EMERGENT ECOLOGICAL PATTERNS AND MODELLING OF GUT MICROBIOMES IN HEALTH AND IN DISEASE

Jacopo Pasqualini[1], **Sonia Facchin**[2], **Andrea Rinaldo**[3,4],
Amos Maritan[1], **Edoardo Vincenzo Savarino**[2], **Samir Suweis**[1]
[1] *Dipartimento di Fisica "G. Galilei", Universita di Padova, Italy*
[2] *Dipartimento di Scienze Chirurgiche, Oncologiche e Gastroenterologiche (DiSCOG), University of Padova, Italy*
{jacopo.pasqualini@unipd.it, samir.suweis@unipd.it}

ABSTRACT

The Human Gut Microbiome (HGM) harbors a huge diversity in terms of species. Recent work has shown the existence of macroecological laws describing the variation and diversity in microbial communities using 16s OTU data. Here we investigate whether such patterns could be used to discriminate between different models of population dynamics, and characterize gut microbiomes in healthy vs. disease states. In this work, by using shotgun metagenomics data, we first discuss how a simple scaling relation (*Taylor's law*) plays a crucial role when building theoretical models for the gut microbiome. Moving from recent results, we then introduce the *Symmetric Schlömlich Model*, which is related to a widely used prior for relative species abundance (RSA), i.e. the *Dirichlet Model*. We show that the RSA can be obtained by constraining the solution of the *Stochastic Logistic Growth Model*, thus endowing all parameters with a clear ecological interpretation. Finally, we characterize the gut microbiomes from an ecological viewpoint, and compare how different models describe the emergent gut microbiome patterns in healthy and disease conditions.

1 Introduction

Next-generation sequencing (*NGS*) has disclosed the possibility of investigating microbial biodiversity with unprecedented resolution. This 'data explosion' is both a challenge and a fascinating opportunity. Indeed, during the last 15 years, researchers in the biomedical field adopted this approach to investigate the human microbiome, i.e., the collective ensemble of microorganism coexisting in and on the human body, in particular through the cataloging of the bacterial species in specific sites and their relative abundances [1, 2, 3, 4], shedding light on a plethora of micro-organisms impossible to access with traditional methods, such as culturing.

All these efforts have been designed as a community resource program to develop broadly available, reference dataset and rapidly released computational and statistical tools, and clinical protocols, to serve as a catalyst for this emerging field. Indeed, the studies published in the last decade have established the importance of the microbiome to human health and development [5, 6, 7, 8]. We have now understood that the human microbiome is a fundamental intermediary for our metabolism, synthesizing for us a multitude of fundamental chemical compounds that we would not be capable to provide ourselves[9].

The mechanism through which the alteration of the microbiota results in disease is generically referred as dysbiosis. A broadly used definition of dysbiosis describes it as a compositional and/or functional alteration in the microbiota that is triggered by a set of environmental and host-related factors that perturbs the microbial ecosystem to an extent that exceeds its resistance and resilience capabilities [10, 11]. In particular, several studies have shown how the alteration of the gut microbiota is linked to a variety of pathological conditions such as inflammatory bowel disease [12, 13, 14].

However, despite the tremendous advancement in microbiota research, our current understanding of the relationships between the microbiome and disease remains limited. More specifically, the attempts to characterize the signatures of healthy microbiomes have only relied on correlative analyses and microbiome-wide associations: effective for many purposes, but also confounding in many other instances and not sufficient to unveil causal drivers behind these correlations [15].

Further crucial factors that also need to be considered are sampling effects, false positives and compositionality in shotgun sequencing. The first comes from the fact that, in order to estimate species abundance from metagenomes, we never access the whole genetic material building up the community but only a subsample, in from of short DNA fragments. This effect also manifests in the sparsity of the dataset, where most of the species appear in few of the samples. However, sparsity may not indicate ecological exclusion of the species [16, 17]. The second effect pertains to the taxonomic profiling of DNA sequences. Due to the eventually large overlap among large regions of phylogenetically distant bacterial genomes, sequence alignment tools tend to introduce a lot of false positive species (typically found in very low abundances), which are not really present in the microbial community [18]. This problem is typically mitigated by choosing an appropriate database for classification or by setting a relative abundance cut-off, which is arbitrarily chosen [18]. Finally, microbiome data are affected by compositionality. Due to both technical (sequencing, sequence alignment errors) and biological variability of samples, it is often the case that taxonomic profiles differs - possibly by one or two order of magnitudes - in the number of aligned sequences, which makes the comparison between different samples difficult. In order to make different samples comparable, it is clear that a proper normalization must be adopted. In this work, we will adopt a sum-to-one normalization. This unavoidable normalization introduces constraints in the data, known as compositionality, which can strongly affects signals in the data, and must be carefully taken into account when analyzing microbiome studies. [19, 20, 21].

From a theoretical point of view, the dynamics of microbial communities can be studied with different approaches that lie on a wide spectrum. On one side we have probabilistic models allowing for the estimation of the relative abundances, also taking into account for sampling effect. A commonly used probabilistic model in microbiome studies is given by the Multinomial Dirichlet Distribution (MDD) [22]. It assumes that the relative abundances of the microbial species follow a Dirichlet distribution, which provides a flexible framework for modeling the variation in microbial community structure across samples and it has been applied in numerous human gut microbiome studies [23]. On the other side, we find consumer resource models, where nutrient uptake, cross-feeding, proteome allocations, and other ecological mechanisms can be considered in the modelling of species populations [24, 25, 26, 27]. The limitation of this approach is the very high number of parameters that should be inferred from data.

In between these two approaches, we find simple phenomenological models, as for example neutral stochastic demographic models [28], and the stochastic logistic model (SLM) for microbiome species dynamics [29]. These models reflect basic and fundamental ecological processes, such as birth and death of species, migration or effect of environmental fluctuations and species interactions on species abundance [16, 30]. In particular, it has been recently shown that the SLM, despite its simplicity, is able to reproduce emergent macroecological patterns in very different microbial communities, highlighting the important role of environmental stochasticity [16].

In this work, firstly we will show that the MDD and the SLM can be indeed related through a general sampling theory that considers ecological dynamics driven by logistic growth and environmental fluctuations. From this approach, we will derive a generalization of the MDD known as *Scaled Dirichlet Distribution* (SDD) [31] [32] through which we will model species abundances in gut microbiomes. There are two empirical patterns that we will exploit to reduce the number of fitting parameters in the models. The first is the relationship between the mean and variance in the abundance of microbial species populations in our samples, that typically displays a power law relationship with exponent 2 and it is known as *Taylor's Law* (TL) [33, 16]. The second is how the mean (relative) abundance of species are distributed (MAD), that in microbial communitiy as described by 16 OTU data, is found to be Lognormal [16].

In the spirit of a longstanding tradition in theoretical ecology and statistical physics, the model we propose is a null model, meaning that it gives expectations of what we should observe in microbial communities where the population of each species is driven by its effective growth rate, resource limitation (given by its carrying capacity) and environmental noise, considering the effect of sampling and compositionality [34, 28].

We will then compare the proposed models with emergent macroecological patterns obtained from a meta-analysis of three studies using shotgun metagenomic data for the gut microbiome in both healthy persons and individuals diagnosed with inflammatory bowel disease (IBD) [4, 13, 35]. The power of this framework is that producing minimal expectations of species abundance shape allows us to measure deviations in the data in an interpretable way. Importantly, how deviation of species associated with phenotype can help us to figure out how microbiome configurations affect health and disease.

We also perform a sensitivity analysis of our modelling results to different population cut-off thresholds, to test the robustness of our results and possible errors induced by false positives.

2 Theoretical Framework

We start by introducing the stochastic logistic model [16, 29], which gives the evolution of abundances of the species in time

$$\frac{dx_i}{dt} = \frac{x_i}{\tau_i}(1 - \frac{x_i}{K_i}) + \sqrt{\frac{\sigma_i}{\tau_i}} x_i \xi_i, \tag{1}$$

where $x_i \in (0, \infty)$, K_i is the carrying capacity of species i, τ_i set the growth time scale and σ_i is the width of environmental noise experienced by the i-th species. The latter captures the fluctuations induced to the species growth rate by the environmental stochasticity and by species interactions [30].

It can be shown that this process, once stationarity is reached, follows a Γ *Distribution*, i.e. $p_\Gamma(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$, which describes the abundance fluctuations of a species among different samples, without considering compositionality [36, 19] and sampling effects [17]. Following a recent proposal [37], we refer to it as the *bare* distribution. If we now consider compositionality and work with the relative populations $v_i = x_i / \sum_{i=1}^S x_i$, then the latter are distributed following the SDD [31], which describes the "dressed" joint probability distribution of $S-1$ species relative abundance, i.e.,

$$P(\vec{v}|\vec{\alpha}, \vec{\beta}) = \frac{1}{Z(\vec{\alpha}, \vec{\beta})} \frac{\prod_{i=1}^S v_i^{\alpha_i - 1}}{(\sum_{i=1}^S \beta_i v_i)^{\alpha_0}} \delta(1 - \sum_{i=1}^S v_i), \quad \textit{with } \alpha_i = \alpha(K_i, \sigma_i),\ \beta_i = \beta(K_i, \sigma_i), \tag{2}$$

where $\alpha_0 = \sum_{i=1}^S \alpha_i$, δ is the Dirac delta and $Z(\vec{\alpha}, \vec{\beta}) = B(\vec{\alpha}) / \prod_{i=1}^S \beta_i^{\alpha_i}$ is the normalization constant. Finally, $B(\vec{\alpha}) = \prod_{i=1}^S \Gamma(\alpha_i) / \Gamma(\sum_{i=1}^S \alpha_i)$ is the S-variate Beta function. Whatever particular form $P(\vec{v}|\vec{\alpha}, \vec{\beta})$ assumes, it describes the behaviour of the *relative species abundances* in a given sample, and lies on the $(S-1)$-dimensional simplex Δ^S. Such a family of distributions has $2S-1$ parameters (see section 4.2.2 of the Supplementary Material), and since $S \in [10^2, 10^3]$ and microbiome dataset include, typically, $R \approx 10^2$ samples, fitting this model is an unfeasible task. On the other hand, the family has an analytical expression, providing a clear ecological interpretation of its parameters. As we will show, we can reduce greatly the number of free parameters by constraining the model employing relationships obtained from empirical macro-ecological patterns, as proposed by [16].

First, we consider the TL. It takes the form of a scaling relation between the mean relative abundance of a species (among samples) and its fluctuations, i.e.

$$\sigma_{v_i}^2 = A \overline{v}_i^\zeta, \tag{3}$$

for $i = 1, ...S$, and where A and ζ do not depend on i. Compositionality does not affect such law if $\zeta = 2$, and thus the same A is found also if we consider $\vec{x}$, instead of the $\vec{v}$ (for a derivation of this result, see SM section 4). Therefore, we can connect the ecological parameters K_i, σ_i with $\zeta = 2$, finding that TL is informative on both intra-species competition (driven by K) and the intensity of the environmental noise (σ). Exploiting this relation, we can thus reduce the number of parameters in our SSD model to S. Considering the parametrization of the Γ distribution in terms of α and β, then the bare distribution can be parametrized just by its first moment $\overline{x}_i$:

$$\alpha_i = \frac{\overline{x}_i^{2-\zeta}}{A} \quad \beta_i = \frac{\overline{x_i}^{1-\zeta}}{A} \quad \textit{where } \overline{x_i} = K_i(1 - \frac{\sigma_i}{2}). \tag{4}$$

The dependence of the αs and βs on the exponent ζ suggests that there exists two interesting behaviour for the SDD. In fact, depending on the value of ζ, SDD will take the shape of a different distribution. For $\zeta = 1$, we have a Poisson-like scaling as variance and mean are proportional and the SDD reduces to the MDD. The other limiting behaviour is the one classically met in theoretical ecology, with $\zeta = 2$ [33, 16]. Therefore, considering these exponents for ζ, after some manipulations, we obtain the following

$$\textit{Poisson-like scaling}, \qquad P_{\zeta=1}(\vec{v}|\vec{\alpha}) = \frac{1}{B(\vec{\alpha})}\prod_{i=1}^{S}(v_i)^{\alpha_i-1}\delta(1-\sum_i v_i) \qquad \alpha_i = \frac{\overline{x}_i}{A}, \beta \sim const \tag{5}$$

$$\textit{Taylor-like scaling}, \qquad P_{\zeta=2}(\vec{v}|\alpha,\vec{\beta}) = \frac{1}{Z(S,\alpha,\vec{\beta})}\frac{\left(\prod_i v_i\right)^{\alpha-1}}{\left(\sum_i \beta_i v_i\right)^{\alpha S}}\delta(1-\sum_i v_i) \qquad \alpha = \frac{1}{A}, \beta_i \sim K_i^{-1}, \tag{6}$$

In the Taylor-like scaling case, due to the fact that $\alpha_i = \alpha$ the dependence of $Z(\vec{\alpha},\vec{\beta})$ from $\vec{\alpha}$ reduces to $Z(S,\alpha,\vec{\beta})$.

The second pattern we exploit to reduce the number of parameters in the SSD model is the species *mean* (relative) *abundance distribution* (MAD). Indeed, it has been shown that different types of microbiomes share the same average (relative) species abundance distribution [16], i.e., $\overline{x}_i \sim P_{MAD}(\mu,\lambda)$ ($\overline{v}_i \sim P_{MAD}(\mu_v,\lambda_v)$), where the parameter μ (μ_v) is defined so that $\mu = \mathbb{E}_{MAD}[\log \overline{x}_i]$ ($\mu_v = \mathbb{E}_{MAD}[\log \overline{v}_i]$); while λ (λ_v) can be calculated analytically given the specific shape of P_{MAD}, which solely depends on μ and λ.

We still have to take into account the effect of sampling, that in microbiome data cannot be neglected [17, 16]. We thus introduce the convolution of our model with a multinomial distribution, to consider the heterogeneous sequencing efforts in the different samples. In this way, we can generate a synthetic microbiome using the relative abundance of species $\vec{v}$ as the densities and the number of reads N as the number of trials of the multinomial distribution (see Fig. 1).

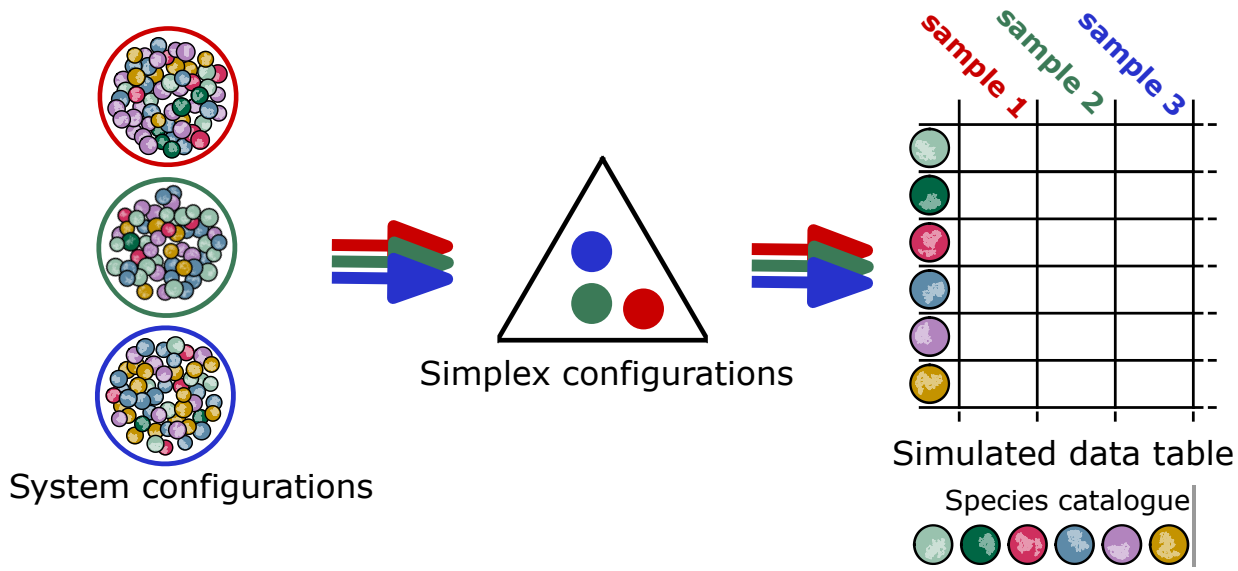

Figure 1: Schematic representation of the proposed theoretical framework to generate synthetic taxonomic data table. Panel a: colored large circles indicate different samples of microbial communities. Each ball inside the circle represents an individual of a given species (color). We normalize the species abundances n_i to densities v_i. In Panel b we show the three samples and three species case, i.e., each point now represents one of three samples, and its projection on the sides of the simplex (where each side is associated with a species), the corresponding relative abundance. Panel c: we eventually introduce sampling, and generate the taxonomic tables through a multinomial distribution with the densities v_i, and the number of individuals as those observed in the empirical data.

Operatively, in order to sample from this null model we implement the following procedure: 1) Extract S average abundances from $P_{MAD}(\mu,\lambda)$ (see Supplementary Material); 2) Generate $\vec{\beta}$ and the $\vec{\alpha}$ (whose dimensions depends on ζ) using Eqs. (4) and the estimate of the coefficient A from the TL (see section 4.1.2 of the Supplementary Materials); 3) Sample the relative species abundances v_i, $i = 1, ..S_R$ for each of the R samples we aim to generate using Eq. (5) or Eq. (6); 4) Sample the species abundances from the multinomial using v_i and N_R; 5) Remove all the species that have relative abundance below a given threshold κ, i.e. $v_i < \kappa$. We will apply three different cut-off $\kappa = 4.5 \cdot 10^-7$ (low), $\kappa = 9 \cdot 10^-6$ (medium), $\kappa = 0.00018$ (high).

Equipped by this null model, our aim is to investigate the statistical properties and macroecological patterns of two different groups of human gut microbial communities. The first coming from a cohort of healthy (H) individuals, while

the second from a cohort of individuals affected by gastrointestinal tract diseases, which we will generally refer to as unhealthy (U).

We will compare the results obtained for both $P_{\zeta=1}$ (MDD), $P_{\zeta=2}$ (SSD) models and also for the compositional-only-on-average stochastic logistic model originally proposed by Grilli [16] (SLG), i.e., $P_{SLG} = \prod_{i=1}^{S} p_\Gamma(\vec{v}|\alpha, \beta_i)\delta(1-\sum_i \bar{v}_i)$. We note that, from a statistical mechanic perspective, $P_{\zeta=2}$ and P_{Gr} are respectively the microcanonical and canonical formulation of the same model.

3 Microbiome Metadata Selection and Analysis

We have selected gut microbiome data from three studies [4, 13, 35], where both sequencing data and sample metadata are available for H controls and three gastrointestinal tract diseases: Chron (CD), Ulcerative Colitis (UC) and Inflammatory Bowels Syndromes (IBS). We have filtered so to have a homogeneous and not biased dataset (see section 5 of the Supplementary Material for details). In general, we have selected the patients less affected (at least at the time of the study) according to medical treatments, in order to limit the impact of different drug treatments on gut microbiome. After this filtering procedure, we ended up with $R_H = 91$ shotgun metagenomic samples from healthy control individuals and $R_U = 202$ samples of dysbiotic microbiome.

We have implemented a computational pipeline to process the throughput sequencing metagenomic data following best practices [38], such as quality filtering (remove reads with Q<20) and human DNA decontamination with the NCBI (GRCH38) human genome assembly [1].

The metagenomic taxonomic profiling tool we have adopted in our analysis is the *Kaiju* classifier [39], which converts metagenomic reads in all possible open reading frames (ORFs) and search for the best match in a protein database. The advantage of this procedure is that, thanks to the degeneracy of genetic code, is robust to random mutations along the genome and as such, to evolutionary divergences between the dataset and the reference catalog of genomes.

As a reference species catalogue, we have used RefSeq [40], which contains protein sequences from complete archaeal and bacterial genomes.

Eventually, we have classified (on average across samples) 39% of reads in H samples and 37% U, at the species level.

From these we build two data-tables, one for each of the two classes (H and U), having as rows the different species (S) and as columns (R) the samples. Each $\{i, j\}$ entry gives the corresponding relative species abundances v_i of the sample j, so that $\sum_{i=1}^{S} v_i^j = 1$. Relative abundances are obtained by dividing the number of reads assigned to a given species by the total number of reads recognized at the species level for that sample. To implement the relative abundance threshold, we set to zero all species abundances that are smaller than the relative abundance cut off κ, i.e. $v_i^j = 0$ if $v_i^j < \kappa$.

4 Results

4.1 Mean Abundance Distribution and Taylor's Law

We find that an emergent *MAD* is observed in both H and U dataset. However, we observe a dependence of the shape of the distribution depending on the relative abundance cut off κ. In particular for $\kappa < 10^{-5}$ the MAD displays a *Log-Laplace* shape, i.e. $P_{MAD}(\overline{x}|\mu, \lambda) = e^{-\frac{|\log \overline{x}-\mu|}{\lambda}}/(2\lambda\overline{x})$, while for $\kappa > 10^{-5}$ the MAD is a Log-Normal distribution $P_{MAD}(\overline{x}|\mu, \lambda) = e^{-\frac{(\log \overline{x}-\mu)^2}{2\lambda^2}}/(\overline{x}\sqrt{2\pi\lambda^2})$, the same found for OTU 16s data [16] In the y-axis of Figure 2 we show the *BIC* ratio **??**: if it is greater than one, it indicates that the Laplace distribution is a better fit than the Log-Normal, while if *BIC*<1, then the opposite is true. We thus obtain the values of μ and λ for three different thresholds of κ. As also clear from Fig. 2, the Laplace distribution is usually a better description for the MAD, except for large threshold where the MAD clearly dispays a Log-Normal shape (and compatibly with the OTU case [**?**]). As explained in the section 4.2.3 of the Supplementary Material, the SSD model is insensitive to the value of μ to generate syntethic table, while the value of λ is inferred from the data and and we find $\lambda_H = 1.407 \pm 0.0005$ and $\lambda_U = 1.413 \pm 0.0007$ (uncertainty of the fit evaluated with bootstrap procedure) as the best fit for the non-thresholded mean species abundances.

Regarding the TL, we find a remarkable robustness of the exponent $\zeta \approx 2$, while a scale dependence of the amplitude A. In particular, for each value of κ we compare the R^2-score ratio of the best fit power law with the one with fixed

[1] https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.26/

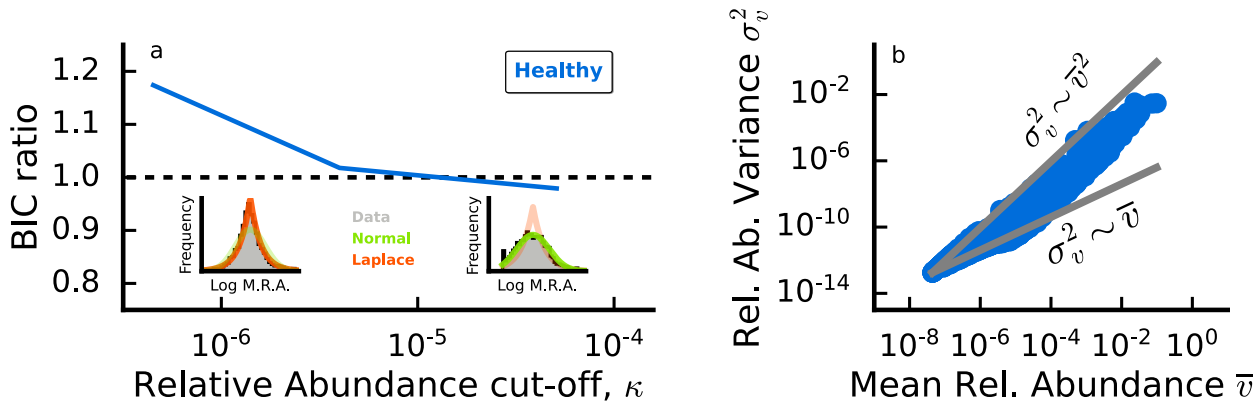

Figure 2: Panel a: *Mean Abundance Distribution* shape displays a dependence on the relative abundance cut-off κ. BIC ratio curve for healthy and disease related data collapse onto the same line. Panel b: Taylor's Law holds in empirical human gut communities, both in healthy and in disease. For rare species, it is difficult to discriminate between Poisson-like and Taylor-like scaling, due to the fact that rare species are present only in few samples. For simplicity, we report the scatter plot only in the healthy case. A brief discussion on the fit procedure can be found in the Supplementary Material, section 4.1.2. The blue straight line represents the model fitted with fixed slope $\zeta = 2$.

$\zeta = 2$ finding $\frac{R^2_{\zeta=2}}{R^2_{fit}} \approx 1$ suggesting a negligible discrepancy between the two models for this scaling relation. Thus, in what follows we will assume $\zeta_{Data} = 2$.

4.2 Emergent Ecological Patterns in healthy and unhealthy microbiomes

In this section, we investigate macro-ecological emergent patterns in gut microbiomes, and test whether our null model can describe them, and if there is any statistically significant deviations in such patterns between H and U samples.

We will focus on the following ecological patterns of the gut microbiomes: 1) α and γ diversity [41], defined as the number of different species in each local community (i.e., samples) and H and U meta-communities (i.e. union of all H/U samples), respectively; 2) The abundance-occupancy distribution, describing the probability for a species with mean relative abundance $\bar{v}$ is found in a fraction $\bar{o}$ of the total number of samples within a meta-community; 3) The species abundance distribution (SAD) of the H and U meta-communities; 4) The relation between the relative number of species observed in a local community and its sequencing depth (i.e. the metagenomic version of species-area relationship).

To compare a given ecological pattern obtained from the data and the corresponding one produced by the null model, we first set the scaling exponent ζ ($\gamma = 2$ for SSD and SLG, $\gamma = 1$ for MDD). Second, we fit the parameters μ, λ, A (in the case of SSD we do not fit μ due to model symmetries -see Supplementary Material, section 4.2.3) from the data with no cut-off ($\kappa = 0$). Then we generate 500 realizations of the two meta-communities (H and U) with the same number of reads (N), of species ($S = \gamma$), and of samples R as found in the data. Eventually, we consider the three relative abundance cut-off κ both in the empirical and simulated data (i.e. all species with $v_i < \kappa$ is set to zero) and for each pattern we calculate a R^2-like score (**??**). The final R^2-score we assign to the model is the average over the instances of the model.

The first patterns we consider are the γ and α diversity. The values of such quantities strongly depend on κ, but reveal a persistent regularity that is independent of it. Indeed, the overall (γ) diversity of specie is found in the unhealthy meta-community is larger than the one found in healthy microbiomes ($\gamma_U < \gamma_H$). On the other hand, we have that the *average* local diversity ($\overline{\alpha}$) found for H samples is larger compared to one for U samples, i.e., ($\overline{\alpha}_H < \overline{\alpha}_D$) (see figure 3)

The second pattern we consider is the SAD [42]. In agreement with previous results obtained with 16s OTU data [16, 43] and also shotgun data [44], we find that the SAD displays a heavy tail, that is compatible for small and medium cut-off with power-law distributions with exponents around 1.7 (see figure 7 of the supplementary methods for more details). For large threshold the SAD is more compatible with a Log-Normal distribution. No significance difference between H and U individuals is observed (see Figure 4a). Moreover, all the models (SLM, SSM and MDD) generate SADs that are compatible with the empirical ones. These results confirm [45] that SADs are not informative patterns of the underlying ecological mechanisms driving species abundances.

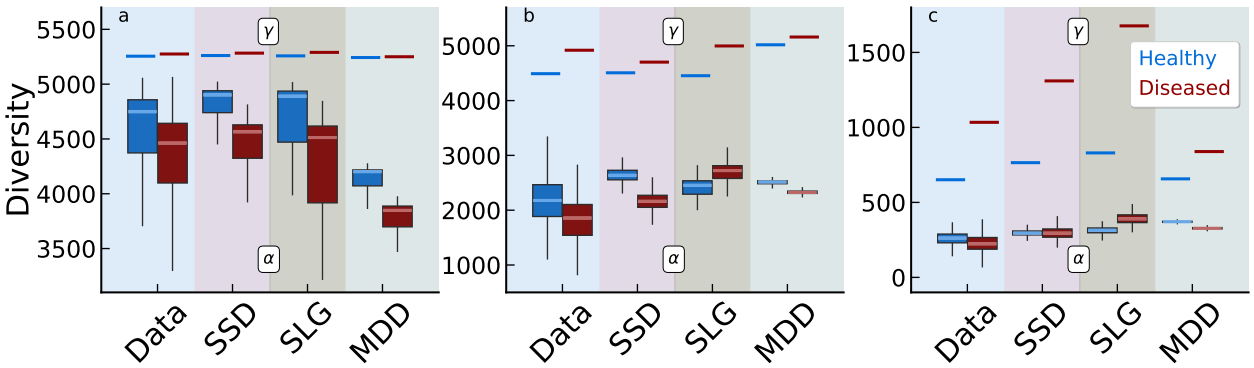

Figure 3: Box-whiskers plots describe average local (α) diversity, while horizontal bars indicate the corresponding metacommunity (γ) diversity. In healthy gut microbiomes (in blue) we find higher average α and lower γ diversity than unhealthy ones (in red). The three panels represent different threshold relative abundance cut-off κ: a) low ($\kappa = 4.5 \cdot 10^{-}7$); b) medium ($\kappa = 9 \cdot 10^{-}6$); c) high ($\kappa = 0.00018$). In each panel, we compare the diversity of the empirical H and U diversity with respect to the one generate by three different null models: SSD, SLG and MDD. SSD is found the best model, especially for low and medium κ.

We then investigate the relation between the log-mean relative abundance of a species and its occupancy. We have already introduced the average relative abundance of species i as $\overline{v}_i = \frac{1}{R}\sum_{j=1}^{R} v_i^j$, while we define the *occupancy* of species i as $\overline{o}_i = \frac{1}{R}\sum_{j=1}^{R} \theta(v_i^j)$, where θ is the Heaviside Theta. The relation between $\overline{v}$ and $\overline{o}$, as shown in Figure 34b, describes how likely it is for a species, given its average relative abundance to be sampled in a realization of the system (also known as intensity-sparsity relation [17]). When low/medium value of κ is set, the curve suddenly saturates, suggesting that a vast proportion of the available $S = \gamma$ species is expected to be sampled. In this scenario, the community is dominated by rare species, and thus almost all sampled individuals belong to different species, thus saturating the diversity very fast. As κ increases we have fewer and fewer species that are rare in relative abundance but, at the same time are harder to sample. As Figure 4b shows, this behaviour is shared by all models, and thus is also not discriminative of underlying ecological processes. However, for low (high) κ the MDD typically underestimates (overestimates) the occupancy of species (not shown here, see supplementary material, figure 1).

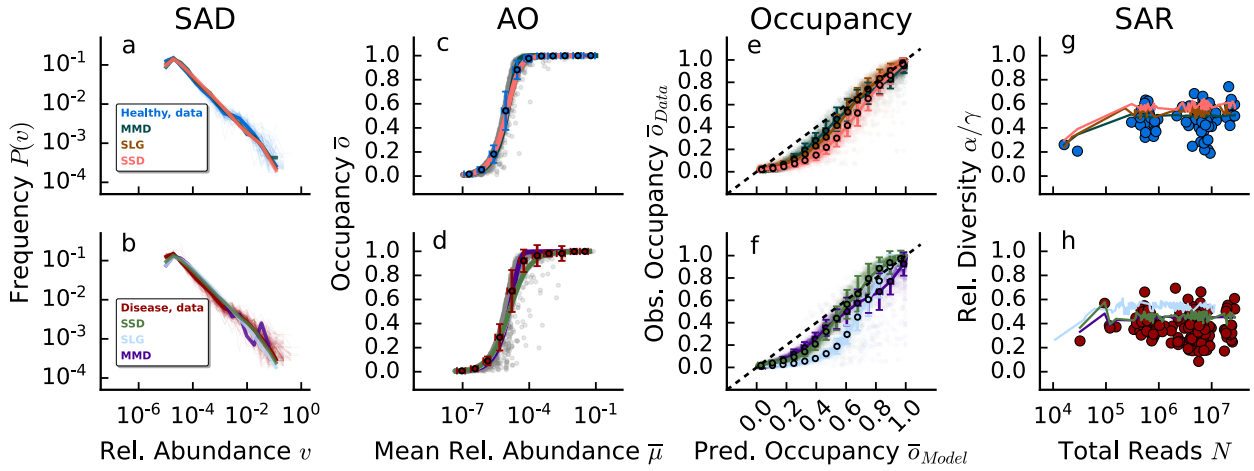

Figure 4: Comparison of emergent empirical ecological patterns in H and U microbiomes and for different models (SSD, SLG and MDD). Panel a-b) Species Abundance distribution (SAD); c-d) Species abundances occupancy curve (AO); e-f) Empirical vs. predicted occupancy curves (O); g-h) Species Area Relationship curves (SAR). Patterns which involve species abundances (SAD and AO) are indistinguishable for the different models. On the other hand, presence/absence patterns (O and SAR) seem to be more informative: here the different models show significative, but still mild deviations from the data. These patterns can be investigated for different cutoff (see Supplementary Method,figures 1-4), here we show the average threshold cut-off $\kappa = 9 \times 10^{-6}$.

We can also compare directly the occupancy curves obtained from presence-absence data, and those predicted by the models (as shown in Figure 4)e-f. Interestingly, contrarily from the previous patterns, in this case there are differences on how good the models describes the data. In particular for H samples, the SLG and MDD models outperform the SSD one, while on the contrary SSD better predict the pattern of U samples. However, such results are not robust for different threshold (see Supplementary Material, figure 1). In fact for low (high) κ we find that the MDD strongly underestimate (overestimate) the empirical occupancy curve, while SLG and SSD perform in a similar way.

We finally consider the "metagenomic" version of the species-area-relation (SAR) [46], i.e. how the diversity increases with increasing sampled area. Here the area is substituted by the total number of classified reads. Therefore, we consider increasing number of reads combining the samples of each group, and calculate the normalized diversity as the number of unique species in the aggregated community divided by γ diversity. Although all models on average slightly overestimate the overall diversity, in H samples they all perform similarly, while for U samples SSD and MMD models (that can generate the same number of total reads as found in the data) are slightly better than the SLG model.

Table 1 summurizes all the results and the comparison between the goodness of fit (measured as explained in section 4.2 of the three models for the presented emergent empirical ecological patterns of gut microbiomes in both health and in disease states.

	AO Curve			Species Occupancy			SAR Curve		
H	schlomilch	0.9971	0.9965	0.9838	0.8508	0.8236	0.8306	0.8195	-0.695
	cricket	0.9987	0.9954	0.9836	0.9083	0.9126	0.8115	0.9844	0.5854
	dirichlet	0.9372	0.9953	0.9445	0.5167	0.9254	0.3378	0.7768	0.0057
U	schlomilch	0.9928	0.9821	0.9874	0.8858	0.8619	0.7767	0.9766	0.8405
	cricket	0.9874	0.9823	0.9842	0.9117	0.5717	0.5691	0.9885	0.5222
	dirichlet	0.902	0.9938	0.8641	0.5627	0.8729	0.4856	0.8741	0.7267

Table 1: R^2 scores for the considered macroecological patterns.

Finally, by fitting the SSD, we are able to gain insight about possible differences in the species population dynamics within microbiomes in healthy and in disease. As reported in figure 5, by fitting the MAD we find λ for both H and U cohorts. We can interpret it as the fluctuation scale of the carrying capacities. This turns out to be slightly higher in the U cohort. This difference, which may seem negligible, may be relevant if we consider the analytic form of the LogLaplace distribution, displaying a critical point in $\lambda = 1$. The most relevant difference, however, comes from the value we infer for σ, the width of the environmental fluctuations. The SLM predicts that the abundance fluctuations distribution (AFD) has a polynomial part with an exponent that is greater than zero if $\sigma > 1$ (as observed in the unhealthy cohort) and less than zero if $\sigma < 1$ (as observed in the healthy cohort). Therefore, the fluctuations of the growth rates are much stronger in the U microbiomes, reflected in the AFD shifting from a modal distribution in H samples to a power law one (with exponential cutoff) in U microbiomes, i.e. in dysbios the probability for a species to be very rare is higher. This result can be interpreted as a factor promoting species turnover and thus explain the α-γ diversity tradeoff that we observe in the data and discussed before.

5 Discussion

The advantages of working with simple and interpretable models to fit the emergent ecological patterns in both H and U cohorts are: 1) We can discriminate which patterns give insights on the underling ecological processes, i.e. can discriminate between models; 2) We can compare patterns of gut microbiomes in health and in disease, infer the parameters of the SSD model that better describes such pattern and try to gain ecological insights on the dysbiois.

We have shown that microbiome species abundance data display a Taylor law with $\zeta = 2$. By construction the MDD does not satisfies this fundamental constraint. And yet, as shown in Table **??**, it is able to generate compatible AO curve and SADs with data at any threshold! Similarly, it can reproduce Occupancy curves at medium threshold and SAR curves at low threshold. Therefore, we show that not all the patterns (and thresholds) are equally informative and can be used to discriminate between models and or underlying ecological processes. However, AO-curves have been previously used to test specific underlying ecological theory [47]. On the contrary, our results suggest that the shape of the AO curve are simply the results of two main ingredients: heterogeneous populations averages and random sampling. Similarly all patterns that are not threshold seem to be strongly affected by a vast number of rare species (that in shotgun data are probably false positives [18]).

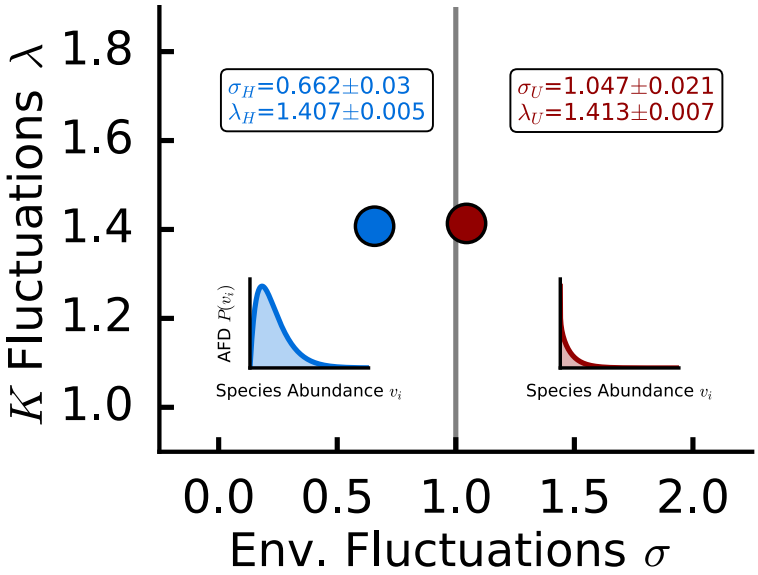

Figure 5: Taylor's Law and Mean abundance distribution parameters are informative of the underlying SLM model. The values obtained from the data suggest that healthy and unhealthy species abundance distribution follow different qualitative behaviours, with the unhealthy case being prone to more extinction. Light dots are obtained with a bootstrap procedure, from which reported averages and standard deviation of λ and σ are obtained.

Something similar occurs for the SAD, where all models are practically indistinguishable from one another. Indeed, the fact that different models (i.e., processes) can lead to very similar SAD patterns has long been known in theoretical ecology [48, 42].

Finally, we observe that the species occupancy and the species diversity are the less easy to capture, especially at medium/hight thresholds and from these we can get some insights by the different models performances. For example, when focusing on species occupancy, we observe that models incorporating Taylor's law ($\zeta = 2$) achieve higher explanation of the data. Similarly, by considering the SAR, the MDD has the poorest performance followed by the SLG, while the SSD achieves the best performance, although the R^2 is not very high. This scenario suggests that Taylor's law with a scaling exponent of $\zeta = 2$ (and thus large species fluctuations) is important to predict presence/absence patterns and species diversity in empirical data. Moreover, it also indicates that SLG, which reproduces ecological communities with a number of individuals that are only on average as those observed in the data, induces significant underestimates of species diversity.

All in all, we suggest that although compositionality and sampling strongly obscure ecological signals making the majority of empirical patterns qualitatively similar, there are indeed quantitative ecological differences between microbial communities of the gut microbiome in health and disease. In particular, by considering only a few relevant patterns like Taylor's law and species occupancy (and by using interpretable analytical models that also include environmental noise, we may propose an interpretation of the observed differences in the taxonomic data, eventually shedding light on underlying ecological processes characterizing informative emergent patterns, such as the specific trend of α and γ diversity in both H and U cohorts. We thus conclude that dysbiosis is characterized by stronger turnover than healthy microbiomes, that is due to larger environmental fluctuations and also reflected in marked heteregenities of species' carrying capacities.

References

[1] A framework for human microbiome research. *nature*, 486(7402):215–221, 2012.

[2] Structure, function and diversity of the healthy human microbiome. *nature*, 486(7402):207–214, 2012.

[3] Peter J Turnbaugh, Ruth E Ley, Micah Hamady, Claire M Fraser-Liggett, Rob Knight, and Jeffrey I Gordon. The human microbiome project. *Nature*, 449(7164):804–810, 2007.

[4] The integrative human microbiome project. *Nature*, 569(7758):641–648, 2019.

[5] Claire Duvallet, Sean M Gibbons, Thomas Gurry, Rafael A Irizarry, and Eric J Alm. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nature communications*, 8(1):1784, 2017.

[6] Jack A Gilbert, Martin J Blaser, J Gregory Caporaso, Janet K Jansson, Susan V Lynch, and Rob Knight. Current understanding of the human microbiome. *Nature medicine*, 24(4):392–400, 2018.

[7] A review of 10 years of human microbiome research activities at the us national institutes of health, fiscal years 2007-2016. *Microbiome*, 7(1):31, 2019.

[8] Ohad Manor, Chengzhen L Dai, Sergey A Kornilov, Brett Smith, Nathan D Price, Jennifer C Lovejoy, Sean M Gibbons, and Andrew T Magis. Health and disease markers correlate with gut microbiome composition across thousands of people. *Nature communications*, 11(1):5206, 2020.

[9] Steven R Gill, Mihai Pop, Robert T DeBoy, Paul B Eckburg, Peter J Turnbaugh, Buck S Samuel, Jeffrey I Gordon, David A Relman, Claire M Fraser-Liggett, and Karen E Nelson. Metagenomic analysis of the human distal gut microbiome. *science*, 312(5778):1355–1359, 2006.

[10] David A Relman. The human microbiome: ecosystem resilience and health. *Nutrition reviews*, 70(suppl_1):S2–S9, 2012.

[11] Jack A Gilbert and Susan V Lynch. Community ecology as a framework for human microbiome research. *Nature medicine*, 25(6):884–889, 2019.

[12] Jack A Gilbert et al. Microbiome-wide association studies link dynamic microbial consortia to disease. *Nature*, 535(7610):94–103, 2016.

[13] Jason Lloyd-Price, Cesar Arze, Ashwin N Ananthakrishnan, Melanie Schirmer, Julian Avila-Pacheco, Tiffany W Poon, Elizabeth Andrews, Nadim J Ajami, Kevin S Bonham, Colin J Brislawn, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*, 569(7758):655–662, 2019.

[14] Yong Fan and Oluf Pedersen. Gut microbiota in human metabolic health and disease. *Nature Reviews Microbiology*, 19(1):55–71, 2021.

[15] B. T. Tierney and et al. Systematically assessing microbiome-disease associations identifies drivers of inconsistency in metagenomic research. *Plos Biol*, 20:e3001556, 2022.

[16] Jacopo Grilli. Macroecological laws describe variation and diversity in microbial communities. *Nature communications*, 11(1):1–11, 2020.

[17] Giacomo Baruzzo, Ilaria Patuzzi, and Barbara Di Camillo. Beware to ignore the rare: how imputing zero-values can improve the quality of 16s rrna gene studies results. *BMC bioinformatics*, 22(Suppl 15):618, 2021.

[18] Anna Tovo, Peter Menzel, Anders Krogh, Marco Cosentino Lagomarsino, and Samir Suweis. Taxonomic classification method for metagenomics based on core protein families with core-kaiju. *Nucleic acids research*, 48(16):e93–e93, 2020.

[19] Gregory B Gloor, Jean M Macklaim, Vera Pawlowsky-Glahn, and Juan J Egozcue. Microbiome datasets are compositional: and this is not optional. *Frontiers in microbiology*, 8:2224, 2017.

[20] Matthew CB Tsilimigras and Anthony A Fodor. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Annals of epidemiology*, 26(5):330–335, 2016.

[21] Dionne Swift, Kellen Cresswell, Robert Johnson, Spiro Stilianoudakis, and Xingtao Wei. A review of normalization and differential abundance methods for microbiome counts data. *Wiley Interdisciplinary Reviews: Computational Statistics*, 15(1):e1586, 2023.

[22] Joshua G Harrison, W John Calder, Vivaswat Shastry, and C Alex Buerkle. Dirichlet-multinomial modelling outperforms alternatives for analysis of microbiome and other ecological count data. *Molecular ecology resources*, 20(2):481–497, 2020.

[23] Ahmed A Metwally, Hani Aldirawi, and Jie Yang. A review on probabilistic models used in microbiome studies. *Communications in Information and Systems*, 18(3):173–191, 2018.

[24] Joshua E Goldford, Nanxi Lu, Djordje Bajić, Sylvie Estrela, Mikhail Tikhonov, Alicia Sanchez-Gorostiaga, Daniel Segrè, Pankaj Mehta, and Alvaro Sanchez. Emergent simplicity in microbial community assembly. *Science*, 361(6401):469–474, 2018.

[25] Chengyi Tu, Samir Suweis, Jacopo Grilli, Marco Formentin, and Amos Maritan. Reconciling cooperation, biodiversity and stability in complex ecological communities. *Scientific reports*, 9(1):5580, 2019.

[26] Robert Marsland III, Wenping Cui, and Pankaj Mehta. A minimal model for microbial biodiversity can reproduce experimentally observed ecological patterns. *Scientific reports*, 10(1):3308, 2020.

[27] Leonardo Pacciani-Mori, Samir Suweis, Amos Maritan, and Andrea Giometto. Constrained proteome allocation affects coexistence in models of competitive microbial communities. *The ISME Journal*, 15(5):1458–1477, 2021.

[28] Sandro Azaele, Simone Pigolotti, Jayanth R Banavar, and Amos Maritan. Dynamical evolution of ecosystems. *Nature*, 444(7121):926–928, 2006.

[29] Lana Descheemaeker and Sophie De Buyl. Stochastic logistic models reproduce experimental time series of microbial communities. *Elife*, 9:e55650, 2020.

[30] Samir Suweis, Francesco Ferraro, Sandro Azaele, and Amos Maritan. Generalized lotka-volterra systems with time correlated stochastic interactions. *arXiv preprint arXiv:2307.02851*, 2023.

[31] Gianna Serafina Monti, Glòria Mateu i Figueras, Vera Pawlowsky-Glahn, and Juan José Egozcue. The shifted-scaled dirichlet distribution in the simplex. In © *International Workshop on Compositional Data Analysis (4th: 2011: Sant Feliu de GuÍxols, Girona). CODAWORK 2011: International Workshop on Compositional Data Analysis, hold on May 9-13rd. 2011, Sant Feliu de Guíxols, Girona*. Universitat Politècnica de Catalunya. Centre Internacional de Mètodes . . . , 2011.

[32] David DK Chow. Schl\" omilch integrals and probability distributions on the simplex. *arXiv preprint arXiv:2201.11013*, 2022.

[33] Andrea Giometto, Marco Formentin, Andrea Rinaldo, Joel E Cohen, and Amos Maritan. Sample and population exponents of generalized taylor's law. *Proceedings of the National Academy of Sciences*, 112(25):7755–7760, 2015.

[34] Nicholas J Gotelli and Gary R Graves. *Null models in ecology*. 1996.

[35] Ruben AT Mars, Yi Yang, Tonya Ward, Mo Houtti, Sambhawa Priya, Heather R Lekatz, Xiaojia Tang, Zhifu Sun, Krishna R Kalari, Tal Korem, et al. Longitudinal multi-omics reveals subset-specific mechanisms underlying irritable bowel syndrome. *Cell*, 182(6):1460–1473, 2020.

[36] John Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160, 1982.

[37] Onofrio Mazzarisi, Amanda de Azevedo-Lopes, Jeferson J Arenzon, and Federico Corberi. Maximal diversity and zipf's law. *Physical Review Letters*, 127(12):128301, 2021.

[38] Rob Knight, Alison Vrbanac, Bryn C Taylor, Alexander Aksenov, Chris Callewaert, Justine Debelius, Antonio Gonzalez, Tomasz Kosciolek, Laura-Isobel McCall, Daniel McDonald, et al. Best practices for analysing microbiomes. *Nature Reviews Microbiology*, 16(7):410–422, 2018.

[39] Peter Menzel, Kim Lee Ng, and Anders Krogh. Fast and sensitive taxonomic classification for metagenomics with kaiju. *Nature communications*, 7(1):11257, 2016.

[40] Nuala A O'Leary, Mathew W Wright, J Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, et al. Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, 44(D1):D733–D745, 2016.

[41] Elizabeth K Costello, Keaton Stagaman, Les Dethlefsen, Brendan JM Bohannan, and David A Relman. The application of ecological theory toward an understanding of the human microbiome. *Science*, 336(6086):1255–1262, 2012.

[42] Sandro Azaele, Samir Suweis, Jacopo Grilli, Igor Volkov, Jayanth R Banavar, and Amos Maritan. Statistical mechanics of ecological systems: Neutral theory and beyond. *Reviews of Modern Physics*, 88(3):035003, 2016.

[43] Ilaria Patuzzi, Giacomo Baruzzo, Carmen Losasso, Antonia Ricci, and Barbara Di Camillo. metasparsim: a 16s rrna gene sequencing count data simulator. *BMC bioinformatics*, 20(9):1–13, 2019.

[44] Claudia Sala, Silvia Vitali, Enrico Giampieri, Ìtalo Faria do Valle, Daniel Remondini, Paolo Garagnani, Matteo Bersanelli, Ettore Mosca, Luciano Milanesi, and Gastone Castellani. Stochastic neutral modelling of the gut microbiota's relative species abundance from next generation sequencing data. *BMC bioinformatics*, 17(2):179–188, 2016.

[45] James Rosindell, Stephen P Hubbell, Fangliang He, Luke J Harmon, and Rampal S Etienne. The case for ecological neutral theory. *Trends in ecology & evolution*, 27(4):203–208, 2012.

[46] Bernard D Coleman. On random placement and species-area relations. *Mathematical Biosciences*, 54(3-4):191–215, 1981.

[47] Michael Sieber, Lucía Pita, Nancy Weiland-Bräuer, Philipp Dirksen, Jun Wang, Benedikt Mortzfeld, Sören Franzenburg, Ruth A Schmitz, John F Baines, Sebastian Fraune, et al. Neutrality in the metaorganism. *PLoS Biology*, 17(6):e3000298, 2019.

[48] Drew W Purves, Stephen W Pacala, et al. Ecological drift in niche-structured communities: neutral pattern does not imply neutral process. *Biotic interactions in the tropics*, pages 107–138, 2005.