Title:

Health and disease imprinted in the time variability of the human microbiome

Running title:

Microbiota, are you sick?

Jose Manuel Martí[1,2], Daniel M. Martínez[1,2,3], Manuel Peña[2], César Gracia[1,2], Amparo Latorre[1,3,4,5], Andrés Moya[1,3,4,5] & Carlos P. Garay[1,2,#]

[1]Institute for Integrative Systems Biology (I2SysBio), 46980, Spain.

[2]Instituto de Fisica Corpuscular, CSIC-UVEG, P.O. 22085, 46071, Valencia, Spain.

[3]FISABIO, Avda de Catalunya, 21, 46020, Valencia, Spain.

[4]Cavanilles Institute of Biodiversity and Evolutionary Biology, Univ. de Valencia, 46980, Spain.

[5]CIBER en Epidemiología y Salud Pública (CIBEResp), Madrid, Spain

Words count for the Abstract section: 134 of 250 max

Words count for the Importance section: 105 of 150 max

Words count for the rest of text: 5095 of 5000 max

[#] Corresponding author: penagaray@gmail.com

Abstract

Human microbiota plays an important role in determining changes from health to disease. Increasing research activity is dedicated to understand its diversity and variability. We analyse 16S rRNA and whole genome sequencing (WGS) data from the gut microbiota of 97 individuals monitored in time. Temporal fluctuations in the microbiome reveal significant differences due to factors that affect the microbiota such as dietary changes, antibiotic intake, early gut development or disease. Here we show that a fluctuation scaling law describes the temporal variability of the system and that a noise-induced phase transition is central in the route to disease. The universal law distinguishes healthy from sick microbiota and quantitatively characterizes the path in the phase space, which opens up its potential clinical use and, more generally, other technological applications where microbiota plays an important role.

Importance

Human microbiota is tightly associated to the health status of a person. Here we analyse the microbial composition of several subjects under different conditions, over a time span that ranges from days to months. Using the Langevin equation as the basis of our mathematical framework in order to evaluate microbial temporal stability, we prove that we are capable to distinguish stable from unstable microbiotas. This first step will help us to determine how microbiota temporal stability is related to the healthiness of the people, and it will allow the development of a more complete framework in order to deepen the knowledge of this complex system.

Keywords— microbiome, systems biology, ecological modelling, community composition, stability

Introduction

The desire to understand the factors that influence human health and cause diseases has always been one of the major driving forces of biological research. Modern high-throughput sequencing and bioinformatics tools provide a powerful means of understanding how the human microbiome contributes to health and its potential as a target for therapeutic interventions. High throughput methods for microbial 16S ribosomal RNA gene and WGS have now begun to reveal the composition of archaeal, bacterial, fungal and viral communities located both, in and on the human body. Biology has recently acquired new technological and conceptual tools to investigate, model and understand living organisms at the system level, thanks to the spectacular progress in quantitative techniques, large-scale measurement methods and the integration of experimental and computational approaches. Systems Biology has mostly been devoted to the study of well-characterized model organisms but, since the early days of the Human Genome Project it has become clear that applications of system-wide approaches to Human Biology would bring huge opportunities in Medicine. Here we present the imprints of disease in macroscopic properties of the system, by studying the temporal variability in the microbiome.

Results

We have analysed the microbiome temporal variability to extract global properties of the system. As fluctuations in total counts are plagued by systematic errors we worked on temporal variability of relative abundances for each taxon. Our first finding was that, in all cases, changes in relative abundances of taxa follow a universal pattern known as the fluctuation scaling law (*9*) or Taylor's power law (*10*), i.e., microbiota of all detected taxa follows a power law dependence between mean relative abundance x_i and dispersion σ_i, $\sigma_i = V \cdot x_i^{\beta}$. While the law is universal, spanning six orders of magnitude in the observed relative abundances, the power law (or scaling) index β and the variability V (hereafter Taylor parameters) appear to be correlated with the stability of the community and the

health status of the host, which is the main finding in this letter (see Figure 1). Taylor parameters describing the temporal variability of the gut microbiome in our sampled individuals are shown in ST1-6. Our results hint at a universal behaviour. Firstly, the variability (which corresponds to the maximum amplitude of fluctuations) is large, which suggests resilient capacity of the microbiota, and the scaling index is always smaller than one, which means that, more abundant taxa are less volatile than less abundant ones. Secondly, Taylor parameters for the microbiome of healthy individuals in different studies are compatible within estimated errors. This enables us to define the health zone in the Taylor parameter space. We can better visualize the results of individuals from different studies by standardizing their Taylor parameters, where standardization means that each parameter is subtracted by the mean value and divided by the standard deviation of the group of healthy individuals in the study (see Supplementary Section 12 and ST1-6). The zone of health and the standardized Taylor parameters for individuals whose gut microbiota is threatened (i.e., suffering from kwashiorkor, altered diet, antibiotics, IBS) is shown in Figure 1. Children developing kwashiorkor show smaller variability than their healthy twins. A meat/fish-based diet increases the variability significantly when compared to a plant-based diet. All other cases presented increased variability, which is particularly severe, and statistically significant at more than 95% CL, for obese patients grade III on a diet, individuals taking antibiotics or IBS–diagnosed patients. A global property emerges from all worldwide data collected: Taylor parameters characterize the statistical behaviour of microbiome changes. We have verified that our conclusions are robust to systematic errors due to taxonomic assignment.

Taylor's power law has been explained in terms of various effects, all without general consensus. It can be shown to have its origin in a mathematical convergence similar to the central limit theorem, so virtually any statistical model designed to produce a Taylor law converge to a Tweedie distribution (*11*), providing a mechanistic explanation based on the statistical theory of errors (*12–14*). To unveil the generic mechanisms that drive different scenarios in the β–V space, we model the system by assuming that taxon

relative abundance follows a Langevin equation with a deterministic term that captures the fitness of each taxon and a randomness term with Gaussian random noise (*15*). Both terms are modelled by power laws, with coefficients that can be interpreted as the taxon fitness F_i and the variability V (see Section 1 in supplemental material). When V is sufficiently low, abundances are stable in time. Differences in variability V can induce a noise-induced phase transition in relative abundances of taxa. The temporal evolution of the probability of a taxon having abundance x_i given its fitness is governed by the Fokker–Planck equation. The results of solving this equation show that the stability is best captured by fitness F and amplitude of fluctuations V phase space (see Figure 2).

The model predicts two phases for the gut microbiome: a stable phase with large variability that permits some changes in the relative abundances of taxa and an unstable phase with larger variability, above the phase transition, where the order of abundant taxa varies significantly with time. The microbiome of all healthy individuals was found to be in the stable phase, while the microbiome of several other individuals was shown to be in the unstable phase. In particular, individuals taking antibiotics and IBS–diagnosed patient P2 had the most severe symptoms. In this phase diagram, each microbiota state is represented by a point at its measured variability V and inferred fitness F. The model predicts high average fitness for all taxa, i.e., taxa are narrowly distributed in F. The fitness parameter has been chosen with different values for demonstrative purposes. Fitness is larger for the healthiest subjects and smaller for the IBS–diagnosed patients.

Discussion

We have quantitatively characterized whether the microbiota belongs to a healthy individual or a subject corresponding to an altered or pathological state (i.e., altered diet, antibiotic treatment, early gut development, diagnosed IBS). Deciphering the mechanisms of disease requires in depth knowledge of the underlying biological mechanisms. We describe here the macroscopic behavior of disease by a noise-induced phase transition with a control parameter that can be measured by the temporal variability of the

microbiome. The microbiota of healthy individuals and of individuals with pathologies represent different phases separated by this noise-induced phase transition. Improved high?throughput sequencing of samples from individuals monitored over time and taxonomic assigning methods will provide a better distinction among pathologies or altered states of the microbiota.

Temporal evolution of model parameters

We have studied the time dependence of the variability V and power law index β (see Section) by using a sliding window approach. The total number of time points are divided in subsets of five points, where next subset is defined by adding next time sampling and by eliminating the earliest one. Both parameters were calculated for each subset against the average time lapse. Figure 3 shows the variability V as a function of time for the largest sampling: two individuals in the Caporaso's study (*1*) corresponding to the gut microbiota of a male (upper plot) and a female (lower plot). Figure 4 shows the time evolution of V for patient P2 of the IBS study (*6*) (upper plot) and patient D in the antibiotics study (*5*) (lower plot).

Materials and Methods

We have analysed more than 35000 time series of taxa from the gut microbiomes of 97 individuals (sampling from three up to 332 time points), obtained from publicly available high throughput sequencing data on: healthy individuals over a long term span (*1*), people with various degrees of obesity (*2*), twin pairs discordant for kwashiorkor (*3*), response to diet changes (*4*) or to antibiotic perturbation (*5*), and subjects diagnosed with irritable bowel syndrome (IBS) (*6*). We engineered a complete software framework and a web platform, ComplexCruncher, ready to be implemented by other users.

We summarize our dataset in ST1-6 (Tables 1-6 in suplemental material). The bacteria and archaea taxonomic assignations were obtained by analysing 16S rRNA sequences,

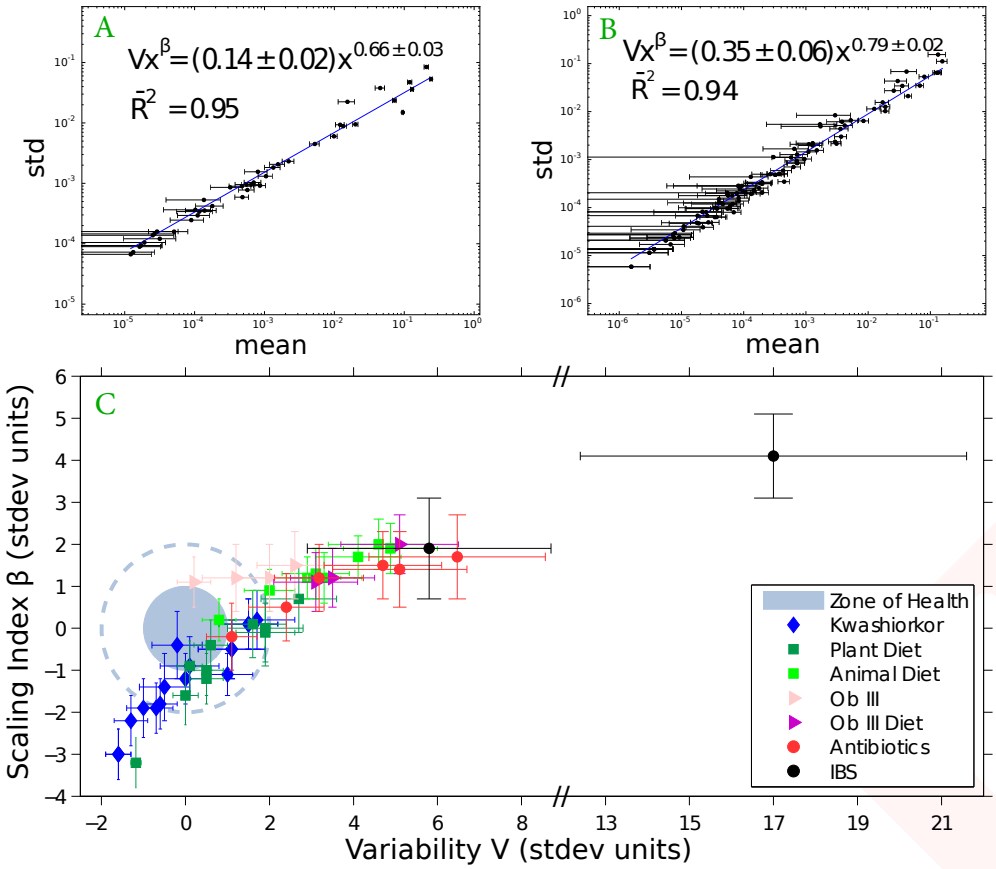

Figure 1. Taylor's law parameter space. X-weighted power-law fits of the standard deviations versus the mean values for each bacterial genus monitored in time. We show the fit for samples from a healthy subject (Figure A) and from a subject diagnosed with irritable bowel syndrome (Figure B) studied in our lab (**?**). We have compiled all data studied in this work in Figure C. The coloured region (dashed line) corresponds to 68% (95%) CL region of healthy individuals in the Taylor parameter space. Points with errors place each individual gut microbiome in the Taylor space. Note that the parameters have been standardized (stdev units) to the healthy group in each study for demonstrative purposes.

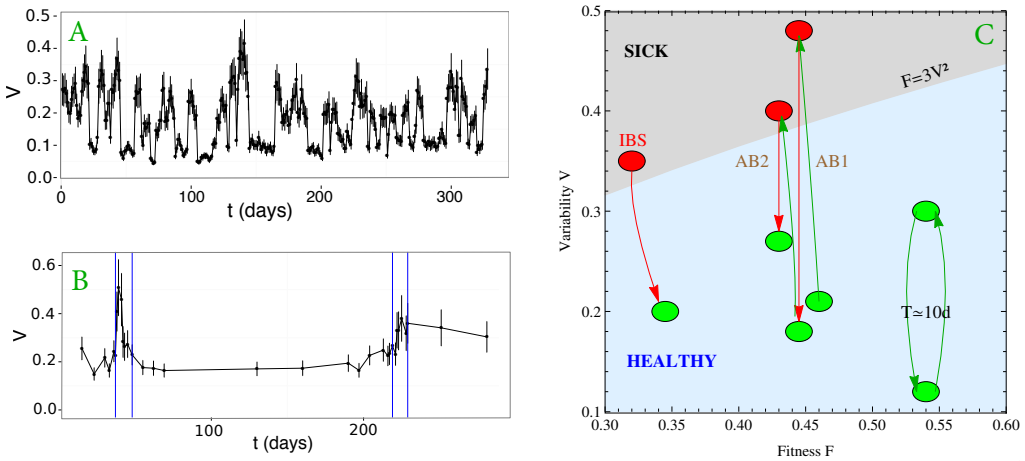

Figure 2. Phase space of microbiota stability. We show the Variability parameter as a function of time for the male in Caporaso's study (*1*) (Figure A) and for the patient D in the antibiotics study (*5*) (Figure B). Periods of antibiotic treatment are shown by blue vertical lines. Microbiota states can be placed in the phase space F-V (Figure C). The light blue shaded region corresponds to the stable phase, while the grey shaded region is the unstable phase (the phase transition line is calculated for $\alpha = \beta = 0.75$). We place healthy individuals (green) and individuals whose gut microbiota is threatened (antibiotics, IBS) in the phase space fitness - variability. Gut microbiota of healthy individuals over a long term span show a quasi–periodical variability (central period is ten days). We show that taking antibiotics (AB1 and AB2 correspond to first and second treatment respectively) induces a phase transition in the gut microbiota, which impacts its future changes. We also show an IBS–diagnosed patient transiting from the unstable to the stable phase.

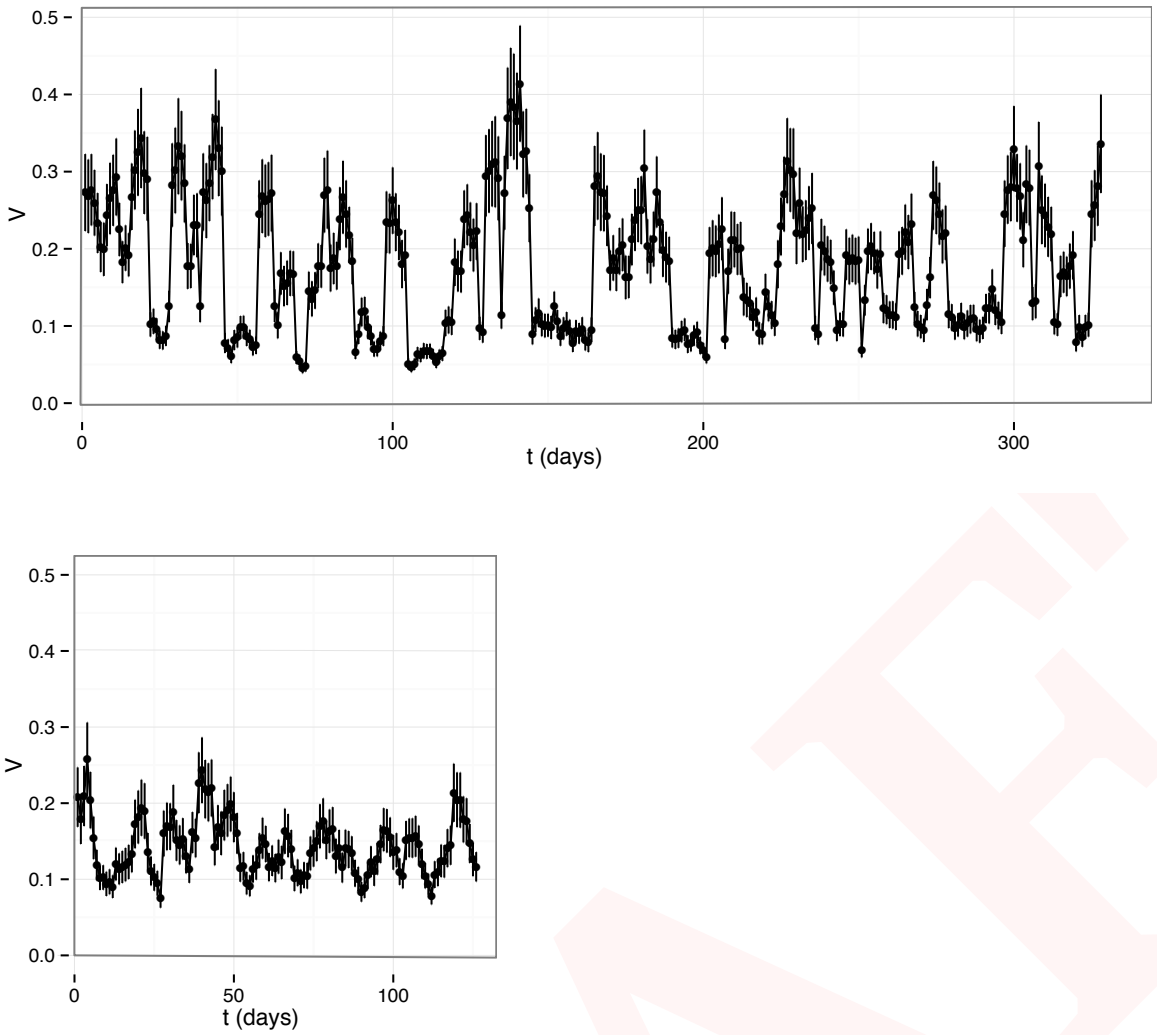

Figure 3. *V* as a function of time for the two individuals in the Caporaso's study (*1*): samples of gut microbiome of a male (upper plot) and a female (lower plot). Both samples show changes in the variability V with quasi–periodic behavior peaked at about 10 days. Variability grows more for the gut microbiota of the male and share a minimal value around 0.1 with the gut microbiota of the female.

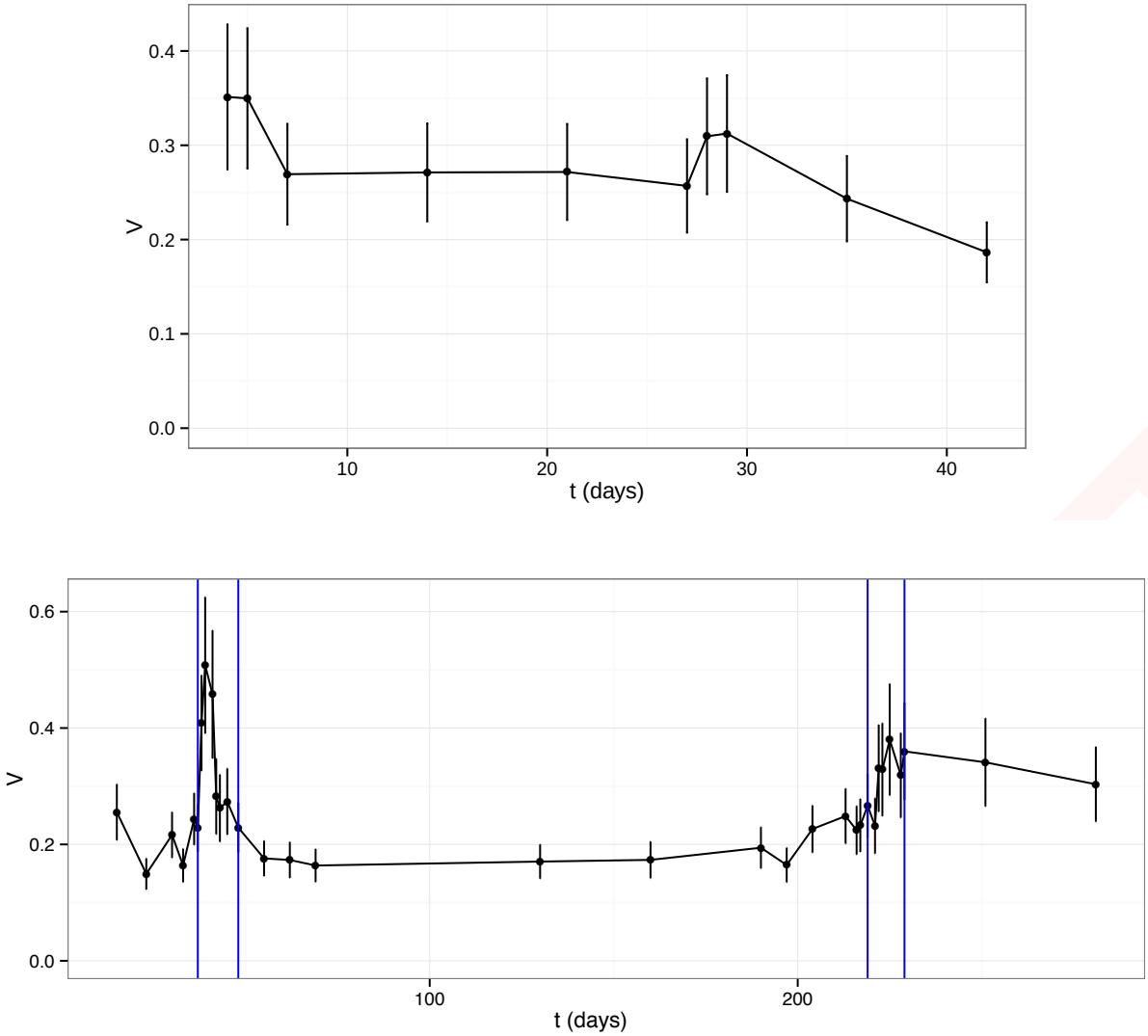

Figure 4. *V* as a function of time for patient P2 of the IBS study (*6*) (upper plot) and patient D in the antibiotics study (*5*) (lower plot). The variability of the gut microbiota of P2 decreases from above 0.3 to below 0.2, showing a slow tendency to increase the order of the system. Antibiotic intake leaks to a quick increase of variability which lasts for a few days to recover ordering. The second antibiotic treatment shows some memory (lower increase of variability) with a slower recovery. NOTE: The blue vertical lines in the lower plot are showing the periods of antibiotic treatment.

which were clustered into operational taxonomic units (OTUs) sharing 97 % sequence identity using QIIME (*7*). WGS data (*3*) were analysed and assigned at strain level by the Livermore Metagenomic Analysis Toolkit (LMAT) (*8*), according to their default quality threshold. Genus, with best balance between error assignment and number of taxa, was chosen as our reference taxonomic level. We have verified that our conclusions are not significantly affected by selecting family or species as the reference taxonomic level (see Figure 1 in supplemental material).

Specify, in each study treated, the nature of the samples (conditions, timespan between timepoints, subjects). Specify, and it is very important, what we consider ?healthy? in each study (for example: pre-antibiotics is healthy)

Model

We model the microbial abundances across time along the lines of Blumm *et al.* (*15*). The dynamics of taxon relative abundances is described by the Langevin equation:

$$\dot{x_i} = F_i \cdot x_i^{\alpha} + V \cdot x_i^{\beta} \xi_i(t) - \phi(t) \cdot x_i, \tag{1}$$

where F_i captures the fitness of the taxon i, V corresponds to the noise amplitude and ξ_i(t) is a Gaussian random noise with zero mean $< \xi_i(t) > = 0$ and variance uncorrelated in time, $< \xi_i(t)\xi_i(t') > = \delta(t'-t)$, . The function $\phi(t)$ ensures the normalization at all times, $\sum x_i(t) = 1$, and corresponds to $\phi(t) = \sum F_i x_i^{\alpha} + \sum V x_i^{\beta} \xi_i(t)$. The temporal evolution of the probability that a taxon i has a relative abundance $x_i(t)$, P(x_i,t), is determined by the Fokker-Planck equation:

$$\frac{\partial P}{\partial t} = -\frac{\partial}{\partial x_i}[(F_i \cdot x_i^{\alpha} - \phi(t) \cdot x_i) \cdot P] + \frac{1}{2}\frac{\partial^2}{\partial x_i^2}(V^2 \cdot x_i^{2\beta} \cdot P). \tag{2}$$

The microbiota evolves towards a steady-state with a time-independent probability depending on the values of α, β, F_i and V. For $\alpha < 1$ (otherwise, systems are always unsta-

ble), the steady-state probability may be localized in a region around a preferred value or broadly distributed over a wide range, depending on whether the fitness F_i dominates or is overwhelmed by the noise amplitude V. The steady-state solution of the Fokker-Planck equation is given by:

$$P_0(x_i) = C_{ne}(\alpha,\beta,F_i,V)\cdot x_i^{-2\beta}\cdot\exp\left[\frac{2F_i}{V^2}\frac{x_i^{1+\alpha-2\beta}}{1+\alpha-2\beta}-\frac{\phi_0}{V^2}\frac{x_i^{2-2\beta}}{1-\beta}\right] \quad \text{if} \quad 2\beta\neq 1+\alpha,$$

$$P_0(x_i) = C_e(\alpha,\beta,F_i,V)\cdot x_i^{\frac{2F_i}{V^2}-2\beta}\cdot\exp\left[\frac{\phi_0}{V^2}\frac{x_i^{2-2\beta}}{1-\beta}\right] \quad \text{if} \quad 2\beta=1+\alpha,$$

where $\phi_0=(\sum_i F_i^{1/(1-\alpha)})^{1-\alpha}$ and C_{ne} and C_e are integrals that should be solved numerically for the parameters of interest. The ordered phase happens when the solution has a maximum in the physical interval $(0<x_i<1)$. For larger V, the transition to a disordered phase happens when the maximum shifts to the unphysical region $x_i<0$, which sets the phase transition region $V(\alpha,\beta,F_i)$. The phase transition region can be calculated analytically in particular cases:

$$F_i^2 = 4\beta\phi_0 V^2 \quad \text{if} \quad \beta=\alpha\neq 1,$$

$$F_i = \beta V^2 \quad \text{if} \quad 2\beta=1+\alpha,$$

where the first case, simplifies to $F=3V^2$ if $\beta=0.75$ and the fitness of this taxon dominates in ϕ_0. In many physical systems (Brownian motion is the classical example), the two terms of the Langevin equation are related. The *fluctuation–dissipation theorem* states a general relationship between the response to an external disturbance and the internal fluctuations of the system (*16*). The theorem can be used as the basic formula to derive the fitness from the analysis of fluctuations of the microbiota, assuming that it is in equilibrium (the ordered phase).

Explain better the fluctiation-dissipation theorem

Selection and Methods

Sample selection

We have chosen studies about relevant pathologies containing metagenomic sequencing time data series of bacterial populations from humans in different healthy and non-healthy states. We have selected only those individuals who had three or more time points of data available in databases. Metadata of each study is provided in Supplementary Tables 1 to 6. All used 16S rRNA gene sequencing except for the study of the discordant kwashiorkor twins (*3*) (see Supplementary Tables 4 and 5) where shotgun metagenomic sequencing (SMS) and 16S rRNA were used. In the latter case we selected to work with SMS data to show that our method is valid regardless of the source of taxonomic information. Each one of the datasets was treated as follows:

16rRNA sequences processing

Reads from the selected studies were first quality filtered using the FastX toolkit (*17*), allowing only those reads which had more than 25 of quality along the 75% of the complete sequence. 16S rRNA reads were then clustered at 97% nucleotide sequence identity (97% ID) into operational taxonomic units (OTUs) using QIIME package software (*7*) (version 1.8) We followed open reference OTU picking workflow in all cases. The clustering method used was uclust, and the OTUs were matched against Silva database (*18*) (version 111, July 2012) and were assigned to taxonomy with an uclust-based consensus taxonomy assigner. The parameters used in this step were: similarity 0.97, prefilter percent id 0.6, max accepts 20, max rejects 500.

Metagenomic sequences processing

Metagenomic shotgun (and 16S too) sequences were analyzed with LMAT (Livermore Metagenomics Analysis Toolkit) software package (*8*) (version 1.2.4, with Feb'15 release of data base *LMAT-Grand*). LMAT was run using a Bull shared-memory node belonging

to the team's HPC (high performance computing) cluster. It is equipped with 32 cores (64 threads available using Intel Hyper-threading technology) as it has 2 Haswell-based Xeons, the E5-2698v3@2.3 GHz, sharing half a tebibyte (0.5 TiB, that is, 512 gibibytes) of DRAM memory. This node is also provided with a card PCIe SSD as NVRAM, the P420m HHHL, with 1.4 TB, and 750000 reading IOPS, 4 KB, achieving 3.3 GB/s, which Micron kindly issued free of charge, as a sample for testing purposes. The computing node was supplied with a RAID-0 (striping) scratch disk area. We used the "Grand" database[I], release Feb'15, provided by the LMAT team. Previously to any calculation, the full database was loaded in the NVRAM. With this configuration the observed LMAT sustained sequence classification rate was 20 kpb/s/core. Finally, it is worth mentioning that a complete set of Python scripts have been developed as back-end and front-end of the LMAT pipeline in order to manage the added complexity of time series analysis.

Taxa level selection

We selected genus as taxonomic level for the subsequent steps of our work. In order to ensure that, between adjacent taxonomic levels, there were not crucial differences which could still be of relevance after standardization (see Section), we tested two different data sets. In the former, the antibiotics study (*5*) with 16S data, we tested the differences between genus and family levels. The latter dataset tested was the kwashiorkor discordant twins study (*3*) for both genus and species taxonomic levels. The Supplementary Figures 5 (overview) and 6 (detail) plot the comparison between studies (and so, 16S and SMS) and between adjacent taxonomic levels.

ComplexCruncher

A complete software framework, named 'ComplexCruncher', has been engineered to support the analysis of the dynamics of ranking processes in complex systems. Although the

[I] In this context, "Grand" refers to a huge database that contains k-mers from all viral, prokaryote, fungal and protist genomes present in the NCBI database, plus Human reference genome (hg19), plus GenBank Human, plus the 1000 Human Genomes Project (HGP). This represent about 31.75 billion k-mers occupying 457.62 GB.

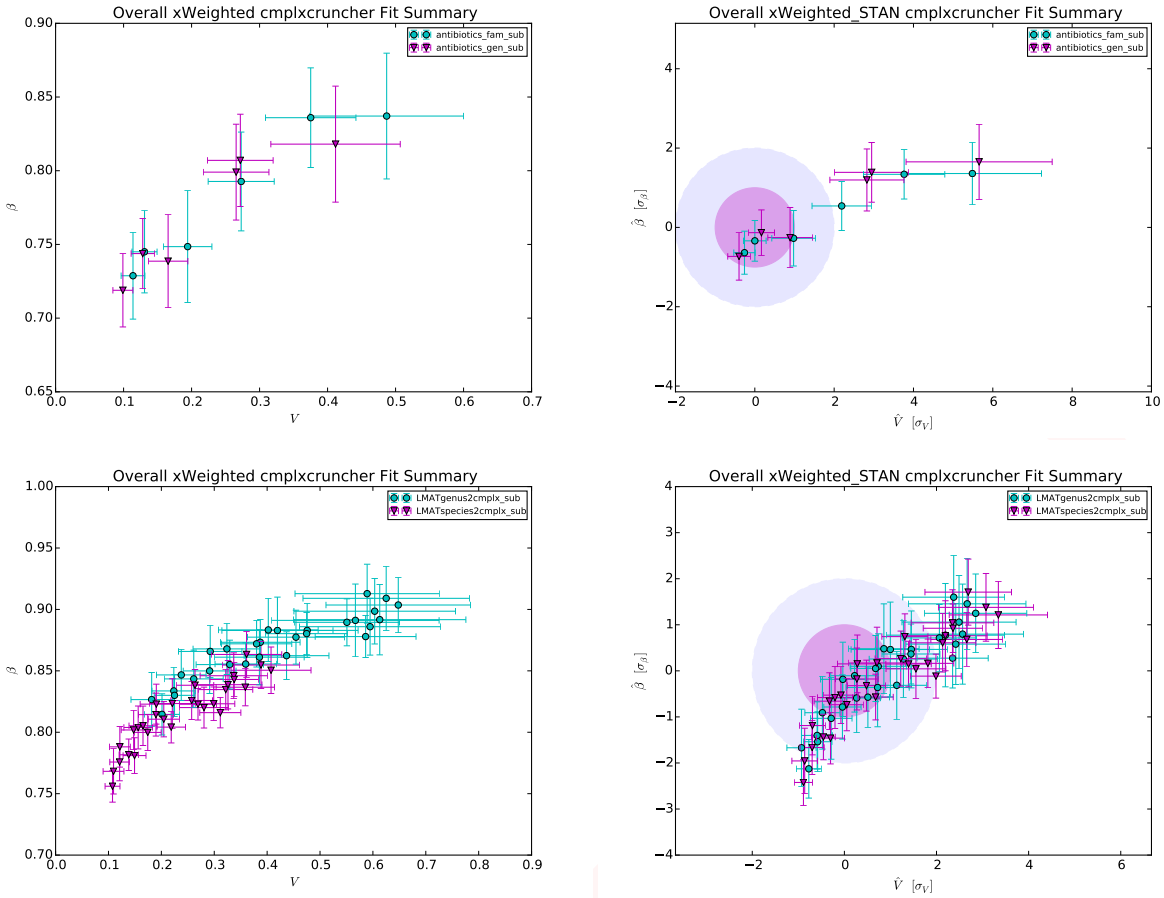

Figure 5. Overview of comparison of different approaches based on adjacent taxonomic levels using plots in the Taylor-parameters space. For 16S (former row of subfigures), the levels are family vs. genus, whereas for SMS (latter row of subfigures) levels are genus vs. species. The left column shows the raw results and the right column plots the standardized results (see Section)

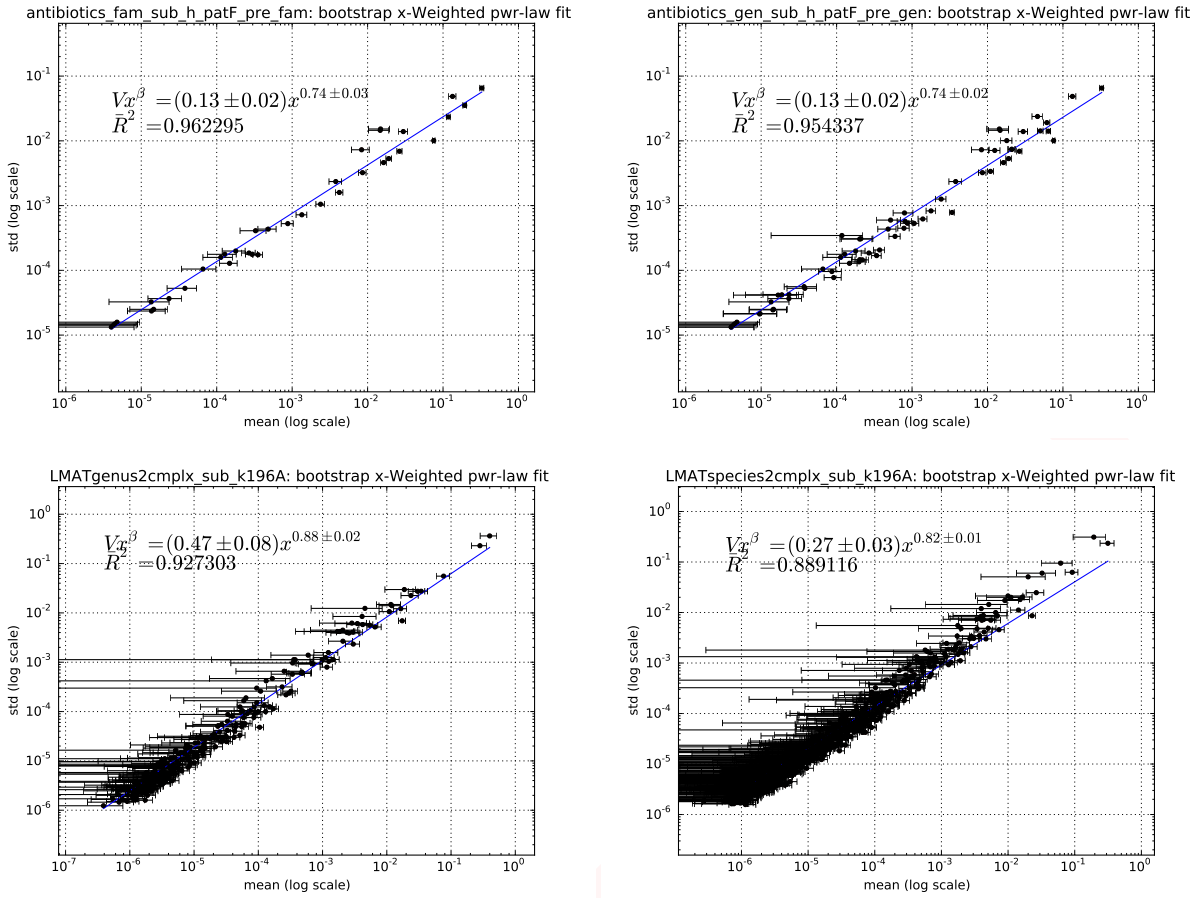

Figure 6. Detail of comparison of different approaches based on adjacent taxonomic levels using plots of X-weighted power-law fits (see Section). The former row of subfigures shows examples for 16S, whereas the latter row of subfigures plots examples for SMS. The left column shows results for the superior taxonomic level (family for 16S, genus for SMS), while the right column shows results for the inferior level (genus for 16S, specie for SMS).

Metadata	V	β	$\bar{R}^2$	V_{st}	β_{st}
A	0.26 ± 0.05	0.826 ± 0.025	0.918	3.1 ± 0.9	1.2 ± 0.6
A	0.32 ± 0.06	0.857 ± 0.025	0.924	4.4 ± 1.1	2.0 ± 0.6
A	0.194 ± 0.033	0.813 ± 0.024	0.918	1.9 ± 0.6	0.9 ± 0.6
A	0.24 ± 0.04	0.824 ± 0.020	0.924	2.7 ± 0.7	1.2 ± 0.5
A	0.34 ± 0.06	0.855 ± 0.024	0.931	4.7 ± 1.1	1.9 ± 0.6
A	0.30 ± 0.05	0.847 ± 0.022	0.921	3.9 ± 1.0	1.7 ± 0.5
A	0.133 ± 0.021	0.784 ± 0.023	0.916	0.7 ± 0.4	0.2 ± 0.6
A	0.25 ± 0.04	0.831 ± 0.024	0.929	3.0 ± 0.8	1.4 ± 0.6
P	0.23 ± 0.05	0.804 ± 0.035	0.885	2.6 ± 0.9	0.7 ± 0.8
P	0.097 ± 0.018	0.705 ± 0.031	0.891	0.03 ± 0.34	-1.6 ± 0.7
P	0.037 ± 0.006	0.642 ± 0.025	0.881	-1.12 ± 0.11	-3.1 ± 0.6
P	0.118 ± 0.019	0.723 ± 0.025	0.895	0.4 ± 0.4	-1.2 ± 0.6
P	0.17 ± 0.04	0.78 ± 0.04	0.842	1.5 ± 0.7	0.1 ± 0.9
P	0.123 ± 0.020	0.757 ± 0.026	0.914	0.5 ± 0.4	-0.4 ± 0.6
P	0.19 ± 0.05	0.77 ± 0.04	0.871	1.8 ± 0.9	-0.0 ± 0.9
P	0.121 ± 0.020	0.736 ± 0.027	0.921	0.5 ± 0.4	-0.9 ± 0.6
P	0.187 ± 0.034	0.771 ± 0.030	0.908	1.8 ± 0.7	-0.1 ± 0.7
P	0.097 ± 0.015	0.735 ± 0.025	0.922	0.05 ± 0.28	-0.9 ± 0.6

Table 1. Taylor parameters. Individuals with either animal-based (A) or plant-based (P) diets (*4*). Previous to diet, the population sampled is described by $\bar{V} = 0.09 \pm 0.05, \bar{\beta} = 0.77 \pm 0.04$.

Metadata	V	β	$\bar{R}^2$	V_{st}	β_{st}
Ab	0.35 ± 0.07	0.81 ± 0.04	0.925	4.3 ± 1.4	1.3 ± 0.9
Ab	0.41 ± 0.09	0.82 ± 0.04	0.908	5.6 ± 1.8	1.6 ± 0.9
Ab	0.23 ± 0.04	0.770 ± 0.031	0.920	2.1 ± 0.8	0.5 ± 0.7
Ab	0.165 ± 0.029	0.738 ± 0.031	0.928	0.9 ± 0.6	-0.3 ± 0.7
Ab	0.34 ± 0.06	0.812 ± 0.032	0.936	4.1 ± 1.2	1.5 ± 0.7
Ab	0.26 ± 0.05	0.798 ± 0.033	0.931	2.8 ± 0.9	1.1 ± 0.8

Table 2. Taylor parameters for individuals taking antibiotics (*5*). Prior to antibiotics intake, the population sampled is described by $\bar{V} = 0.12 \pm 0.05, \bar{\beta} = 0.75 \pm 0.04$.

Metadata	V	β	$\bar{R}^2$	V_{st}	β_{st}
IBS	0.204 ± 0.034	0.739 ± 0.029	0.916	7.6 ± 3.7	1.9 ± 1.2
IBS	0.35 ± 0.05	0.793 ± 0.023	0.935	23.1 ± 5.9	4.0 ± 0.9

Table 3. Taylor parameters for persons diagnosed with irritable bowel syndrome (IBS) (*6*). Healthy individuals sampled in this study are characterized by $\bar{V} = 0.134 \pm 0.009, \bar{\beta} = 0.691 \pm 0.025$.

Metadata	V	β	$\bar{R}^2$	V_{st}	β_{st}
DH	0.27 ± 0.04	0.835 ± 0.016	0.925	0.2 ± 0.4	-1.0 ± 0.6
DH	0.36 ± 0.06	0.858 ± 0.015	0.929	1.1 ± 0.6	-0.2 ± 0.5
DH	0.35 ± 0.06	0.859 ± 0.014	0.926	1.0 ± 0.5	-0.1 ± 0.5
DH	0.25 ± 0.04	0.829 ± 0.014	0.911	0.0 ± 0.4	-1.2 ± 0.5
DH	0.30 ± 0.05	0.844 ± 0.014	0.920	0.5 ± 0.4	-0.7 ± 0.5
DH	0.29 ± 0.05	0.850 ± 0.016	0.915	0.4 ± 0.5	-0.5 ± 0.5
DH	0.28 ± 0.05	0.848 ± 0.016	0.921	0.3 ± 0.5	-0.5 ± 0.6
DH	0.35 ± 0.07	0.861 ± 0.017	0.918	0.9 ± 0.6	-0.0 ± 0.6
DH	0.31 ± 0.04	0.833 ± 0.012	0.916	0.6 ± 0.4	-1.1 ± 0.4
DH	0.33 ± 0.05	0.843 ± 0.013	0.925	0.8 ± 0.5	-0.7 ± 0.5
DH	0.31 ± 0.05	0.852 ± 0.014	0.925	0.6 ± 0.5	-0.4 ± 0.5
DH	0.31 ± 0.05	0.853 ± 0.015	0.930	0.6 ± 0.5	-0.4 ± 0.5
DH	0.203 ± 0.033	0.815 ± 0.015	0.907	-0.44 ± 0.32	-1.7 ± 0.5

Table 4. Taylor parameters for the healthy subject of the discordant twins (*3*). This table continues in Supplementary Table 5. The population of healthy twins is characterized by $\bar{V} = 0.25 \pm 0.10, \bar{\beta} = 0.863 \pm 0.028$.

Metadata	V	β	$\bar{R}^2$	V_{st}	β_{st}
DK	0.40 ± 0.07	0.859 ± 0.017	0.926	1.5 ± 0.7	-0.1 ± 0.6
DK	0.44 ± 0.08	0.868 ± 0.016	0.919	1.8 ± 0.8	0.2 ± 0.6
DK	0.196 ± 0.031	0.819 ± 0.014	0.916	-0.50 ± 0.30	-1.5 ± 0.5
DK	0.160 ± 0.026	0.798 ± 0.015	0.904	-0.85 ± 0.25	-2.3 ± 0.5
DK	0.30 ± 0.05	0.845 ± 0.014	0.924	0.5 ± 0.4	-0.6 ± 0.5
DK	0.23 ± 0.04	0.834 ± 0.014	0.908	-0.1 ± 0.4	-1.0 ± 0.5
DK	0.27 ± 0.05	0.848 ± 0.015	0.930	0.2 ± 0.4	-0.5 ± 0.5
DK	0.35 ± 0.07	0.860 ± 0.019	0.916	1.0 ± 0.7	-0.1 ± 0.7
DK	0.34 ± 0.05	0.835 ± 0.012	0.917	0.9 ± 0.5	-1.0 ± 0.4
DK	0.25 ± 0.04	0.831 ± 0.012	0.912	0.0 ± 0.4	-1.1 ± 0.4
DK	0.36 ± 0.06	0.858 ± 0.013	0.918	1.1 ± 0.5	-0.2 ± 0.5
DK	0.31 ± 0.06	0.851 ± 0.016	0.924	0.6 ± 0.6	-0.4 ± 0.6
DK	0.149 ± 0.022	0.799 ± 0.013	0.905	-0.96 ± 0.22	-2.2 ± 0.5

Table 5. Taylor parameters for the kwashiorkor part of the discordant twins (*3*). This is a continuation of Supplementary Table 4. The population of healthy twins is characterized by $\bar{V} = 0.25 \pm 0.10, \bar{\beta} = 0.863 \pm 0.028$.

software was devised with a clear bias towards metagenomics, it is general enough to be able to cope with a ranking process in any complex system. Implemented in Python using well-known open-source community software, the software solution is composed of two parts that can be used together or apart: a web-based graphic front-end connected to a database, and a computing kernel. Used together, this software enables other users to reproduce our results easily and, furthermore, upload and analyse their own data or experiment with the preloaded metagenomics data sets.

'ComplexCruncher WebPortal' (CCWebPortal) is a web platform designed to allow the user to interact with a data repository of selected and well-documented metagenomics data sources. Through a few simple steps, the user can perform advanced searches on the complete set of records in the metagenomics repository. The web application provides advanced filters that allow the user to reduce the search to a small set of interest. After this first step, the user can refine the search and discard those records that do not meet certain requirements.

Metadata	V	β	$\bar{R}^2$	V_{st}	β_{st}
OW	0.59 ± 0.12	0.894 ± 0.034	0.920	6.6 ± 2.0	2.6 ± 1.0
OW	0.22 ± 0.04	0.830 ± 0.030	0.904	0.5 ± 0.6	0.7 ± 0.9
OBI	0.28 ± 0.04	0.855 ± 0.022	0.958	1.5 ± 0.6	1.4 ± 0.6
OBI	0.33 ± 0.07	0.870 ± 0.031	0.916	2.4 ± 1.1	1.9 ± 0.9
OBII	0.223 ± 0.032	0.823 ± 0.023	0.938	0.6 ± 0.5	0.5 ± 0.7
OBII	0.208 ± 0.029	0.844 ± 0.022	0.935	0.4 ± 0.5	1.1 ± 0.7
OBIII	0.34 ± 0.05	0.855 ± 0.025	0.943	2.5 ± 0.9	1.4 ± 0.7
OBIII	0.26 ± 0.04	0.845 ± 0.026	0.954	1.1 ± 0.7	1.2 ± 0.8
OBIII	0.33 ± 0.06	0.870 ± 0.027	0.908	2.4 ± 1.0	1.9 ± 0.8
OBIII	0.200 ± 0.026	0.843 ± 0.020	0.949	0.2 ± 0.4	1.1 ± 0.6
OBIII	0.30 ± 0.05	0.846 ± 0.026	0.929	1.9 ± 0.8	1.2 ± 0.7
OBIII	0.176 ± 0.029	0.826 ± 0.026	0.894	-0.2 ± 0.5	0.6 ± 0.8
OBIII	0.30 ± 0.06	0.841 ± 0.031	0.896	1.8 ± 0.9	1.0 ± 0.9
OBIII	0.28 ± 0.04	0.857 ± 0.025	0.941	1.5 ± 0.7	1.5 ± 0.7
OBIII	0.122 ± 0.018	0.822 ± 0.024	0.930	-1.05 ± 0.30	0.5 ± 0.7
OBIIId	0.47 ± 0.08	0.872 ± 0.023	0.945	4.7 ± 1.3	1.9 ± 0.7
OBIIId	0.38 ± 0.06	0.846 ± 0.023	0.951	3.2 ± 1.0	1.2 ± 0.7
OBIIId	0.36 ± 0.06	0.842 ± 0.022	0.954	2.9 ± 0.9	1.1 ± 0.6

Table 6. Taylor parameters for individuals with different degrees of overweight and obesity (*2*). Healthy people in this study, whom were not obese, are characterized by $\bar{V} = 0.19 \pm 0.06, \bar{\beta} = 0.806 \pm 0.034$.

The web application allows calculations to be done directly by the stable release of the *cmplxcruncher* computing kernel. At the end of the calculations, the results are displayed to the user on the same browser which runs the web application. Then, the user can interact over the series of generated graphics thus allowing flexible comparison among them. In addition, CCWebPortal enables direct download of generated data (plots, spreadsheets, etc). The web application generates a report file summarizing all the results in PDF format. If the user has login permissions, CCWebPortal enables the option of insert new database records in addition to editing and deleting existing ones.

CCWebPortal is a web application that runs on current versions of many browsers. Additional software is not needed and only requires javaScript to be enabled on the browser to run applications. CCWebPortal is implemented following the client—server distributed programming model, where the javaScript client application connects to a remote server that enables the execution of calculations and transactions through a centralized database management system. A set of relational tables allows the structuring of the metagenomics repository to establish relationships between records. Thus the search and information threshing is optimized for queries launched from the client interface. Access to the database on the server is implemented through Django framework, an open-source framework written in Python using the model-view-controller (MVC) architectural pattern for implementing user interfaces.

The effective data analysis has been performed with a Python tool developed from scratch to more than 4200 lines of code. Implemented following the Object Oriented Programming, paradigm, this software is the back-end of the website described above. However, it could be run as an independent piece of software since it is built as a Python package provided with a command-line front-end (*cmplxcruncher*.py). Once installed, the tool can be run interactively but also in automatic mode, which uses parallel computation to speed up the analysis of several data sources.

cmplxcruncher performs the power-law fit described in the *Blumm, N. et al.* paper, but by fitting the best model, i.e. choosing between fitting a power-law using linear regression

versus nonlinear regression (*19*). In the power-law fit plots we also show the generalized coefficient of determination computed for continuous models (*20, 21*).

Un-weighted power-law fit

Fitting the best model

As already mentioned, to choose between fitting power laws ($y = Vx^{\beta}$) using linear regression on log-transformed (LLR) data versus non-linear regression (NLR), we mainly follow *General Guidelines for the Analysis of Biological Power Laws* (*19*). It consists of the following three steps:

1. Determining the appropriate error structure by likelihood analysis.
   (a) Fit the Non-Linear Regression (NLR) model and obtain V_{NLR}, β_{NLR} and σ^2_{NLR}.
   (b) Calculate the loglikelihood that the data (n is sample size) are generated from a normal distribution with additive error:
   - The likelihood of a normal distribution is:

$$\mathscr{L}_{\mathrm{norm}} = \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi\sigma^2_{\mathrm{NLR}}}} \exp\left(-\frac{\left(y_i - V_{\mathrm{NLR}} x_i^{\beta_{\mathrm{NLR}}}\right)^2}{2\sigma^2_{\mathrm{NLR}}} \right) \right]$$

   - So, the loglikelihood of a normal distribution is:

$$\begin{aligned} \log \mathscr{L}_{\mathrm{norm}} &= -\frac{n}{2} \log\left|2\pi\sigma^2_{\mathrm{NLR}}\right| - \frac{1}{2\sigma^2_{\mathrm{NLR}}} \underbrace{\sum_{i=1}^{n} \left(y_i - V_{\mathrm{NLR}} x_i^{\beta_{\mathrm{NLR}}}\right)^2}_{\mathrm{RSS}_{\mathrm{NLR}}} \\ &= -\frac{n}{2} \log\left|2\pi\sigma^2_{\mathrm{NLR}}\right| - \frac{\mathrm{RSS}_{\mathrm{NLR}}}{2\sigma^2_{\mathrm{NLR}}} \end{aligned}$$

   (c) Calculate the *corrected Akaike's Information Criterion* for the NLR model:

$$\mathrm{AIC}_{c_{\mathrm{NLR}}} = 2k - 2\log \mathscr{L}_{\mathrm{norm}} + \frac{2k(k+1)}{n-k-1}$$

   (d) Fit the Log-transformed Linear Regression (LLR) model and obtain V_{LLR}, β_{LLR}

and σ^2_{LLR}.

(e) Calculate the loglikelihood that the data (n is sample size) are generated from a lognormal distribution with multiplicative error:

- The likelihood of a lognormal distribution is:

$$\mathscr{L}_{\text{logn}} = \prod_{i=1}^{n} \left[\frac{1}{y_i \sqrt{2\pi\sigma^2_{\text{LLR}}}} \exp\left(-\frac{(\log|y_i| - \log|V_{\text{LLR}}| - \beta_{\text{LLR}} \log|x_i|)^2}{2\sigma^2_{\text{LLR}}} \right) \right]$$

- So, the loglikelihood of a lognormal distribution is:

$$\begin{aligned} \log \mathscr{L}_{\text{logn}} &= -\frac{n}{2} \log\left|2\pi\sigma^2_{\text{LLR}}\right| - \sum_{i=1}^{n} \log|y_i| - \\ &\qquad -\frac{1}{2\sigma^2_{\text{LLR}}} \underbrace{\sum_{i=1}^{n} (\log|y_i| - \log|V_{\text{LLR}}| - \beta_{\text{LLR}} \log|x_i|)^2}_{\text{RSS}_{\text{LLR}}} \\ &= -\frac{n}{2} \log\left|2\pi\sigma^2_{\text{LLR}}\right| - \frac{\text{RSS}_{\text{LLR}}}{2\sigma^2_{\text{LLR}}} - \sum_{i=1}^{n} \log|y_i| \end{aligned}$$

(f) Calculate the *corrected Akaike's Information Criterion* for the LR model:

$$\text{AIC}_{\text{c}_{\text{LLR}}} = 2k - 2\log \mathscr{L}_{\text{logn}} + \frac{2k(k+1)}{n-k-1}$$

2. Compare $\text{AIC}_{\text{c}_{\text{NLR}}}$ with $\text{AIC}_{\text{c}_{\text{LLR}}}$:

- If $\text{AIC}_{\text{c}_{\text{NLR}}} - \text{AIC}_{\text{c}_{\text{LLR}}} < -2$, the assumption of normal error is favoured compared to lognormal error, so proceed with the results obtained from the NLR fit.
- If $\text{AIC}_{\text{c}_{\text{NLR}}} - \text{AIC}_{\text{c}_{\text{LLR}}} > 2$, the assumption of lognormal error is favoured compared to normal error, so proceed with the results obtained from the LLR fit.
- If $\left|\text{AIC}_{\text{c}_{\text{NLR}}} - \text{AIC}_{\text{c}_{\text{LLR}}}\right| \leq 2$, no model is favoured, so proceed with model averaging:

$$\begin{aligned} B_{\text{av}} &= w_{\text{NLR}} V_{\text{NLR}} + w_{\text{LLR}} V_{\text{LLR}} \\ \beta_{\text{av}} &= w_{\text{NLR}} \beta_{\text{NLR}} + w_{\text{LLR}} \beta_{\text{LLR}} \end{aligned}$$

where:

$$w_{\mathrm{NLR}} = \frac{1}{1+\mathrm{e}^{\frac{1}{2}\left(\mathrm{AIC}_{\mathrm{c}_{\mathrm{NLR}}}-\mathrm{AIC}_{\mathrm{c}_{\mathrm{LLR}}}\right)}}$$
$$w_{\mathrm{LLR}} = \frac{1}{1+\mathrm{e}^{\frac{1}{2}\left(\mathrm{AIC}_{\mathrm{c}_{\mathrm{LLR}}}-\mathrm{AIC}_{\mathrm{c}_{\mathrm{NLR}}}\right)}}$$

which are obtained to fulfill the next condition: $w_{\mathrm{NLR}} + w_{\mathrm{LLR}} = 1$. The CIs for B_{av} and β_{av} are to be generated by ordinary bootstrapping[II].

3. Assess the validity of the underlying statistical assumptions with diagnostic plots because while it is rare for all the assumptions to be fully satisfied by real-life data sets, major violations indicate the lack of appropriateness of the model and, thus, the potential invalidity of the results.

Calculating the coefficient of determination

We think the best approach in this situation is to apply the generalized R^2 that, for continuous models, was defined as (*20*):

$$R^2 = 1 - \left(\frac{\mathscr{L}(0)}{\mathscr{L}(\hat{\theta})}\right)^{\frac{2}{n}}$$

where $\mathscr{L}(\hat{\theta})$ and $\mathscr{L}(0)$ denote the likelihoods of the fitted and the "null" model, respectively, and n is the sample size. In terms of the loglikelihoods, the generalized coefficient of determination would be:

$$R^2 = 1 - \mathrm{e}^{-\frac{2}{n}\left(\log\mathscr{L}(\hat{\theta})-\log\mathscr{L}(0)\right)}$$

We have the likelihoods calculated from the previous section, but what about the "null" models? We understand that they are the models with only the intercept. So for the

[II] *cmplxcruncher* has available the next bootstrapping alternatives (*22*): ordinary, "Resampling Residuals" method, "Wild" method, and "Monte-Carlo" method.

Gaussian additive error model:

$$\mathscr{L}_{\text{norm}}(0) = \prod_{i=1}^{n}\left[\frac{1}{\sqrt{2\pi\sigma^2_{\text{NLR0}}}}\,\exp\left(-\frac{(y_i-\bar{y})^2}{2\sigma^2_{\text{NLR0}}}\right)\right]$$

So:

$$\begin{aligned}\log\mathscr{L}_{\text{norm}}(0) &= -\frac{n}{2}\log\left|2\pi\sigma^2_{\text{NLR0}}\right| - \frac{1}{2\sigma^2_{\text{NLR0}}}\sum_{i=1}^{n}(y_i-\bar{y})^2\\ &= -\frac{n}{2}\left(\log\left|2\pi\sigma^2_{\text{NLR0}}\right|+1\right)\end{aligned}$$

since $\sigma^2_{\text{NLR0}} = \frac{1}{n}\sum(y_i-\bar{y})^2 = \frac{1}{n}\text{TSS}_{\text{NLR}}$. Now, coming back to the coefficient of determination, we have:

$$\begin{aligned}R^2_{\text{NLR}} &= 1-\mathrm{e}^{\frac{2}{n}\left(\log\mathscr{L}_{\text{NLR}}(0)-\log\mathscr{L}_{\text{NLR}}(\hat{\theta})\right)} = 1-\exp\left(\frac{\log(\text{RSS}_{\text{NLR}})}{\log(\text{TSS}_{\text{NLR}})}\right) =\\ &= 1-\frac{\text{RSS}_{\text{NLR}}}{\text{TSS}_{\text{NLR}}} = 1-\frac{\sum_{i=1}^{n}\left(y_i - V_{\text{NLR}}x_i^{\beta_{\text{NLR}}}\right)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}\end{aligned}$$

recovering the traditional expression for R^2. Using the same approach for calculating R^2_{LLR}, then:

$$\mathscr{L}_{\text{logn}}(0) = \prod_{i=1}^{n}\left[\frac{1}{y_i\sqrt{2\pi\sigma^2_{\text{LLR0}}}}\,\exp\left(-\frac{(\log|y_i|-\log|B_{\text{LLR0}}|)^2}{2\sigma^2_{\text{LLR0}}}\right)\right]$$

So:

$$\begin{aligned}\log\mathscr{L}_{\text{logn}}(0) &= -\frac{n}{2}\log\left|2\pi\sigma^2_{\text{LLR0}}\right| - \frac{1}{2\sigma^2_{\text{LLR0}}}\sum_{i=1}^{n}\left(\log|y_i| - \overline{\log|y|}\right)^2 - \sum_{i=1}^{n}\log|y_i|\\ &= -\frac{n}{2}\left(\log\left|2\pi\sigma^2_{\text{LLR0}}\right|+1\right) - \sum_{i=1}^{n}\log|y_i|\end{aligned}$$

since $\sigma^2_{\mathrm{LLR0}} = \frac{1}{n}\sum\left(\log|y_i| - \overline{\log|y|}\right)^2 = \frac{1}{n}\mathrm{TSS}_{\mathrm{logn}}$. Again, recalling the expression for the generalized coefficient of determination, we have:

$$\begin{aligned} R^2_{\mathrm{LLR}} &= 1 - \mathrm{e}^{\frac{2}{n}\left(\log\mathscr{L}_{\mathrm{LLR}}(0) - \log\mathscr{L}_{\mathrm{LLR}}(\hat{\theta})\right)} = 1 - \exp\left(\frac{\log(\mathrm{RSS}_{\mathrm{LLR}})}{\log(\mathrm{TSS}_{\mathrm{LLR}})}\right) = \\ &= 1 - \frac{\mathrm{RSS}_{\mathrm{LLR}}}{\mathrm{TSS}_{\mathrm{LLR}}} = 1 - \frac{\sum_{i=1}^{n}\left(\log|y_i| - \log|V_{\mathrm{LLR}}| - \beta_{\mathrm{LLR}}\log|x_i|\right)^2}{\sum_{i=1}^{n}\left(\log|y_i| - \overline{\log|y|}\right)^2} \end{aligned}$$

X-weighted power-law fit

When fitting the power-law of std vs. mean, we can take into account that every mean has uncertainty and estimate it for a sample size n by the SEM (*Standard Error of the Mean*):

$$\mathrm{SEM} = \frac{s}{\sqrt{n}}$$

where s is the sample standard deviation. So, the vector of weights is computed with:

$$\mathbf{w} = \frac{1}{\overrightarrow{\mathrm{SEM}}} = \frac{\sqrt{\mathbf{n}}}{\mathbf{s}}$$

Here, the uncertainties affect the independent variable, so the fit is not so trivial as a Y-weighted fit, where the uncertainties affect the dependent variable. A standard approach to do this fit is: a) invert your variables before applying the weights, b) then perform the weighted fit, and finally, c) revert the inversion. This method is deterministic, but the approximate solution worsens with smaller R^2. For comparison, we develop a stochastic method by using a bootstrapping-like strategy that avoids the inversion and is applicable regardless of R^2. Both methods, detailed below, are implemented in *cmplxcruncher*.

Method 1: By inverting the data

In the case of the log-LR model, we have:

$$\log y = \log V + \beta \log x \quad \rightarrow \quad \underbrace{\log x}_{\tilde{y}} = \overbrace{-\frac{1}{\beta} \log V}^{b} + \overbrace{\frac{1}{\beta}}^{m} \underbrace{\log y}_{\tilde{x}}$$

where m determines the slope or gradient of the fitted line, and b determines the point at which the line crosses the y-axis, otherwise known as the y-intercept. Once the model is fitted, the original parameters can be retrieved easily:

$$\begin{aligned} \beta &= \frac{1}{m} \\ V &= \mathrm{e}^{-\beta b} = \mathrm{e}^{-\frac{b}{m}} \end{aligned}$$

Their respective uncertainties are to be obtained using *error propagation*:

$$\begin{aligned} \sigma_\beta &= \left| \frac{\mathrm{d}\beta}{\mathrm{d}m} \right| \sigma_m = \frac{1}{m^2} \sigma_m \\ \sigma_V &= \sqrt{\left(\frac{\partial V}{\partial b} \right)^2 \sigma_b^2 + \left(\frac{\partial V}{\partial m} \right)^2 \sigma_m^2} = \frac{1}{m} \mathrm{e}^{-\frac{b}{m}} \sqrt{\sigma_b^2 + \frac{b^2}{m^2} \sigma_m^2} \end{aligned}$$

Method 2: Bootstrapping-like strategy

The basic idea of bootstrapping is that inference about a population from sample data (sample → population) can be modeled by resampling the sample data and performing inference on (resample → sample). To adapt this general idea to our problem, we resample the x-data array using its errors array. That is, for each replicate, a new x-data array is computed based on:

$$x_i^* = x_i + v_i$$

where v_i is a Gaussian random variable with mean $\mu_i = 0$ and standard deviation $\sigma_i = \mathrm{SEM}_i$, as defined previously in this supplementary material. For each replicate a complete un-weighted power-law fit is performed, as described in the previous section. It is

worth mentioning that each replicate is filtered to avoid values of x_i^* under *eps* (obtained by `np.finfo(np.double).eps`) in order to keep away from the error of getting log of negatives or zero during the fit.

We devised and implemented a multi-step algorithm to estimate the fit parameters that finishes when a relative error of less than 10^{-4} is achieved. It also ends if the number of steps reaches 100 to avoid too much time lapse, to prevent any pathologic numeric case which, in fact, we still have not detected in all the data sets analyzed.

In the previous version of the algorithm, for each step, the method generated 10 replicates for each x-data point, in other words, it was computing the fit for 10 times the length of the x-data array replicates, with a maximum of 10000 fits per step. Nevertheless, we found that such an approach depending on the length of the x-data array did not perform better, so we decided to simplify the method and fix the number of fits per step in 100. This latter approach improved the performance.

The parameters of the X-weighted fit are then estimated by averaging through all the replicate fits performed, and their errors are estimated by computing the standard deviation also for all the fits. At the end of each step, the relative error is calculated by comparing the fit parameters estimation in the last step with the previous one.

Finally, both the coefficient of determination of the fit and the coefficient of correlation between the fit parameters are estimated by averaging.

Rank Stability Index (RSI)

The Rank Stability Index is shown as a percentage in a separate bar on the right of the rank matrix plot provided by *cmplxcruncher*. The RSI is strictly 1 for an element whose range never changes over time, and is strictly 0 for an element whose rank oscillates between the extremes from time to time. So, RSI is calculated, per element, as 1 less the quotient of the number of true rank hops taken between the number of maximum

Case	Condition	Colour
1	$1 \geq \mathrm{RSI} > 0.99$	blue
2	$\mathrm{RSI} > 0.90$	green
3	$\mathrm{RSI} > 0.75$	orange
4	$\mathrm{RSI} > 0.25$	red
5	$0.25 \geq \mathrm{RSI} \geq 0$	**black**

Table 7. Colour code of the RSI percentage text shown in rank plots, following the first condition satisfied.

possible rank hops, all powered to p:

$$\mathrm{RSI} = \left(1 - \frac{\text{true rank hops}}{\text{possible rank hops}}\right)^p = \left(1 - \frac{D}{(N-1)(t-1)}\right)^p$$

where D is the total of rank hops taken by the studied element, N is the number of elements that have been ranked, and t is the number of time samples. The power index p is arbitrarily chosen to increase the resolution in the stable region; the value in the current version of the code is $p = 4$.

As an example of this "zooming" effect in the stable region, to match a linear ($p = 1$) RSI of 0.9 to a powered one of 0.1, we should select $p = 21.8543$. An alternative way to obtain this effect and exactly map a linear RSI of 0.9 to a non-linear RSI (RSI') of 0.1, is by applying the following function:

$$\mathrm{RSI}' = \frac{10^{10\left(1-\frac{D}{(N-1)(t-1)}\right)} - 1}{10^{10} - 1} \approx 10^{-10\left(\frac{D}{(N-1)(t-1)}\right)}$$

where the approximation is valid because $10^{10} \gg 1$ but, the small price to pay for it is that, in the worst instability case, the RSI' would not be strictly 0 but 10^{-10}.

The colour code of the RSI percentage text in the rank plot of *cmplxcruncher* is chosen following the first condition satisfied from those shown in Table 7 (see page 29).

cmplxcruncher output

In the previous sections, we have discussed details about the methods used in *cmplxcruncher*. In this section we review the output from the package, which aims at summarizing the as yet undescribed functionality.

Fit Plots

Both for the unweighted fit (detailed in the section) and for the X-weighted fit (detailed in the section), two plots are generated: the former with logarithmic scale, the latter with lineal scale. Figure 7 shows an example of the unweighted fit plots and Figure 8 does the same with the X-weighted fit plots. Additionally, for the unweighted fit, a complete residues analysis is performed, and a 4-in-1 figure is generated as shown in Figure 9, corresponding to the fit of Figure 7. Among other tests, it allows to check for normality and homoscedasticity of the residues.

Histogram Plots

cmplxcruncher generates three different histogram plots:

—**Absolute frequencies plot** : This plot is useful to visually assess the validity of the time points in terms of the accumulated absolute frequency of the elements (taxa), since absolute frequencies far (much higher or much lower) from those typically observed could mean a sampling problem. As an example, Figure 10 shows this histogram for the pre-treatment data (first 7 times) of patient “D” in the antibiotics study (*5*).

—**2D deviation plot** : The 2D semi-logarithmic histogram representing deviations from the mean versus the mean itself, is a useful tool in the analysis of the stability of ranking processes in complex systems (*15*). Figure 11 shows this plot for the data used in the fit shown in Figure 7.

—**Zero relative frequency plot** : We could define the ZRF (Zero Relative Frequency, thereby ranging from 0 to 1) of an element (taxon) as the portion of times where

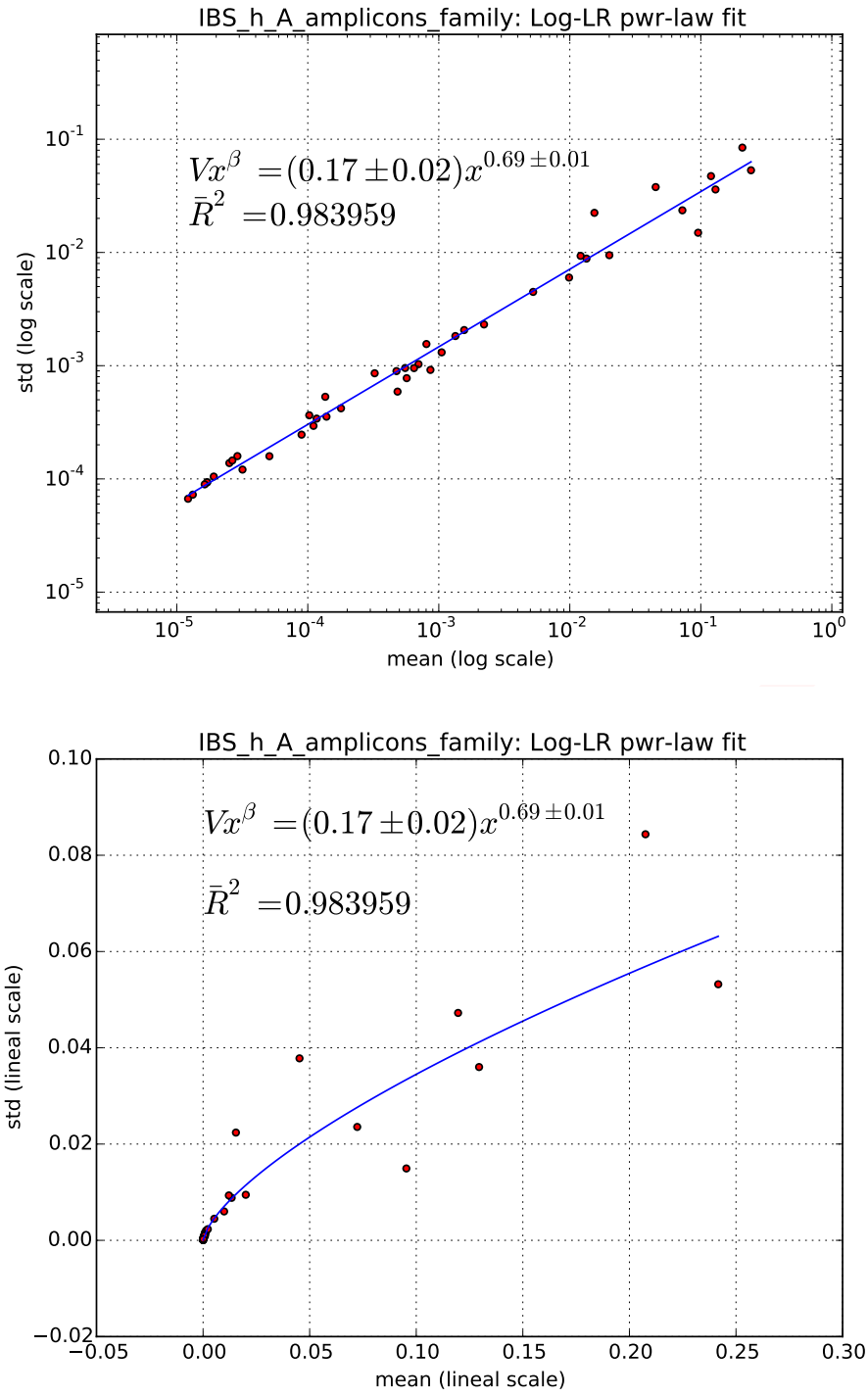

Figure 7. Example of unweighted fit, shown both in logarithmic and lineal scale plots

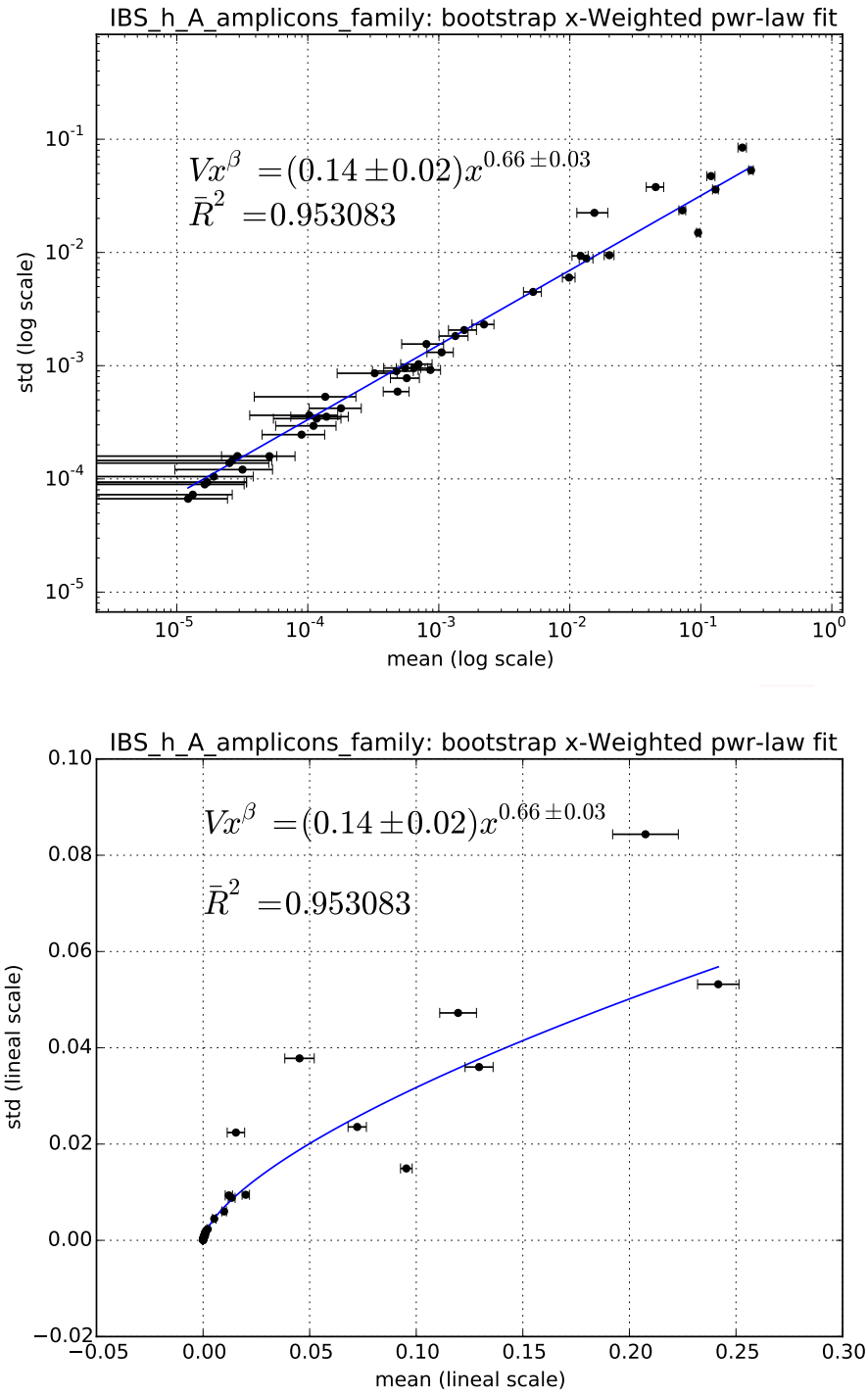

Figure 8. The X-weighted fit log and lineal plots corresponding to the fit of Figure 7

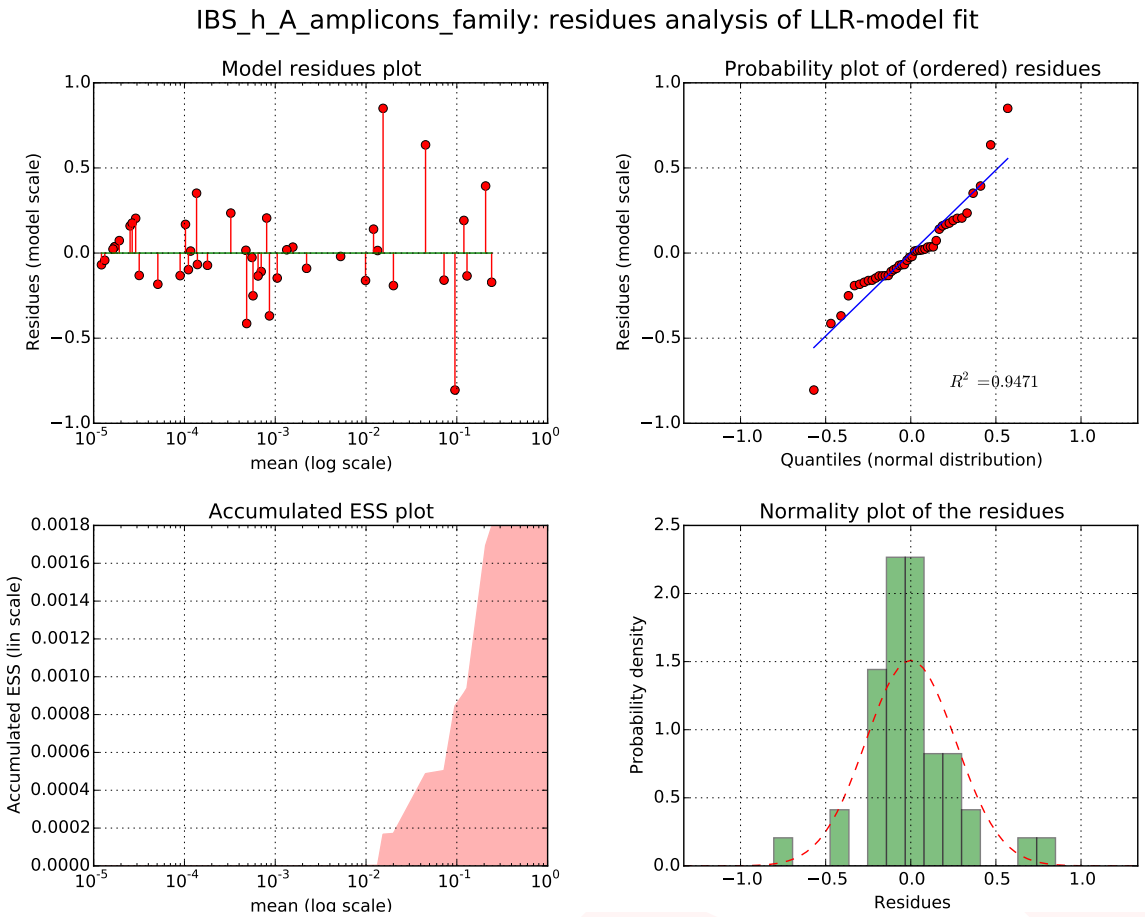

Figure 9. Residues analysis plot corresponding to the fit of Figure 7. The top-left subplot is a simple residues plot. The top-right subplot is a Normal quantiles plot with linear fitting (value of coefficient of determination is provided). The bottom-left subplot shows an accumulated ESS (Explained Sum of Squares) plot. Finally, the bottom-right subplot is a residues Normal histogram plot. This set of subplots allows to check for normality and homoscedasticity of the residues.

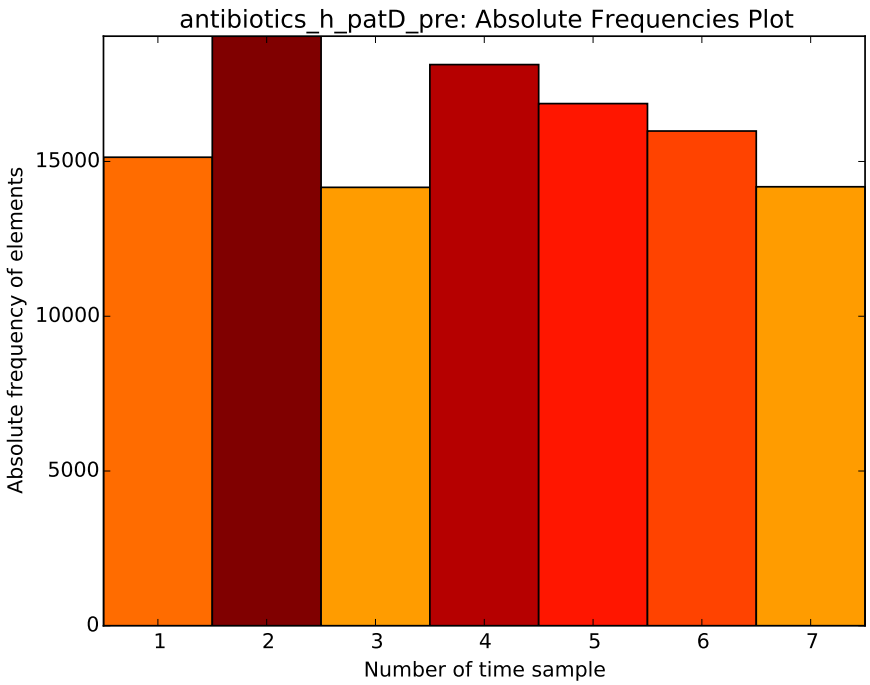

Figure 10. Histogram with the absolute frequencies of the pre-treatment data (7 first times) of patient "D" in the antibiotics study (*5*)

it is zero, i.e., it is not found. Attending to all the elements (taxa), we can plot the ZRF histogram, which then lies on the horizontal axis of the plot. The vertical axis shows the number of elements (taxa), so the height of a bar represents the number of elements that have determinate ZRF. In this respect, the bar over 0.0 counts the number of elements (taxa) that are present at every time point of the data set (aka "core"), while the bar over 1.0 would count the number of elements (taxa) that are never found (this bar never appears because all these "null" elements are automatically filtered by the code). Figure 12 shows an example of this plot. There, we can see that 12 taxa are present at all the time points of the time series while 9 taxa basically appear only once. So, this plot is clearly useful to notice how the "core" is distributed.

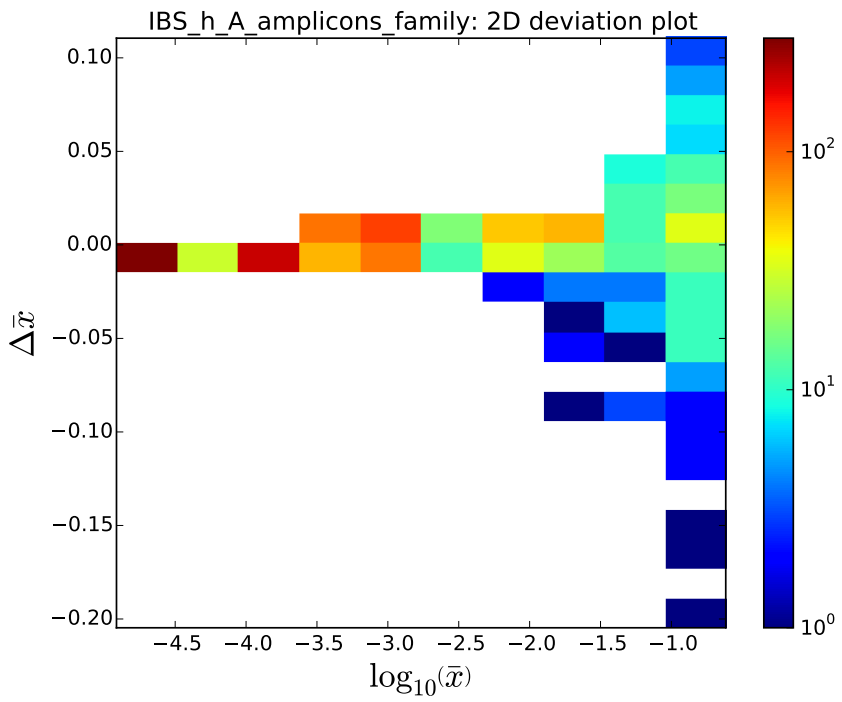

Figure 11. 2D histogram deviation plot of the data used in the fit shown in Figure 7

Correlation and Rank Plots

cmplxcruncher generates three different plots falling under this category, as well as Excel files with the resulting matrices:

—**Elements correlation matrix** : This plot shows a correlation matrix among the elements (taxa), calculated with the time as independent variable. For these calculations, the data set is not normalized to avoid entering an additional constraint. As an example, Figure 13 shows this matrix for the "core" elements (taxa) present in the pre-treatment data (first seven times) of patient "D" in the antibiotics study (*5*).

—**Times correlation matrix** : This plot presents a correlation matrix among the time points of the data set, calculated with the elements (taxa) as independent variable. Again, the data set is not normalized. Figure 14 shows this matrix for the "core" elements (taxa) present in pre-treatment data (first seven times) of patient "D" in the antibiotics study (*5*).

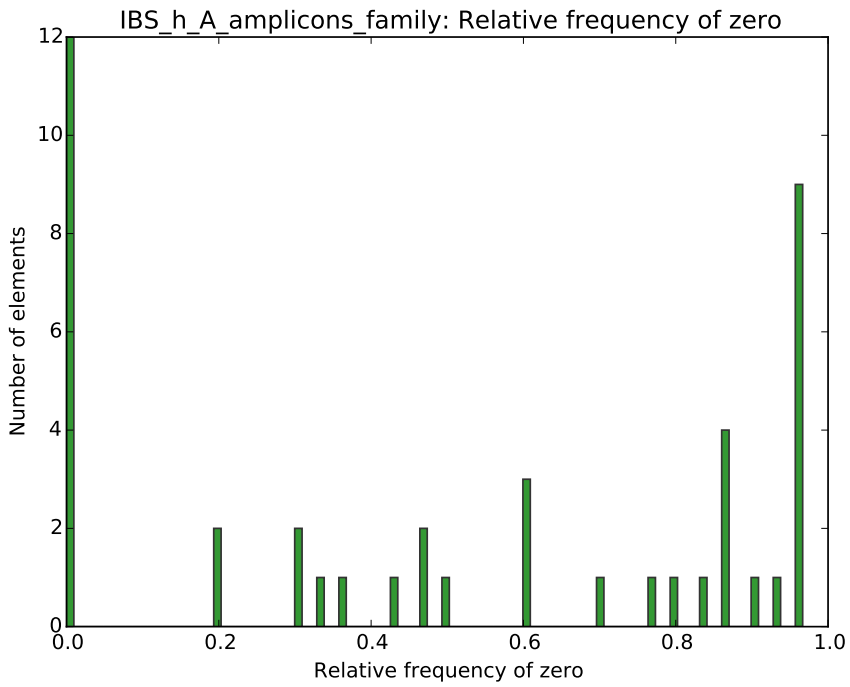

Figure 12. Histogram with the relative frequency of zero for the elements (taxa) in the data used in the fit shown in Figure 7

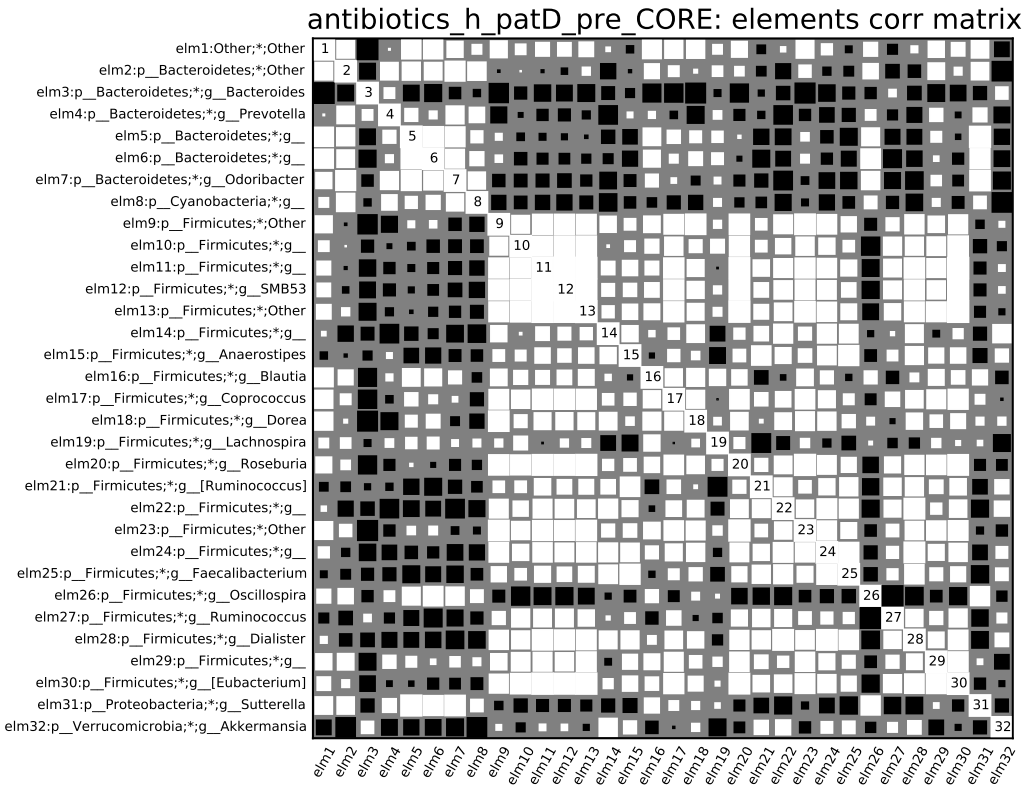

Figure 13. Element correlation plot of the pre-treatment data (first seven times) of patient "D" in the antibiotics study (*5*) for "core" taxa

—**Rank dynamics and stability plot** : This plot shows the variation in the rank with time for the most dominant elements (taxa) and their calculated RSI, as discussed in Section . Figure 15 shows this plot for the elements (taxa) in the data used in the fit shown in Figure 7.

Standardization

In order to show all the studies properly under common axes, we decided to standardize the Taylor parameters using the group of healthy individuals for each study. With this approach, all the studies can be visualized in a shared plot with units of Taylor-parameters standard-deviation on their axes.

For a Taylor parameter, e.g. V, the estimate of the mean ($\widehat{V}$) for the healthy subpop-

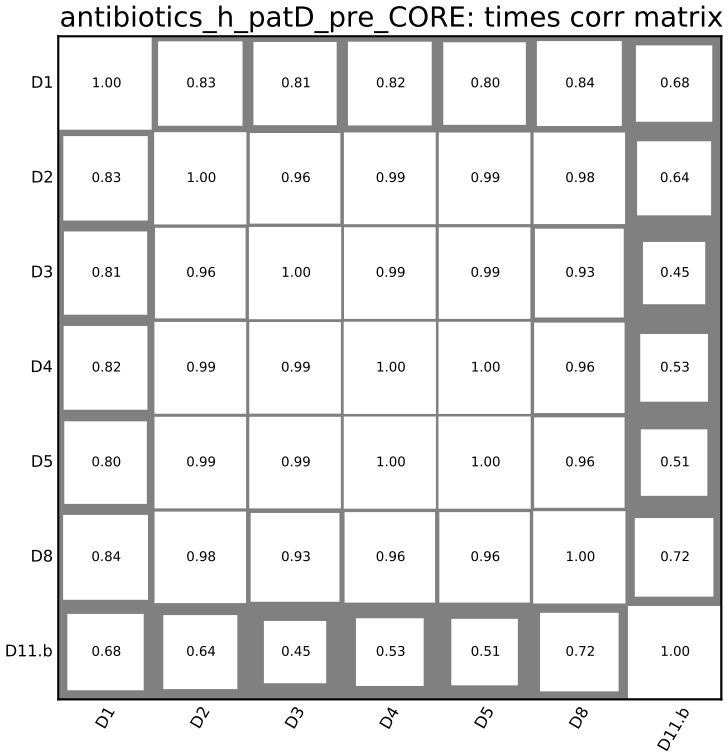

Figure 14. Times correlation plot of the pre-treatment data (first seven times) of patient "D" in the antibiotics study (*5*) for "core" taxa

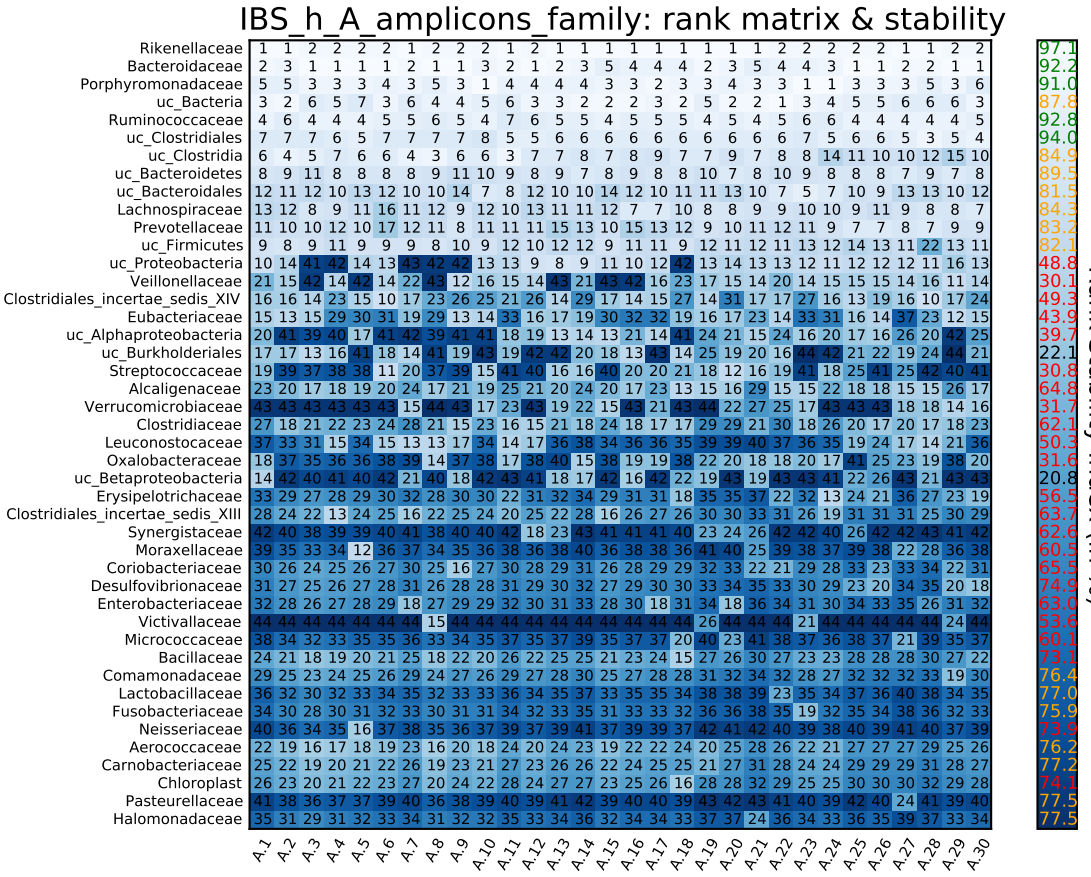

Figure 15. Matrix showing the rank variation throughout time for the most dominant elements (taxa) and their calculated RSI (as discussed in Section) in the data used in the fit shown in Figure 7

ulation, composed of h individuals, is:

$$\widehat{V} = \frac{1}{W_1}\sum_{i=1}^{h} V_i\omega_i = \sum_{i=1}^{h} V_i\omega_i$$

as $W_1 = \sum_i^h \omega_i = 1$, since ω_i are normalized weights calculated as:

$$\omega_i = \frac{\frac{1}{\sigma^2_{V_i}}}{\sum_i^h \frac{1}{\sigma^2_{V_i}}}$$

being σ_{V_i} the estimation of the uncertainty in V_i obtained together with V_i from the X-weighted power-law fit described in Section , for healthy individuals.

Likewise, the estimation of the standard deviation for the healthy population ($\widehat{\sigma}_V$) is:

$$\widehat{\sigma}_V = \sqrt{\frac{1}{W_1 - \frac{W_2}{W_1}}\sum_{i=1}^{h}\left[\omega_i\left(V_i - \hat{V}\right)^2\right]}$$

being $W_2 = \sum_i^h \omega_i^2$, which finally yields to:

$$\widehat{\sigma}_V = \sqrt{\frac{1}{1 - \sum_i^h \omega_i^2}\sum_{i=1}^{h}\left[\omega_i\left(V_i - \hat{V}\right)^2\right]}$$

Acknowledgments

Authors declare that there are no competing financial interests in relation to the work described here.

Funding Information

This work is supported by Generalitat Valenciana Prometeo Grants II/2014/050, II/2014/065, by the Spanish Grants FPA2011-29678, BFU2012-39816-C02-01 of MINECO and by PITN-GA-2011-289442-INVISIBLES. JMM & DMM acknowledge FPI and FISABIO fellowships.

514 Modificar becas de JMM y DMM ¿poner algún grant más?

References

1. Caporaso, J.G. et al. Moving pictures of the human microbiome. *Genome Biol.* **12,** R50 (2011).
2. Faith, J.J. et al. The long-term stability of the human gut microbiota. *Science* **341,** 1237439 (2013).
3. Smith M.I. et al. Gut microbiomes of Malawian twin pairs discordant for kwashiorkor. *Science* **339,** 548-54 (2013).
4. David, L.A. et al. Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505,** 559-63 (2014).
5. Dethlefsen L., Relman D. A. Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proc. Nat. Acad. Sci. USA* **108,** 4554-61 (2011).
6. Durban, A. *et al.* Structural alterations of faecal and mucosa-associated bacterial communities in irritable bowel syndrome. *FEMS Microbiol. Ecol.* **86**, 581-9 (2013).
7. Caporaso, J.G. et al. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7,** 335-6 (2010).
8. Ames, S.K. et al. Scalable metagenomic taxonomy classification usng a reference genome database. *Bioinformatics* **29,** 2253-2260 (2013).
9. Eisler,Z., Bartos,I., Kertesz,J. Fluctuation scaling in complex systems: Taylor's law and beyond. *Adv. Phys.* **57,** 85 (2008).
10. Taylor, L.R. Aggregation, Variance and the mean. *Nature* **189,** 732-35 (1961).
11. Jorgensen,B., Martinez,J.R., Tsao,M. Asymptotic behaviour of the variance function. *Scand. J. Statist.*. **21,** 223-243 (1994).
12. Fronczak,A., Fronczak,P. Origins of Taylor's power law for fluctuation scaling in complex systems. *Phys. Rev. E* **81,** 066112 (2010).
13. Kendal, W.S., Jorgensen,B. Taylor's power law and fluctuation scaling explained by a central-limit-like convergence. *Phys. Rev. E* **83,** 066115 (2011).

14. Kendal, W.S., Jorgensen,B. Tweedie convergence: A mathematical basis for Taylor's power law. *Phys. Rev. E* **84,** 066120 (2011).
15. Blumm, N. et al. Dynamics of ranking processes in complex systems. *Phys. Rev. Lett.* **109,** 128701 (2012).
16. Weber, J. *et al.* Fluctuation dissipation theorem. *Phys. Rev.* **101**, 1620-6 (1956).
17. Gordon, A., Hannon, G.J. FASTX-Toolkit. FASTQ/A shortreads pre-processing tools (2010). http://hannonlab.cshl.edu/fastx_toolkit/ (accessed 23 Feb 2015).
18. Quast C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools (2013)
19. Xiao Xiao, Ethan P. White, Mevin B. Hooten, and Susan L. Durham. On the use of log-transformation vs. nonlinear regression for analyzing biological power laws. *Ecology* **92**, 10, 1887-1894 (2011).
20. Magee L., R^2 measures based on wald and likelihood ratio joint significance tests. *The American Statistician* **44**, 3, 250-253 (1990).
21. Nagelkerke N.J.D., A note on a general definition of the coefficient of determination. *Biometrika* **78**, 3, 691-692 (1991).
22. Wu, C.F.J. Jackknife, bootstrap and other resampling methods in regression analysis. (with discussions) *The Annals of Statistics* **14**: 1261?1350 (1986)

Eliminar *et al.* y poner la referencia completa como exige la guía de estilo de la revista...