Tjørve, E. 2009. Shapes and functions of species–area curves (II): a review of new models and parameterizations. Journal of Biogeography 36:1435–1445.

Ulrich, W., and J. Buszko. 2003. Self-similarity and the species-area relation of Polish butterflies. Basic and Applied Ecology 4:263–270.

van der Gast, C. J. 2013. Microbial biogeography and what Baas Becking should have said. Microbiology Today 40:108–111.

Wang, J., and H. Jia. 2016. Metagenome-wide association studies: fine-mining the microbiome. Nature Reviews Microbiology 14:508–522.

Xiao, L., et al. 2015. A catalog of the mouse gut metagenome. Nature Biotechnology 33:1103–1108.

Xiao, L., et al. 2016. A reference gene catalogue of the pig gut microbiome. Nature Microbiology 1:16161.

Zhou, J., S. Kang, C. W. Schadt, and C. T. Garten. 2008. Spatial scaling of functional gene diversity across various microbial taxa. Proceedings of the National Academy of Sciences of the United States of America 105:7768–7773.

Zhu, W., A. Lomsadze, and M. Borodovsky. 2010. Ab initio gene identification in metagenomic sequences. Nucleic Acids Research 38:e132.

Data Availability Statement

All data sets analyzed in the study are available in public domain and their sources are noted in Table 1. Computational code for implementing *m*-DAR is provided in Data S1, file name of "R-Code and DemoData.txt."

Supporting Information

Additional Supporting Information may be found online at: http://onlinelibrary.wiley.com/doi/10.1002/ecs2.3807/full