PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

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 Savarino[2], Samir Suweis[1]*

1 Dipartimento di Fisica "G. Galilei" e INFN sezione di Padova, University of Padova, Padova, Italy, **2** Dipartimento di Scienze Chirurgiche, Oncologiche e Gastroenterologiche (DiSCOG), University of Padova, Padova, Italy, **3** Dipartimento di Ingegneria Civile, Edile e Ambientale (ICEA), University of Padova, Padova, Italy, **4** Laboratory of Ecohydrology, École Polytechnique Fédérale Lausanne, Lausanne, Switzerland

* jacopo.pasqualini@unipd.it (JP); samir.suweis@unipd.it (SS)

OPEN ACCESS

Citation: Pasqualini J, Facchin S, Rinaldo A, Maritan A, Savarino E, Suweis S (2024) Emergent ecological patterns and modelling of gut microbiomes in health and in disease. PLoS Comput Biol 20(9): e1012482. https://doi.org/10.1371/journal.pcbi.1012482

Editor: Nic Vega, Emory University Department of Biology, UNITED STATES OF AMERICA

Received: November 2, 2023

Accepted: September 11, 2024

Published: September 27, 2024

Data Availability Statement: Shotgun sequencing for human gut microbiome were obtained by publically available studies, available at https://doi.org/10.5281/zenodo.13757689. The GRCH38 Human Genome Assembly for host decontamination can be found at https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.26/ The code for taxonomic profiling of metagenomic sequences has been deposited at https://github.com/jacopopasqualini/MetaGym. All codes for simulations and figures are available at https://

Abstract

Recent advancements in next-generation sequencing have revolutionized our understanding of the human microbiome. Despite this progress, challenges persist in comprehending the microbiome's influence on disease, hindered by technical complexities in species classification, abundance estimation, and data compositionality. At the same time, the existence of macroecological laws describing the variation and diversity in microbial communities irrespective of their environment has been recently proposed using 16s data and explained by a simple phenomenological model of population dynamics. We here investigate the relationship between dysbiosis, i.e. in unhealthy individuals there are deviations from the "regular" composition of the gut microbial community, and the existence of macro-ecological emergent law in microbial communities. We first quantitatively reconstruct these patterns at the species level using shotgun data, and addressing the consequences of sampling effects and statistical errors on ecological patterns. We then ask if such patterns can discriminate between healthy and unhealthy cohorts. Concomitantly, we evaluate the efficacy of different statistical generative models, which incorporate sampling and population dynamics, to describe such patterns and distinguish which are expected by chance, versus those that are potentially informative about disease states or other biological drivers. A critical aspect of our analysis is understanding the relationship between model parameters, which have clear ecological interpretations, and the state of the gut microbiome, thereby enabling the generation of synthetic compositional data that distinctively represent healthy and unhealthy individuals. Our approach, grounded in theoretical ecology and statistical physics, allows for a robust comparison of these models with empirical data, enhancing our understanding of the strengths and limitations of simple microbial models of population dynamics.

Author summary

In this study, we explore emerging ecological properties in gut microbiomes. Our aim here is to determine whether these patterns can be informative of the gut microbiome (healthy or diseased) and unveil essential ingredients driving its population dynamics.

github.com/jacopopasqualini/EmergetPatternsHumanGut.

Funding: S.S. acknowledges Iniziativa PNC0000002-DARE - Digital Lifelong Prevention. A.M. also acknowledges the support of the NBFC to the University of Padova, funded by the Italian Ministry of University and Research, PNRR, Missione 4, Componente 2, "Dalla ricerca all'impresa", Investimento 1.4, Project CN00000033. A.R. acknowledges funding from Fondazione Cassa di Risparmio di Padova e Rovigo (IT) through its grant 55722. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Leveraging on metagenomic data and interpretable statistical models based on ecological processes, we show that not all ecological patterns are informative to characterize its states, while few are (e.g., species diversity). Eventually, thanks to the ecological interpretability of the inferred models' parameters, our analysis provides insights into the role of environmental fluctuations and carrying capacities of the gut microbiomes in both health and disease. This study offers valuable knowledge, bridging theoretical concepts with practical implications for human health.

Introduction

Next-generation sequencing has expanded our capacity to explore microbial biodiversity in a previously unachievable depth. This 'data explosion' presents both challenges and exciting prospects. Over the past 15 years, biomedical researchers have leveraged this technology to delve into the human microbiome—the complex ecosystem of microorganisms coexisting in and on the human body [1–4]. This approach has illuminated countless microorganisms, once inaccessible via conventional culturing methods. All these efforts have aimed to establish a community resource program to build comprehensive reference datasets and develop computational tools and clinical protocols. Although several recent studies underscore the critical role of the microbiome in human health [5–11], our understanding of how the microbiome influences disease is still limited. Current methods, primarily focused on correlations and associations within the microbiome, are useful but often fail to identify the actual causes behind these patterns [12]. In part, this is also due to several technical challenges in species classification and abundance estimations, like sampling effects, false positives species (type 1 statistical error in species detection) and data compositionality. In fact, capturing only sample fragments of the entire genetic material leads to sparse datasets, where zero abundances do not always imply species absence [13, 14]. Taxonomic profiling introduces false positive species due to genome sequence overlaps, often resolved by selecting appropriate databases or setting abundance cut-offs [15]. Additionally, normalization, such as sum-to-one, is essential due to the compositional nature of microbiome data, significantly impacting data analysis [16–18]. Nevertheless, microbiomes data display several emergent ecological patterns that are suitable to be explained through population dynamics models. Such models range from probabilistic ones, like the Multinomial Dirichlet Distribution [19], which estimates relative abundances and considers sampling effects, to more complex interaction-based models like the generalized Lotka-Volterra model or other phenomenological/computational models based on inferring species interactions [20–23]. Finally, non-interacting stochastic models reflecting basic ecological processes, like the stochastic logistic model [24], have demonstrated their effectiveness in reproducing microbiomes macroecological patterns [13].

However, an investigation of a possible link between such emergent patterns and the state space of the human gut microbiome, including dysbiosis, is still missing. This study intends to address this gap by integrating theoretical frameworks in population dynamics modelling with empirical data on gut microbiomes in health and in disease. In particular, our work aims to: 1) Quantitatively reconstruct gut microbiome emergent patterns [25] in health and in disease at the species level using shotgun data through a recently proposed taxonomic classifier [26]; In particular, our analysis includes a meta-analysis of studies on gut microbiomes in healthy individuals and those with gastrointestinal diseases [4, 10, 27]. 2) Examine the consequences of sampling effects and false positives on such ecological patterns; 3) Develop and compare how well different statistical interpretable ecological models can describe such patterns and test

possible differences between healthy and diseased cohorts; 4) Understand possible relationships of the inferred models parameters (having a well-defined ecological interpretation) with gut microbiome state (e.g., health or disease), so to be able to generate synthetic compositional data with statistically significant differences between healthy and unhealthy individuals. In this way, we can distinguish which patterns are "inevitable" versus those that are potentially informative about disease states or other biological drivers.

Materials and methods

Theoretical framework

We now implement the mathematical framework, which can be summarized as follows. Starting from a set of equations describing the dynamics of species abundances, we develop three different statistical models aimed at generating synthetic data that match the observed data. We then will use macroecological relationships to constrain these models, leaving a small number of free parameters that can be fit to empirical data. By fitting these models and comparing their predictions against data from healthy and unhealthy microbiomes, we will eventually be able to gain insight into how different ecological patterns may arise and if and how the parameters inferred from the models are related to the state (H or U) of the gut microbiome.

We therefore introduce the stochastic logistic model [13, 24], which gives the evolution of S species abundances in time

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

where $x_i \in (0, \infty)$, K_i is the carrying capacity of species $i = 1...S$, τ_i sets 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 environment (e.g. host) and by species interactions [28].

It can be shown that this process, once stationarity is reached, follows a Gamma *Distribution*, i.e. $p_\Gamma(x_i) = \frac{\beta_i^{\alpha_i}}{\Gamma(\alpha_i)} x_i^{\alpha_i - 1} e^{-\beta_i x_i}$, which describes the abundance fluctuations of species i among different samples, without considering compositionality [16, 29] and sampling effects [14]. Generally, we will refer to the distribution of abundance of a given species across different samples as abundance fluctuation distribution [13]. The parameters α_i and β_i can be related to the statistical mean and variance of the species abundances and the ecological parameters given by the model as

$$\alpha_i = \frac{\bar{x}_i^2}{\sigma_{x_i}^2} = \frac{2 - \sigma_i}{\sigma_i}, \tag{2}$$

$$\underbrace{\beta_i}_{\text{Parameter}} = \underbrace{\frac{\bar{x}_i}{\sigma_{x_i}^2}}_{\text{Observables}} = \underbrace{\frac{2}{\sigma_i K_i}}_{\text{Ecology}} \tag{3}$$

Also, we can write the first two moments of the Gamma distribution as $\bar{x}_i = K_i\left(1 - \frac{\sigma_i}{2}\right)$ and $\sigma_{x_i}^2 = K_i^2 \frac{\sigma_i}{2}\left(1 - \frac{\sigma_i}{2}\right)$. In other words, Eqs (2) and 3 show how the parameters of the Gamma distribution (α_i, β_i) are related to the observables $(\bar{x}_i, \sigma_{x_i}^2)$ and to the ecological parameters K_i and σ_i. If we now consider compositionality and work with the relative abundances

$\nu_i = x_i / \sum_{i=1}^{S} x_i$, then the latter are distributed following the Scaled Dirichlet Distribution [30]

$$P(\vec{\nu}|\vec{\alpha},\vec{\beta}) = \frac{1}{Z(\vec{\alpha},\vec{\beta})} \frac{\prod_{i=1}^{S} \nu_i^{\alpha_i - 1}}{\left(\sum_{i=1}^{S} \beta_i \nu_i\right)^{\alpha_0}}, \tag{4}$$

where the distribution is defined with the constraint $\sum_{i=1}^{S} \nu_i = 1$. We have also introduced $\alpha_0 = \sum_{i=1}^{S} \alpha_i$, and the normalization constant $Z(\vec{\alpha},\vec{\beta}) = B(\vec{\alpha}) / \prod_{i=1}^{S} \beta_i^{\alpha_i}$. Finally, $B(\vec{\alpha}) = \prod_{i=1}^{S} \Gamma(\alpha_i) / \Gamma(\sum_{i=1}^{S} \alpha_i)$ is the S-variate Beta function. An exhaustive derivation of this result can be found in S2 Text. The obtained family of probability distributions $P(\vec{\nu}|\vec{\alpha},\vec{\beta})$ describes the behavior of the *relative species abundances* in a given sample, has $2S$ parameters, and lies in the $(S-1)$-dimensional simplex Δ^S. Since the number of observed species is large ($S \in [10^2, 10^3]$) and the microbiome dataset typically includes $R \approx 10^2$ samples, fitting this model is an unfeasible task. As we will show, we can greatly reduce the number of free parameters by constraining the model through relationships obtained from empirical macro-ecological patterns, as proposed by [13].

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

$$\sigma_{x_i}^2 = A \bar{x}_i^{\zeta}, \tag{5}$$

for $i = 1 \ldots S$, where A and ζ do not depend on i. Compositionality does not affect this law if and only if $\zeta = 2$, and thus in this case the same A is also found if we consider $\vec{\nu}$, instead of $\vec{x}$ (otherwise a correction of the slope A should be taken into account, see S2 Text). 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 environmental noise (σ).

Exploiting Eqs (2) and 3, we can thus reduce the number of parameters in our Scaled Dirichlet Distribution model to $2 + S$, since α_i and β_i are functions of $\bar{x}_i$, ζ and A:

$$\alpha_i = \frac{\bar{x}_i^{2-\zeta}}{A}, \quad \beta_i = \frac{\bar{x}_i^{1-\zeta}}{A}. \tag{6}$$

The dependence of the αs and βs on the exponent ζ suggests that there exist two interesting behaviors for the Scaled Dirichlet Distribution. In fact, for $\zeta = 1$ we have a Poisson-like scaling as variance and mean are proportional, and the Scaled Dirichlet Distribution reduces to the Dirichlet distribution. The other limiting behavior with $\zeta = 2$ is classically encountered in theoretical ecology [31]. In the following, we will only consider these two limiting cases, reducing the number of free parameters to 1 (the TL's amplitude A) + S (the species mean abundances). Therefore, after some manipulations, we obtain the following distributions:

$$P_{\zeta=1}(\vec{\nu}|\vec{\alpha}) = \frac{1}{B(\vec{\alpha})} \prod_{i=1}^{S} \nu_i^{\alpha_i - 1} \tag{7}$$

$$P_{\zeta=2}(\vec{\nu}|S,\alpha,\vec{\beta}) = \frac{1}{Z(S,\alpha,\vec{\beta})} \frac{\left(\prod_i \nu_i\right)^{\alpha-1}}{\left(\sum_i \beta_i \nu_i\right)^{\alpha S}}, \tag{8}$$

where $\sum_i \nu_i = 1$. The $\zeta = 1$ prescribes a Poisson-like scaling with $\alpha_i = \frac{\bar{x}_i}{A}$ ($\bar{x}_i$ and A can be directly obtained from the data), and all the βs are proportional to a constant, and are canceled out from the analytic expression of the Dirichlet distribution given by Eq (7). On the other hand, for $\zeta = 2$ we find that all $\alpha_i = \alpha = (2 - \sigma)/\sigma = A^{-1}$ are constant and $\beta_i \sim K_i^{-1}$. Due to the invariance of the corresponding distribution by rescaling of the βs, the proportionality constant is

irrelevant (see SM section 4.2.4). Also, because in the latter case $\alpha_i = \alpha$, then the dependence of $Z(\vec{\alpha}, \vec{\beta})$ on $\vec{\alpha}$ reduces to $Z(S, \alpha, \vec{\beta})$ and thus the corresponding joint pdf given by Eq (8) is the Symmetric Scaled Dirichlet distribution.

The second pattern that we exploit to reduce the number of the models free parameters is the species *mean* (relative) *abundance distribution* (MAD), which describes the frequencies of the species abundances averaged over samples. Indeed, it has been shown that different types of microbiomes share the same average (relative) species abundance distribution [13], i.e., $\bar{x}_i \sim P_{\mathrm{MAD}}(\mu, \lambda)$ ($\bar{\nu}_i \sim P_{\mathrm{MAD}}(\mu_\nu, \lambda_\nu)$), where the parameters μ (μ_ν) and λ (λ_ν) can be inferred from the data. In S2 Text, we show that in the large S limit $\lambda_\nu \approx \lambda$. Typically, the distribution P_{MAD} has fat tails, and it is compatible with a Log-Normal distribution [13]. Regardless of the particular form P_{MAD} takes, its presence allows us to generate all the $\bar{x}_i$ ($i = 1, \ldots S$) once μ and λ are fitted from the data.

To take into account the effect of sampling, which cannot be neglected in microbiome data [13, 14], we introduce the convolution of the Dirichlet and Symmetric Scaled Dirichlet distributions with a multinomial one. In the first case, convolving the Dirichlet distribution with multinomial sampling, we obtain the Multinomial Dirichlet Distribution [19] (MD). The independent parameters of this model are thus the MAD parameters (μ, λ) and the TL amplitude A. The MD model is compositional, but it does not satisfy the TL. In the second case, by convolving the Symmetric Scaled Dirichlet distribution with the multinomial distribution, we obtain a novel model with ecologically grounded constraints (TL), which we will refer to as Multinomial Symmetric Scaled Dirichlet (MSSD). In this way, we can generate a synthetic microbiome using the relative abundance of species $\vec{\nu}$ as the densities and the number of reads N as the number of trials of the multinomial distribution (see Fig 1). In this case, the number of independent parameters to fit from the data is two, λ and A (or, equivalently, λ and σ).

We will compare the results obtained for both MD and MSSD models and also for the compositional-only-on-average stochastic logistic model originally proposed by Grilli [13], i.e., $P_{SLG} = \prod_{i=1}^{S} p_\Gamma(\nu_i|\alpha_i, \beta_i)$, with $\sum_i \bar{\nu}_i = 1$. To consider sampling effects and since the corresponding joint species abundance distribution does not have a sum-to-one hard constraint, we convolve P_{SLG} with the Poisson distribution and call this model the Poisson Stochastic Logistic Growth model (PSLG). In this model, the ingredient that ensures (on average) compositionality is the fact that mean abundances are constrained to sum up to one. We note that, from a statistical mechanics perspective, $P_{\zeta=2}$ and P_{SLG} are, respectively, the microcanonical and canonical formulations of the same model. Similarly to the MD, the parameters required to fully specify the model are μ, λ and A.

Operatively, to sample from the three above described model models, we implement the following procedure (see also Fig 1): 1) Depending on the model, fit μ, λ, A from the data (for a detailed discussion about fitting procedures, see S2 Text). 2) Extract S average abundances from $P_{\mathrm{MAD}}(\mu, \lambda)$; 3) Using Eq (6) and the estimate of the coefficient A from the TL, generate $\vec{\beta}$ and $\vec{\alpha}$ (whose dimensions depend on ζ, see Eq (6)); 4) Sample the relative species abundances ν_i, $i = 1, ..S_R$ for each of the $r = 1. . .R$ samples we want to generate using Eq (7) or Eq (8); 5) For each sample, with the appropriate sampling distribution, generate species counts using relative abundances and ν_i and the number of classified reads in the r–th sample, N_r; 5) Remove all species whose relative abundance is below a given threshold κ. In order to understand the effect of false positive species on the patterns and models outcomes, we will apply three different cut-off $\kappa_l = 4.5 \times 10^{-7}$ (low), $\kappa_m = 9 \times 10^{-6}$ (medium), $\kappa_h = 1.8 \times 10^{-4}$ (high). The three cut-off values were obtained considering how the diversity of the dataset changes with κ (for a derivation of these values see Fig E in S1 File).

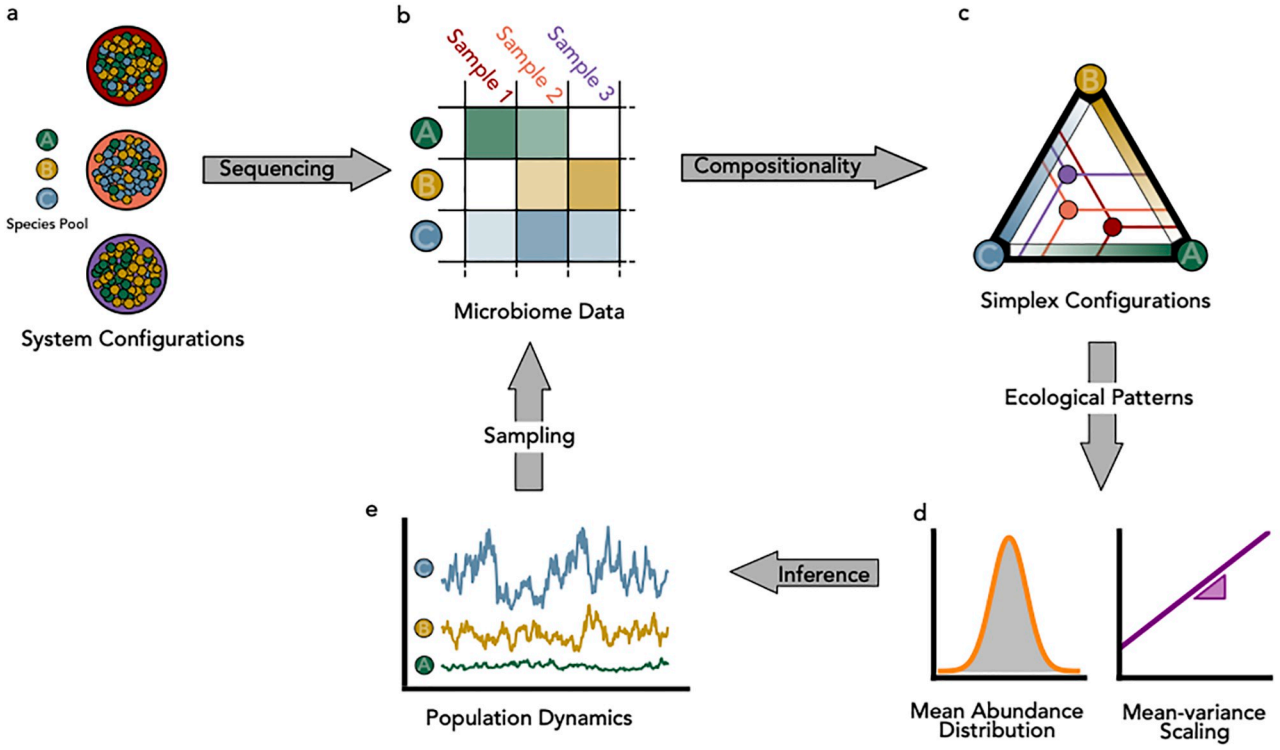

Fig 1. Schematic representation of the proposed theoretical framework to generate synthetic taxonomic data tables. Panel a: large coloured circles indicate different samples of microbial communities. Each small ball inside the circle represents an individual of a given species (A,B,C). We normalize the species abundances n_i to densities ν_i ($\sum_i^S \nu_i = 1$). Panel b: We conceptualise the empirical data generation process in terms of sequential steps that produce tabular count data. Species with low relative abundance may not be sampled, such as species B in sample 1. In panel c we show the densities phase space for the case of three species and three samples, which can be represented by a simplex. Each point within the simplex now represents one of the three samples, while each vertex represents a configuration where only the corresponding species is present in the sample. The relative abundance contribution of each species can be represented by a gradient. For a given sample, in order to obtain the relative abundance of a species, one has to project parallel to the simplex sides (as denoted by the lines). In panel d, we show the two macroecological patterns (Mean Relative Abundance distribution and the mean-variance species abundances scaling relation) that can be used to infer the parameters of the stochastic logistic growth model. Finally, in panel e, we show how the inferred population dynamics model, combined with sampling (e.g. the Multinomial Distribution), can be used to generate synthetic tabular data.

https://doi.org/10.1371/journal.pcbi.1012482.g001

Equipped with these models, our aim is to investigate the statistical properties and macroecological patterns of two distinct groups of human gut microbial communities. The first is a cohort of healthy (H) individuals, while the second is a cohort of individuals affected by gastrointestinal tract diseases, which we will generally refer to as unhealthy (U).

Microbiome metadata selection and analysis

We have selected gut microbiome data from three studies [4, 10, 27], where both sequencing data and sample metadata are available for controls and three gastrointestinal tract diseases: Crohn's Disease, Ulcerative Colitis and Inflammatory Bowels Syndromes. In the following, we will refer to the controls as the H group and all samples from pathological conditions as the U group. We have filtered so as to have a homogeneous and not biased dataset (see S1 Data for details). In general, we have selected patients who were less affected (at least at the time of the study) by medical treatments, to limit the impact of different drug treatments on the 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 from dysbiotic microbiomes.

We have implemented (for details, see S3 Text) a computational pipeline to process throughput sequencing metagenomic data following best practices [32], such as quality filtering (removing reads with $Q < 20$) and human DNA decontamination using the NCBI human genome assembly (GRCH38).

The metagenomic taxonomic profiling tool we have adopted in our analysis is the *Kaiju* classifier [26], which converts metagenomic reads in all possible open reading frames and searches for the best match in a protein database. The advantage of this approach is that, due to the degeneracy of the genetic code, it is robust to random mutations along the genome and, as such, to evolutionary divergences between the dataset and the reference catalogue of genomes. As a reference species catalogue, we have used RefSeq [33], which contains protein sequences from complete archaeal and bacterial genomes. Metagenomic samples were profiled on October 13th 2020. 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 the different species (S) as rows and the samples (R) as columns. 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 less than the relative abundance cut-off κ, i.e. $v_i^j = 0$ if $v_i^j < \kappa$.

Results

For ease of reference, Table 1 provides definitions for all the used acronyms, see also S1 Table.

Mean abundance distribution and Taylor's law

We find that a similar P_{MAD} is observed in both the H and U dataset, and its shape depends on the relative abundance cut-off κ. In particular, for $\kappa < 10^{-5}$ the MAD displays a *Log-Laplace* shape, i.e. $P_{\mathrm{MAD}}(\bar{x}|\mu,\lambda) = e^{-\frac{|\log \bar{x}-\mu|}{\lambda}}/(2\lambda\bar{x})$, while for $\kappa > 10^{-5}$ the MAD is a Log-Normal distribution $P_{\mathrm{MAD}}(\bar{x}|\mu,\lambda) = e^{-\frac{(\log \bar{x}-\mu)^2}{2\lambda^2}}/(\bar{x}\sqrt{2\pi\lambda^2})$, the same found for OTU 16s data [13]. These distributions indicate a high heterogeneity in mean abundances having both heavy tails. On the y-axis of Fig 2 we show the Byesian Information Criterion (*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

Table 1. List of abbreviations and their descriptions.

Abbrev.	Description
H, U	Healthy, Unhealthy
OTU	Operational Taxonomic Unit
MAD	Mean Abundance Distribution
TL	Taylor's Law
SAD	Species Abundance Distribution
AO	Abundance-Occurrence
SAR	Species-Area Relation
BIC	Bayesian Information Criterion
PSLG	Poisson Stochastic Logistic Growth, Model
MD	Multinomial Dirichlet, Distribution
MSSD	Multinomial Symmetric Scaled Dirichlet, Distribution

https://doi.org/10.1371/journal.pcbi.1012482.t001

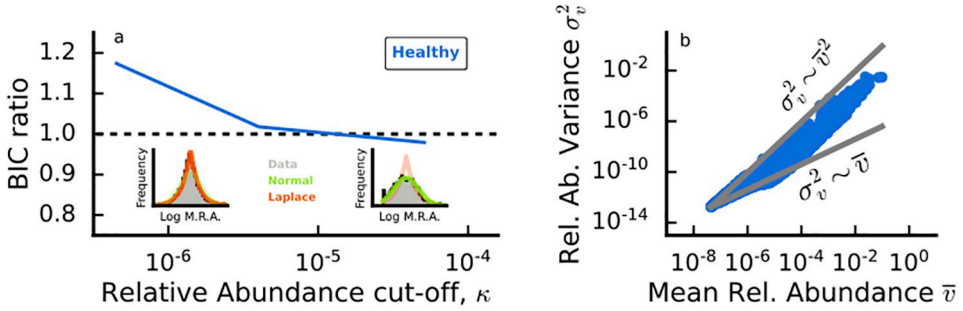

Fig 2. Panel a: *Mean Abundance Distribution* shape displays a dependence on the relative abundance cut-off κ. The BIC ratio curve for healthy and disease-related data collapse onto the same line. Panel b: Taylor's Law holds in empirical human gut communities, both in health and in disease. For rare species, it is difficult to discriminate between Poisson-like and Taylor-like scaling, due to the fact that rare species are present only in a few samples. For simplicity, we report the scatter plot only in the healthy case. A brief discussion of the fitting procedure can be found in S2 Text.

https://doi.org/10.1371/journal.pcbi.1012482.g002

κ. As can also be seen from Fig 2, the Laplace distribution is usually a better description of the MAD, except for a large threshold where the MAD clearly displays a Log-Normal shape (and compatibly with the OTU case [13]). In particular, by fitting P_{MAD} we find $\lambda_H = 1.407 \pm 0.005$ and $\lambda_U = 1.413 \pm 0.007$ (uncertainty of the fit evaluated with a bootstrap procedure, for details see S2 Text).

Regarding the TL, we find that the value of inferred A depends on the threshold κ. On the other hand, the exponent $\zeta \approx 2$ is remarkably robust for all κ and for H and U samples. In particular, for each value of κ we compare the R^2-score ratio of the best fit power-law to that with fixed $\zeta = 2$ finding $\frac{R^2_{\zeta=2}}{R^2_{fit}} \approx 1$ suggesting a negligible discrepancy between the two models for this scaling relation. Therefore, in the following, we assume $\zeta_{Data} = 2$.

Emergent ecological patterns in healthy and unhealthy microbiomes

In this section, we investigate macro-ecological emergent patterns in gut microbiomes, and test which model can describe them, and whether there are 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 [34], 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 that a species with mean relative abundance $\bar{\nu}$ 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 number of species observed in a local community normalized to the meta-community one (α/γ-diversity, also known as Whittaker's beta diversity) and its sequencing depth (i.e., the metagenomic version of the species-area relationship).

To compare a given ecological pattern obtained from the data with the corresponding one produced by a model, we first set the scaling exponent ζ ($\zeta = 2$ for MSSD and SLG, $\zeta = 1$ for MD). Second, we fit the parameters μ, λ, A (in the case of MSSD, due to model symmetries and as explained in S2 Text, fitting μ is not necessary) from the data with no cut-off ($\kappa = 0$). We then 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-offs κ both in the empirical and simulated data (i.e.,

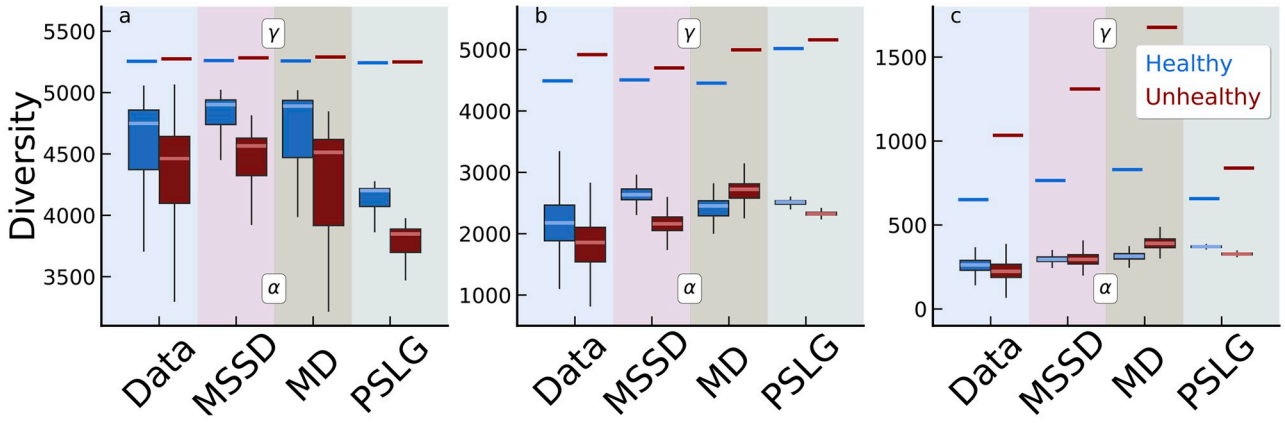

Fig 3. Box-whiskers plots describe the average local (α) diversity, while horizontal bars indicate the corresponding metacommunity (γ) diversity. In healthy gut microbiomes (in blue) we find higher average α and lower γ diversity than unhealthy ones (in red). The three panels represent different threshold relative abundance cut-off κ: a) low ($\kappa_l = 4.5 \times 10^{-7}$); b) medium ($\kappa_m = 9 \times 10^{-6}$); c) high ($\kappa_h = 1.8 \times 10^{-4}$). In each panel, we compare the diversity of the empirical H and U diversity with respect to the one generated by three different null models: MSSD, SLG, and MD. MSSD is found to be the best model, especially for low and medium κ.

https://doi.org/10.1371/journal.pcbi.1012482.g003

all species with $\nu_i < \kappa$ are set to zero) and for each pattern we calculate a R^2-like score (see S1 Text). The final R^2-score we assign to the model is the average of all instances of the model.

The first patterns we consider are the γ and α diversities. The values of such quantities are strongly dependent on κ, but reveal a persistent qualitative regularity among the three regimes. Indeed, the overall (γ) diversity of species found in the unhealthy meta-community is larger than that found in healthy microbiomes ($\gamma_H < \gamma_U$) for all κ. On the other hand, the *average* local diversity ($\bar{\alpha}$) found for the H samples is larger compared to that of the U samples, i.e., ($\bar{\alpha}_H > \bar{\alpha}_U$) (see Fig 3).

The second pattern we consider is the SAD [25], which describes, in a given sample the probability of observing species with a given abundance. In agreement with previous results obtained with 16s OTU data [13, 35] and also with shotgun data [36], we find that the SAD displays a heavy tail, that is compatible with small and medium cut-off with power-law distributions with exponents around 1.7 (see Fig G in S1 File for more details). For large thresholds, the SAD is more compatible with a Log-Normal distribution. No significant differences are observed between the H and U individuals (see Fig 4a). Moreover, all the models (PSLM, MSSD, and MD) generate SADs that are compatible with the empirical ones. These results confirm [37, 38] that SADs are not informative patterns of the underlying ecological mechanisms driving species abundances.

We then investigate the relation between the log-mean relative abundance of a species and its occupancy, which we refer to as the abundance-occupancy (AO) curve. We have already introduced the average relative abundance of species i as $\bar{\nu}_i = \frac{1}{R}\sum_{j=1}^{R} \nu_i^j$, while we define the *occupancy* of species i as $\bar{o}_i = \frac{1}{R}\sum_{j=1}^{R} \theta(\nu_i^j)$, where θ is the Heaviside Theta, which converts relative abundance data into presence/absence ones. The relation between $\bar{\nu}$ and $\bar{o}$, as shown in Fig 4b, describes how likely it is for a species, given its average relative abundance, to be sampled in a realization of the system. This relation is also known as intensity-sparsity relation [14]. When the low/medium value of κ is set, the curve suddenly saturates, suggesting that a large proportion of the available $S = \gamma$ species are 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

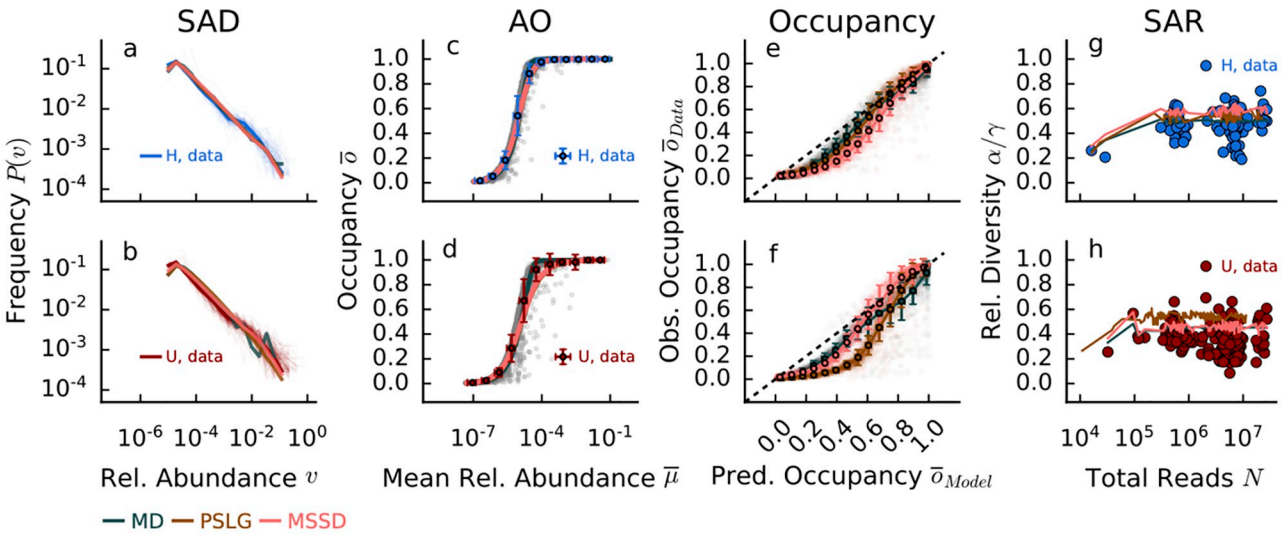

Fig 4. Comparison of emergent empirical ecological patterns in healthy (H, top panels) and unhealthy (bottom panels) microbiomes and for MSSD, SLG and MD models. Panels a-b) Species Abundance distribution (SAD); c-d) Species abundances occupancy curve (AO). Grey shaded points refer to single species mean relative abundance and occupancy; e-f) Empirical vs. predicted occupancy curves (O). Shaded points refer to the predicted/observed occupancy of single species. To compare the occupancy of simulated and observed species, we sort the simulated species according to their mean abundance; g-h) Species Area Relationship (SAR) curves. These patterns can be investigated for different cut-offs (see Fig A-D S1 File), here we show the average threshold cut-off $\kappa_m = 9 \times 10^{-6}$.

https://doi.org/10.1371/journal.pcbi.1012482.g004

fewer species that are rare in relative abundance but, at the same time, are harder to sample. As Fig 4 shows (panels c, d), this behavior is common to all models and thus does not discriminate against the underlying ecological processes. However, for low (high) κ the MD typically underestimates (overestimates) the occupancy of species (for details see Fig A in S1 File).

We can also directly compare the occupancy curves obtained from the presence-absence data with those predicted by the models (as shown in Fig 4e and 4f). Interestingly, contrary to the previous case, here there are differences in how well the models describe the empirical emergent patterns. In particular, for H samples, the PSLG model outperforms the MSSD one, whereas MSSD better predicts the pattern of U samples. Such results and goodness of fit for the MD model are not robust to different thresholds (see Table 2 and Fig A in S1 File).

Finally, we consider the "metagenomic" version of the species-area-relation (SAR) [39], 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 the number of reads combining the samples of each group and calculate the normalized diversity as the number of

Table 2. R^2 scores for the considered macroecological patterns and for three different thresholds: Low κ_l, medium κ_m and high κ_h.

		AO Curve			Species Occupancy			SAR Curve		
		κ_l	κ_m	κ_h	κ_l	κ_m	κ_h	κ_l	κ_m	κ_h
H	MSSD	0.997	0.996	0.983	0.850	0.823	0.830	0.819	0.695	-15.131
	PSLG	0.998	0.995	0.983	0.908	0.912	0.811	0.984	0.585	-22.655
	MD	0.937	0.995	0.944	0.516	0.925	0.337	0.776	0.005	-59.763
U	MSSD	0.992	0.982	0.987	0.885	0.861	0.776	0.976	0.840	-0.779
	PSLG	0.987	0.982	0.984	0.911	0.571	0.569	0.988	0.522	-1.244
	MD	0.902	0.993	0.864	0.562	0.872	0.485	0.874	0.726	-1.302

https://doi.org/10.1371/journal.pcbi.1012482.t002

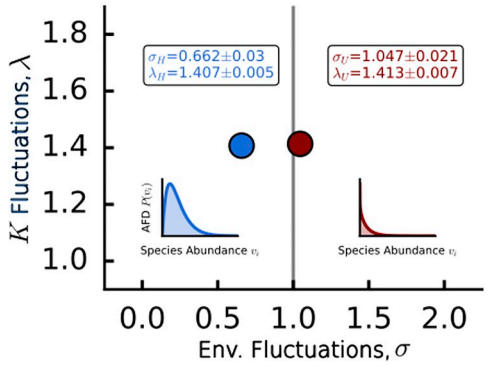

Fig 5. The parameters of Taylor's law and mean abundance distribution are informative about the underlying stochastic logistic model. The values obtained from the data suggest that healthy and unhealthy species abundance distribution follow different qualitative behaviors, with the unhealthy case being prone to more extinction. Average and standard deviation estimates of the parameters have been obtained through a bootstrap procedure.

https://doi.org/10.1371/journal.pcbi.1012482.g005

unique species in the aggregated community divided by the γ diversity. Although all models on average slightly overestimate the overall diversity, in H samples they all perform similarly, while for U samples MSSD and MMD models (which can generate the same number of total reads as found in the data) outperform the SLG model.

Table 2 summarizes all the results and the comparison between the goodness of fit (measured as explained in S1 Text) of the three models for the presented emergent empirical ecological patterns of gut microbiomes in both health and disease states.

Finally, by fitting the MSSD, we can gain insight into possible differences in species population dynamics within microbiomes of healthy and diseased individuals. As reported in Fig 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 statistically indistinguishable between the H and U cohorts. However, the most relevant difference comes from the value we infer for σ, the width of the environmental fluctuations. The stochastic logistic growth model Eq (1) predicts that the abundance fluctuations distribution has a polynomial part with an exponent 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).

Discussion and conclusion

By inferring the parameters of the models that best describe these patterns, we have obtained ecological insights into dysbiosis that would not be directly accessible from the data. In particular, we have found that while the intrinsic logistic population dynamics are similar in the H and U cohorts (they have very similar carrying capacities determined by λ), fluctuations in growth rates due to extrinsic environmental factors (given by σ) are much stronger in the U microbiomes. This is reflected in the abundance fluctuations distribution shifting from a modal distribution in H samples to a power-law one (with exponential cut-off) in U microbiomes, i.e., the probability of a species being very rare is higher in dysbiosis.

This result allows us to explain why in the H cohorts we observe a higher α, but a lower γ diversity: the species in the U microbiomes experience higher fluctuations, and are thus more prone to local extinctions, but are also subject to higher turnover (thus increasing the global diversity of the group). Therefore, we believe that for dysbiois the celebrated Anna Karenina

principle "All happy families are alike; each unhappy family is unhappy in its own way" (Lev Tolstoj, Anna Karenina, 1877) also holds: 'All healthy gut microbiomes are alike; each unhealthy gut microbiome is unhealthy in its own way". In fact, there are convergent observations suggesting that dysbiosis can be attributed to host-specific factors [40, 41]. We also have implemented a stratification analysis (see Fig A-C in S2 File), where we investigated whether the distinct diversity patterns for healthy and unhealthy microbiomes also hold if only specific gastrointestinal diseases are considered. We have observed that while Chrohn's disease (CD) and ulcerative colitis (UC) exhibit trends consistent with our general findings (lower average α diversity and higher γ diversity in U patients), inflammatory bowel syndrome (IBS) presents less systematic tendencies. This distinction is noteworthy given the clinical challenges associated with the diagnosis of IBS. We have also performed a stratified inference of the environmental noise (σ) and carrying capacity heterogeneity (λ) parameters, finding results compatible with the previous ones. For the different cohorts, there is no substantial difference in λ; On the other hand, we have found that CD and UC are associated with an environmental noise strength σ close to one and larger than healthy microbiomes, supporting our interpretation discussed above. Again, IBS has a behavior more similar to that of healthy individuals.

We have also shown that microbiome species abundance data exhibit a Taylor law with a ζ value of 2. Interestingly, the MD model, by its design, does not satisfy this fundamental constraint. However, as shown in Table 2, the MD model is capable of generating AO curves and SADs that are compatible with the data at any threshold. Similarly, it can accurately reproduce Occupancy curves at a medium threshold and SAR curves at a low threshold. This finding highlights that not all patterns and thresholds are equally informative. Some are more effective than others in differentiating between models and underlying ecological processes. Although AO curves have previously been used to test specific underlying ecological theories [42], our results suggest that the shape of the AO curve is simply the result of two main ingredients: heterogeneous population averages and random sampling. Similarly, all patterns at low thresholds are dominated by a large number of rare species (which in shotgun data are probably false positives [15]).

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 [25, 43].

For the SAR curves there is a very strong impact of the thresholds on the models goodness of fit. At low thresholds, all models performed well, thus not providing any discriminatory power for the right choice of the model. However, at the high threshold, there was a significant drop in the goodness of fit for all models. In fact, due to removal of all the rare species, the SAR loses its characteristic shape, and thus it is not useful for models comparison. At the intermediate threshold, the MSSD model performed the best, although it is important to note that the R^2 value is lower compared to its maximum $R^2 = 0.819$. In this case, we also have found that all models fit U samples better than H samples. This effect is probably because in U samples the γ diversity is higher, and we have many more rare species, thus increasing the slope of the SAR (that is, in general, overestimated by the models).

Upon closer examination of the Species Occupancy patterns in Table 2, notable differences emerged between healthy and unhealthy samples at the medium threshold. In the H samples, the SLG model showed the highest goodness of fit, closely followed by the MSSD model. On the contrary, the unhealthy samples showed a different pattern. The SLG model, which performed strongly in the healthy samples, showed a marked decrease in its goodness of fit, indicating potential challenges in capturing the complexity of species occupancy in unhealthy systems at this threshold. The MSSD model also showed a reduction in performance, but remained relatively more consistent compared to the SLG model. The performance of the MD

model for both H and U samples was extremely variable, depending on the cut-off κ. For the medium threshold, the fit—although not as good as the one of the MSSD and SLG models—had a relatively high R^2, whereas for the low and high thresholds, it decreased markedly.

In our study, we thus have found that, unlike to AO curves and SADs, species occupancy and diversity curves provide key insights into the performance of various models. Models incorporating Taylor's law with $\zeta = 2$ (SLG and MSSD) offer a better explanation of the data at the medium threshold (which has proven to be the most informative cut-off for false positives). This suggests that large species fluctuations, as dictated by Taylor's law with a scaling exponent of $\zeta = 2$, are important for accurately predicting presence/absence patterns and species diversity in empirical datasets. We also have found that models that perform well in healthy communities may not necessarily do so in unhealthy ones, and vice versa. This insight is crucial for ecological modeling and could guide future research in developing or choosing models that are tailored to the specific conditions of the ecological systems being studied. Furthermore, we have found that the SLG model, although effectively similar to the MSSD in many respects, underestimates species occupancy (see Fig 4f) and overstimates species diversity (see Fig 4g). The reason is that SLG can generate ecological communities with a number of individuals that is only on average as the one of the corresponding sampled data, while MSSD implements strict compositionality of the data.

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

Supporting information

S1 Table. List of abbreviations.
(PDF)

S1 Data. Data selection, processing and analysis.
(PDF)

S1 File. Supplementary figures.
(PDF)

S2 File. Stratification analysis.
(PDF)

S1 Text. Model testing.
(PDF)

S2 Text. Supplementary methods.
(PDF)

S3 Text. Metagenomic pipeline implementation.
(PDF)

Acknowledgments

CloudVeneto is acknowledged for the use of computing and storage facilities.

Author Contributions

Conceptualization: Jacopo Pasqualini, Samir Suweis.

Data curation: Jacopo Pasqualini, Sonia Facchin, Edoardo Savarino.

Formal analysis: Jacopo Pasqualini, Samir Suweis.

Funding acquisition: Edoardo Savarino, Samir Suweis.

Investigation: Jacopo Pasqualini.

Methodology: Jacopo Pasqualini, Amos Maritan, Samir Suweis.

Project administration: Samir Suweis.

Supervision: Andrea Rinaldo, Amos Maritan, Edoardo Savarino, Samir Suweis.

Visualization: Jacopo Pasqualini.

Writing – original draft: Jacopo Pasqualini, Samir Suweis.

Writing – review & editing: Jacopo Pasqualini, Sonia Facchin, Andrea Rinaldo, Amos Maritan, Edoardo Savarino, Samir Suweis.

References

1. A framework for human microbiome research. nature. 2012; 486(7402):215–221. https://doi.org/10.1038/nature11209 PMID: 22699610
2. Structure, function and diversity of the healthy human microbiome. nature. 2012; 486(7402):207–214. https://doi.org/10.1038/nature11234 PMID: 22699609
3. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The human microbiome project. Nature. 2007; 449(7164):804–810. https://doi.org/10.1038/nature06244 PMID: 17943116
4. The integrative human microbiome project. Nature. 2019; 569(7758):641–648. https://doi.org/10.1038/s41586-019-1238-8 PMID: 31142853
5. Duvallet C, Gibbons SM, Gurry T, Irizarry RA, Alm EJ. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. Nature communications. 2017; 8(1):1784. https://doi.org/10.1038/s41467-017-01973-8 PMID: 29209090
6. Gilbert JA, Blaser MJ, Caporaso JG, Jansson JK, Lynch SV, Knight R. Current understanding of the human microbiome. Nature medicine. 2018; 24(4):392–400. https://doi.org/10.1038/nm.4517 PMID: 29634682
7. A review of 10 years of human microbiome research activities at the US National Institutes of Health, Fiscal Years 2007-2016. Microbiome. 2019; 7(1):31. https://doi.org/10.1186/s40168-019-0620-y
8. Manor O, Dai CL, Kornilov SA, Smith B, Price ND, Lovejoy JC, et al. Health and disease markers correlate with gut microbiome composition across thousands of people. Nature communications. 2020; 11 (1):5206. https://doi.org/10.1038/s41467-020-18871-1 PMID: 33060586
9. Gilbert JA, et al. Microbiome-wide association studies link dynamic microbial consortia to disease. Nature. 2016; 535(7610):94–103. https://doi.org/10.1038/nature18850 PMID: 27383984
10. Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature. 2019; 569(7758):655–662. https://doi.org/10.1038/s41586-019-1237-9 PMID: 31142855
11. Fan Y, Pedersen O. Gut microbiota in human metabolic health and disease. Nature Reviews Microbiology. 2021; 19(1):55–71. https://doi.org/10.1038/s41579-020-0433-9 PMID: 32887946
12. Tierney BT, et al. Systematically assessing microbiome-disease associations identifies drivers of inconsistency in metagenomic research. Plos Biol. 2022; 20:e3001556. https://doi.org/10.1371/journal.pbio.3001556 PMID: 35235560

13. Grilli J. Macroecological laws describe variation and diversity in microbial communities. Nature communications. 2020; 11(1):1–11. https://doi.org/10.1038/s41467-020-18529-y PMID: 32958773

14. Baruzzo G, Patuzzi I, Di Camillo B. Beware to ignore the rare: how imputing zero-values can improve the quality of 16S rRNA gene studies results. BMC bioinformatics. 2021; 22(Suppl 15):618.

15. Tovo A, Menzel P, Krogh A, Cosentino Lagomarsino M, Suweis S. Taxonomic classification method for metagenomics based on core protein families with Core-Kaiju. Nucleic acids research. 2020; 48(16): e93–e93. https://doi.org/10.1093/nar/gkaa568 PMID: 32633756

16. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are compositional: and this is not optional. Frontiers in microbiology. 2017; 8:2224. https://doi.org/10.3389/fmicb.2017.02224 PMID: 29187837

17. Tsilimigras MC, Fodor AA. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. Annals of epidemiology. 2016; 26(5):330–335. https://doi.org/10.1016/j.annepidem.2016.03.002 PMID: 27255738

18. Swift D, Cresswell K, Johnson R, Stilianoudakis S, Wei X. A review of normalization and differential abundance methods for microbiome counts data. Wiley Interdisciplinary Reviews: Computational Statistics. 2023; 15(1):e1586. https://doi.org/10.1002/wics.1586

19. Harrison JG, Calder WJ, Shastry V, Buerkle CA. Dirichlet-multinomial modelling outperforms alternatives for analysis of microbiome and other ecological count data. Molecular ecology resources. 2020; 20 (2):481–497. https://doi.org/10.1111/1755-0998.13128 PMID: 31872949

20. Xiao Y, Angulo MT, Friedman J, Waldor MK, Weiss ST, Liu YY. Mapping the ecological networks of microbial communities. Nature communications. 2017; 8(1):2042. https://doi.org/10.1038/s41467-017-02090-2 PMID: 29229902

21. Maynard DS, Miller ZR, Allesina S. Predicting coexistence in experimental ecological communities. Nature ecology & evolution. 2020; 4(1):91–100. https://doi.org/10.1038/s41559-019-1059-z

22. Ansari AF, Reddy YB, Raut J, Dixit NM. An efficient and scalable top-down method for predicting structures of microbial communities. Nature Computational Science. 2021; 1(9):619–628. https://doi.org/10.1038/s43588-021-00131-x PMID: 38217133

23. Michel-Mata S, Wang XW, Liu YY, Angulo MT. Predicting microbiome compositions from species assemblages through deep learning. Imeta. 2022; 1(1):e3. https://doi.org/10.1002/imt2.3 PMID: 35757098

24. Descheemaeker L, De Buyl S. Stochastic logistic models reproduce experimental time series of microbial communities. Elife. 2020; 9:e55650. https://doi.org/10.7554/eLife.55650 PMID: 32687052

25. Azaele S, Suweis S, Grilli J, Volkov I, Banavar JR, Maritan A. Statistical mechanics of ecological systems: Neutral theory and beyond. Reviews of Modern Physics. 2016; 88(3):035003. https://doi.org/10.1103/RevModPhys.88.035003

26. Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nature communications. 2016; 7(1):11257. https://doi.org/10.1038/ncomms11257 PMID: 27071849

27. Mars RA, Yang Y, Ward T, Houtti M, Priya S, Lekatz HR, et al. Longitudinal multi-omics reveals subset-specific mechanisms underlying irritable bowel syndrome. Cell. 2020; 182(6):1460–1473. https://doi.org/10.1016/j.cell.2020.08.007 PMID: 32916129

28. Suweis S, Ferraro F, Azaele S, Maritan A. Generalized Lotka-Volterra Systems with Time Correlated Stochastic Interactions. arXiv preprint arXiv:230702851. 2023;.

29. Aitchison J. The statistical analysis of compositional data. Journal of the Royal Statistical Society: Series B (Methodological). 1982; 44(2):139–160. https://doi.org/10.1111/j.2517-6161.1982.tb01195.x

30. G S Monti VPG G Mateu-Figueras, Egozcue JJ. The shifted-scaled Dirichlet distribution in the simplex. In: Proceedings of the 4th International Workshop on Compositional Data Analysis; 2011.

31. Giometto A, Formentin M, Rinaldo A, Cohen JE, Maritan A. Sample and population exponents of generalized Taylor's law. Proceedings of the National Academy of Sciences. 2015; 112(25):7755–7760. https://doi.org/10.1073/pnas.1505882112 PMID: 25941384

32. Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, et al. Best practices for analysing microbiomes. Nature Reviews Microbiology. 2018; 16(7):410–422. https://doi.org/10.1038/s41579-018-0029-9 PMID: 29795328

33. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic acids research. 2016; 44(D1):D733–D745. https://doi.org/10.1093/nar/gkv1189 PMID: 26553804

34. Costello EK, Stagaman K, Dethlefsen L, Bohannan BJ, Relman DA. The application of ecological theory toward an understanding of the human microbiome. Science. 2012; 336(6086):1255–1262. https://doi.org/10.1126/science.1224203 PMID: 22674335

35. Patuzzi I, Baruzzo G, Losasso C, Ricci A, Di Camillo B. metaSPARSim: a 16S rRNA gene sequencing count data simulator. BMC bioinformatics. 2019; 20(9):1–13. https://doi.org/10.1186/s12859-019-2882-6 PMID: 31757204

36. Sala C, Vitali S, Giampieri E, do Valle ÌF, Remondini D, Garagnani P, et al. Stochastic neutral modelling of the Gut Microbiota's relative species abundance from next generation sequencing data. BMC bioinformatics. 2016; 17(2):179–188. https://doi.org/10.1186/s12859-015-0858-8 PMID: 26821617

37. Rosindell J, Hubbell SP, He F, Harmon LJ, Etienne RS. The case for ecological neutral theory. Trends in ecology & evolution. 2012; 27(4):203–208. https://doi.org/10.1016/j.tree.2012.01.004 PMID: 22341498

38. Seppi M, Pasqualini J, Facchin S, Savarino EV, Suweis S. Emergent functional organization of gut microbiomes in health and diseases. Biomolecules. 2023; 14(1):5. https://doi.org/10.3390/biom14010005 PMID: 38275746

39. Coleman BD. On random placement and species-area relations. Mathematical Biosciences. 1981; 54 (3-4):191–215. https://doi.org/10.1016/0025-5564(81)90086-9

40. Sonnenburg ED, Smits SA, Tikhonov M, Higginbottom SK, Wingreen NS, Sonnenburg JL. Diet-induced extinctions in the gut microbiota compound over generations. Nature. 2016; 529(7585):212–215. https://doi.org/10.1038/nature16504 PMID: 26762459

41. Gilbert JA, Lynch SV. Community ecology as a framework for human microbiome research. Nature medicine. 2019; 25(6):884–889. https://doi.org/10.1038/s41591-019-0464-9 PMID: 31133693

42. Sieber M, Pita L, Weiland-Bräuer N, Dirksen P, Wang J, Mortzfeld B, et al. Neutrality in the metaorganism. PLoS Biology. 2019; 17(6):e3000298. https://doi.org/10.1371/journal.pbio.3000298 PMID: 31216282

43. Purves DW, Pacala SW, et al. Ecological drift in niche-structured communities: neutral pattern does not imply neutral process. Biotic interactions in the tropics. 2005; p. 107–138. https://doi.org/10.1017/CBO9780511541971.006