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 Martínez-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, UVEG, 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: 5332 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 modeling, metagenomics, 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. We are populated by a myriad of microorganisms that are interacting with us in several physiological processes such as metabolism regulation or maturation of the immune system. Human microbiota has been suggested to be closely related to diseases like type 2 diabetes (*1*), cardiovascular disease (CVD) (*2*), irritable bowel syndrome (*3*), Crohn's disease (*4*) or some affections as obesity (*5*, *6*) or malnutrition (*7*). 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. 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 [ref?].

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 [ref] it has become clear that applications of system-wide approaches to Human Biology would bring huge opportunities in Medicine. Great effort has been placed to unveil the general laws governing the behaviour of this complex system [ref]. Due to his nature, microbiota can be studied under the light of the ecology, where we can find general principles as the Taylor?s law, which relates spatial or temporal variability of the population with its mean. This law, also known as fluctuation scale law, is ubiquitous in the natural world and can be found in several systems as cosmic rays [ref1], stock markets [ref2,3], animal populations [ref4, 5, 6], gene expression [ref7], or in the human genome [ref8]. Taylor?s law has been applied to microbiota in a spatial way in the work of Zhang et al., (2014), where they show that this population tend to be in an aggregated way rather than in a random distribution.

Here we present the imprints of disease in macroscopic properties of the system, by studying the temporal variability in the microbiome. We have analyzed more than 35000 time series of taxa from the gut microbiome of 97 individuals obtained from publicly available high throughput sequencing data on different conditions: diseases, diets, obese status, antibiotic perturbation and healthy individuals. Having seen that all cases follows Taylor?s law, we use this empirical fact to model how the relative abundances of taxa evolves toward time thanks to the Langevin equation, in a similar way as Blumm et al., did in their (2012). We use this mathematical framework to explore the temporal stability of the microbiota in different conditions in order to understand how this affects the healthy status of the subjects. Finally, we have engineered a complete software framework, ComplexCruncher, to support the analysis of the dynamics of ranking processes in complex systems, which is ready to be implemented by other users.

Results

Global 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 ubiquitous pattern known as the fluctuation scaling law (*15*) or Taylor's power law (*16*), i.e., microbiota of all detected taxa follows $\sigma_i = V \cdot x_i^{\beta}$, a power law dependence between mean relative abundance x_i and dispersion σ_i. The law seem to be ubiquitous, spanning even to six orders of magnitude in the observed relative abundances (see Figure 1).

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 related with the health status of the host, which we consider the main finding exposed in this article (see Figure 2).

Taylor parameters describing the temporal variability of the gut microbiome in our sampled individuals are shown in Tables 1 to 6. Our results hint at an ubiquitous behaviour. On the first hand, the variability (which corresponds to the maximum amplitude of fluctuations) is large, which suggests resilient capacity of the microbiota. On the other hand, the scaling index is always smaller than one, which means that more abundant taxa are less volatile than less abundant ones. In addition, Taylor parameters for the microbiome of healthy individuals in different studies are compatible within estimated errors. This enables us to define an area in the Taylor parameter space that we called the *healthy zone*.

In order to jointly visualize and compare the results of individuals from different studies, their Taylor parameters have been standardized, where standardization means that each parameter is subtracted by the mean value and divided by the standard deviation of the group of healthy individuals for each study (for details of the procedure, please see Standardization subsection in Material and Methods). The healthy zone and the standardized Taylor parameters for individuals whose gut microbiota is threatened (i.e., suffering from kwashiorkor, altered diet, antibiotics or IBS) is shown in Figure 2. 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. Furthermore, we have verified that our conclusions are robust to systematic errors due to taxonomic assignment.

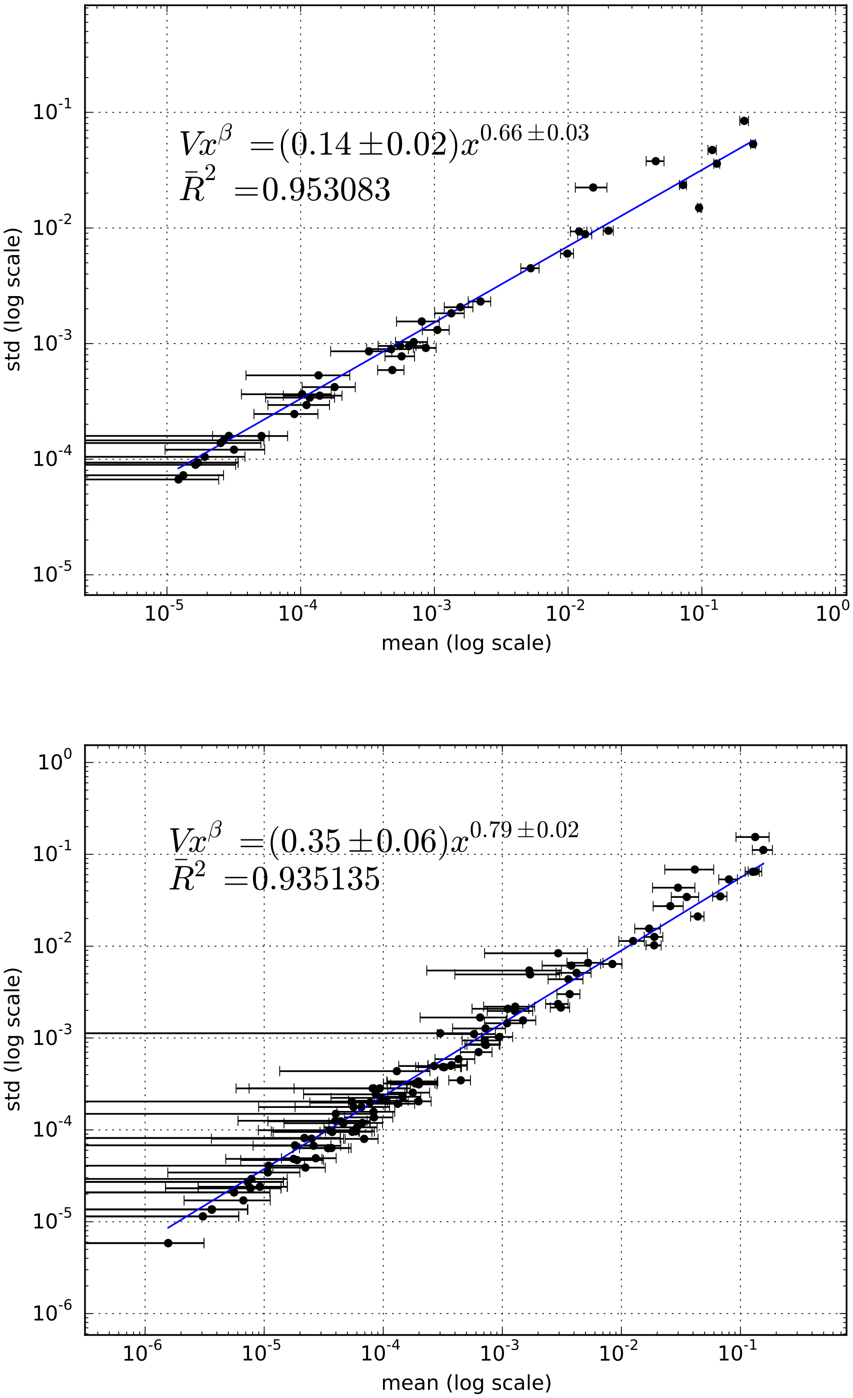


Figure 1. X-weighted power-law fits of the standard deviations versus the mean values for each bacterial genus monitored in time. We show the fit for samples from a healthy subject (top) and from a subject diagnosed with irritable bowel syndrome (bottom), studied in our lab (*3*). Taylor's power law seems to be ubiquitous, spanning to six orders of magnitude.

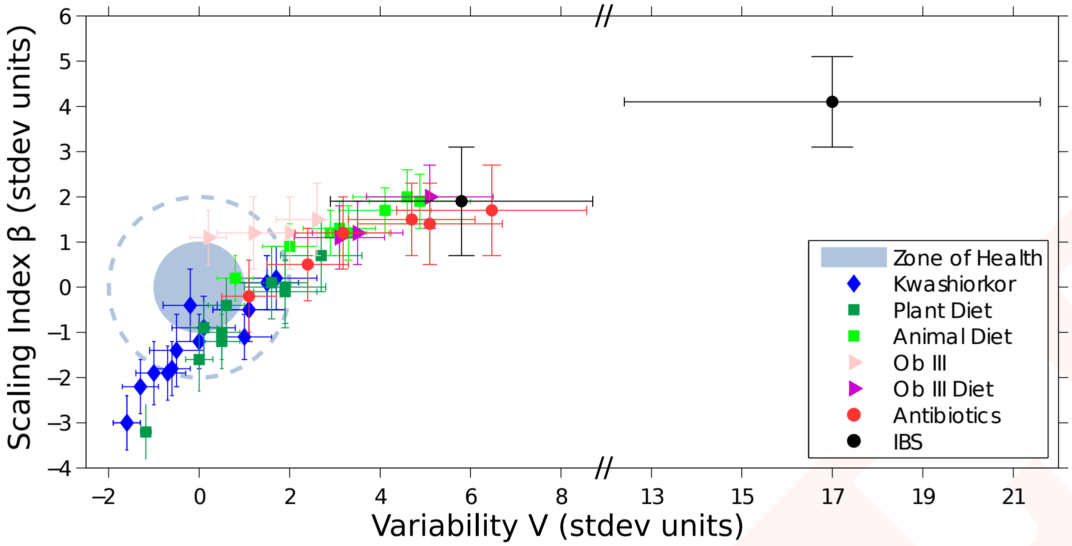


Figure 2. Taylor's law parameter space. We have compiled here all the data studied in this work. The coloured circle corresponds to 68% confidence level (CL) region of healthy individuals in the Taylor parameter space, while dashed line delimites the 98% CL region. Points with errors place each individual gut microbiome in the Taylor space. Note that the parameters have been standardized (stdev units) to the healthy group in each study for demonstrative and comparative purposes.

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 for individuals with either animal-based (A) or plant-based (P) diets (*11*). Previous to diet, the population sampled is described by $\bar{V} = 0.09 \pm 0.05, \bar{\beta} = 0.77 \pm 0.04$, which we used to describe the *healthy zone* for this study.

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 (*12*). Prior to antibiotics intake, the population sampled is described by $\bar{V} = 0.12 \pm 0.05, \bar{\beta} = 0.75 \pm 0.04$, which characterize the *healthy zone* for this study.

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) (*3*). Healthy individuals sampled in this study are characterized by $\bar{V} = 0.134 \pm 0.009$, $\bar{\beta} = 0.691 \pm 0.025$, which we used to define the correspondent *healthy zone*.

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 (*10*). This table continues in Table 5. The population of healthy twins is characterized by $\bar{V} = 0.25 \pm 0.10$, $\bar{\beta} = 0.863 \pm 0.028$, values which we used to describe the *healthy zone* for this study.

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 (*10*). This is a continuation of Table 4, so that the population of healthy twins is also characterized by $\bar{V} = 0.25 \pm 0.10$ and $\bar{\beta} = 0.863 \pm 0.028$.

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 (*17*), providing a mechanistic explanation based on the statistical theory of errors (*18–20*). 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, on the one hand, a deterministic term that captures the fitness of each taxon and, on the other hand, a randomness term associated with Gaussian random noise (*21*). Both terms are modeled by power laws, with coefficients that can be interpreted as the taxon fitness F_i and the variability V (see Model under Material and Methods). In this model, 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

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 (*9*). Healthy people in this study, whom were not obese, are characterized by $\bar{V} = 0.19 \pm 0.06$, $\bar{\beta} = 0.806 \pm 0.034$, which we used to determine the correspondent *healthy zone* for this study.

best captured by a phase space determined by fitness F and amplitude of fluctuations V (see Figure 3).

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.

Specific results

Fit Plots

For each and every dataset included in the study, an unweighted fit (see Material and Methods section for details) and a X-weighted fit (detailed in Material and Methods subsection) have been calculated for standard deviations versus the mean values for each bacterial genus monitored in time. Figure 1 showed the X-weighted fit for samples from a healthy subject (patient A, top) and from a subject diagnosed with irritable bowel syndrome (patient P2, bottom) studied in our lab (*3*), while Figure 4 shows the corresponding unweighted fits. Additionally, for the unweighted fit, a complete residues analysis was performed, and a 4-in-1 figure was generated as shown in Figure 5, corresponding to patient A (top plot in Figure 1 and Figure 4). Among other tests, it allows to check for normality and homoscedasticity of

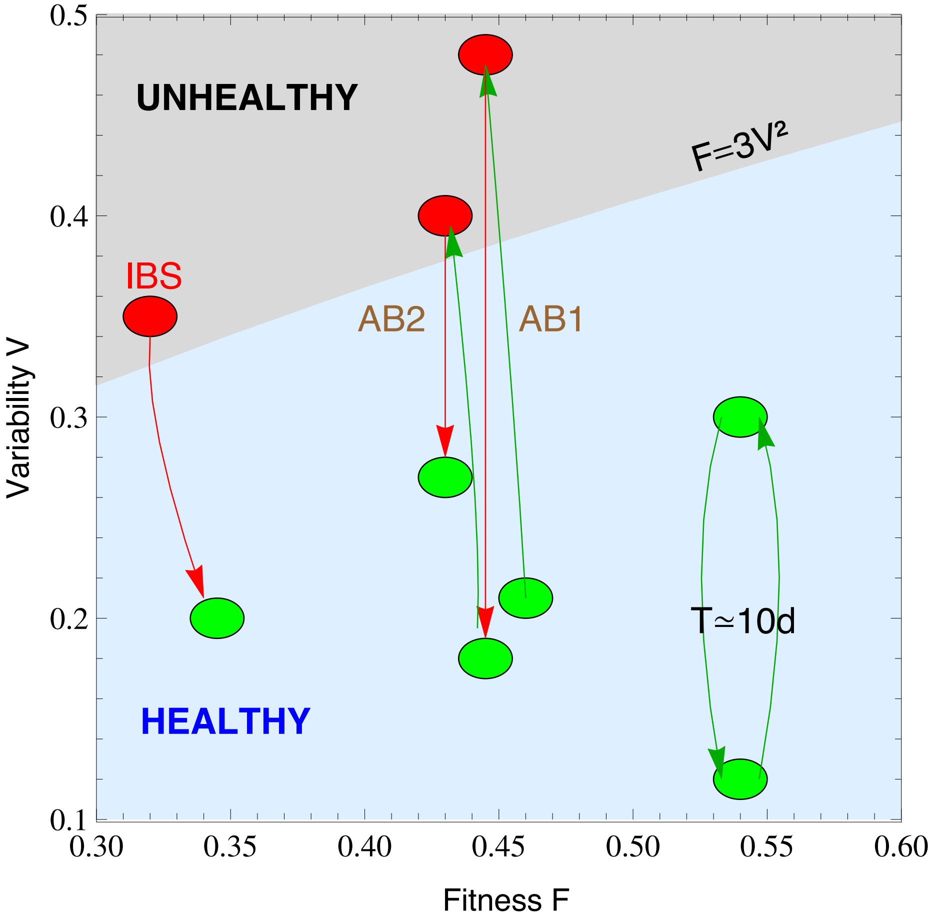


Figure 3. Microbiota states can be placed in the phase space F–V. The light blue shaded region corresponds to the stable phase, while the grey shaded region is the unstable phase (the phase transition line is calculated for $\alpha = \beta = 0.75$). We place healthy individuals (green) and individuals 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.

the residues and, therefore, to further evaluate the goodness of the fit.

Histogram Plots

cmplxcruncher generates two related histogram plots, absolute frequencies plot and zero relative frequency plot. The former 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. In Figure 6 shows this histogram for the pre-treatment data (first 7 times) of patient "D" in the antibiotics study (*12*). On the other hand, we could define the ZRF (Zero Relative Frequency, thereby ranging from 0 to 1) of an element (taxon) as the portion of times where it is zero, i.e., it is not found. Attending to all taxa, we can plot the ZRF histogram, which then lies on the horizontal axis of the plot. The vertical axis shows the number of taxa, so the height of a bar represents the amount of taxa that have determinate ZRF. In this respect, the bar over 0.0 counts the quantity of taxa that are present at every time point of the data set (aka "core"), while the bar over 1.0 would count the total of taxa that are never found (this bar never appears because all these "null" elements are automatically filtered by the software. Figure 7 shows this plot for the healthy patient A of the IBS study performed in our lab (*3*). 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.

A 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 (*21*). Figure 8 shows this 2D deviation plot for the patient A of the IBS study (*3*).

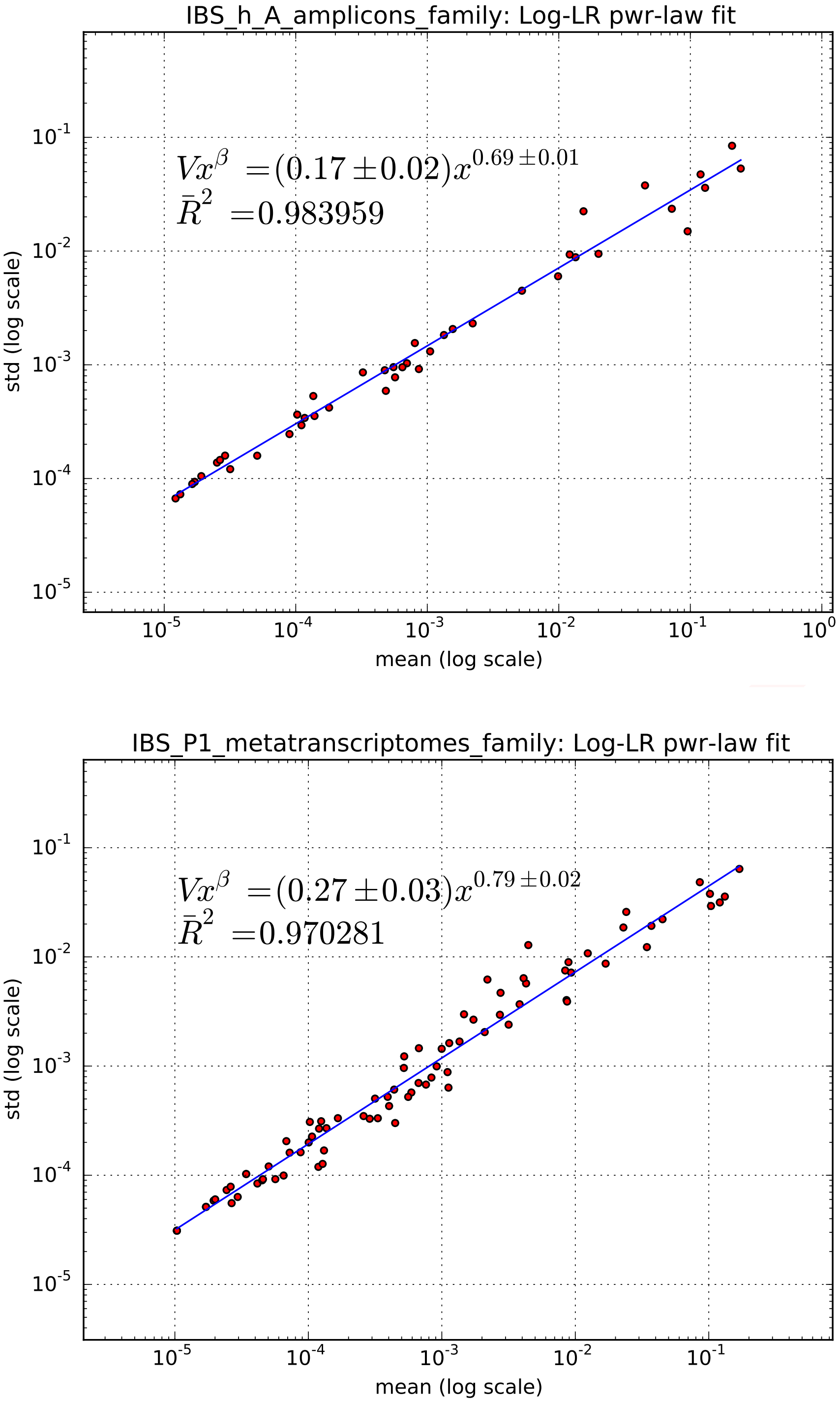


Figure 4. Log plots of unweighted fits corresponding to the datasets shown in Figure 1

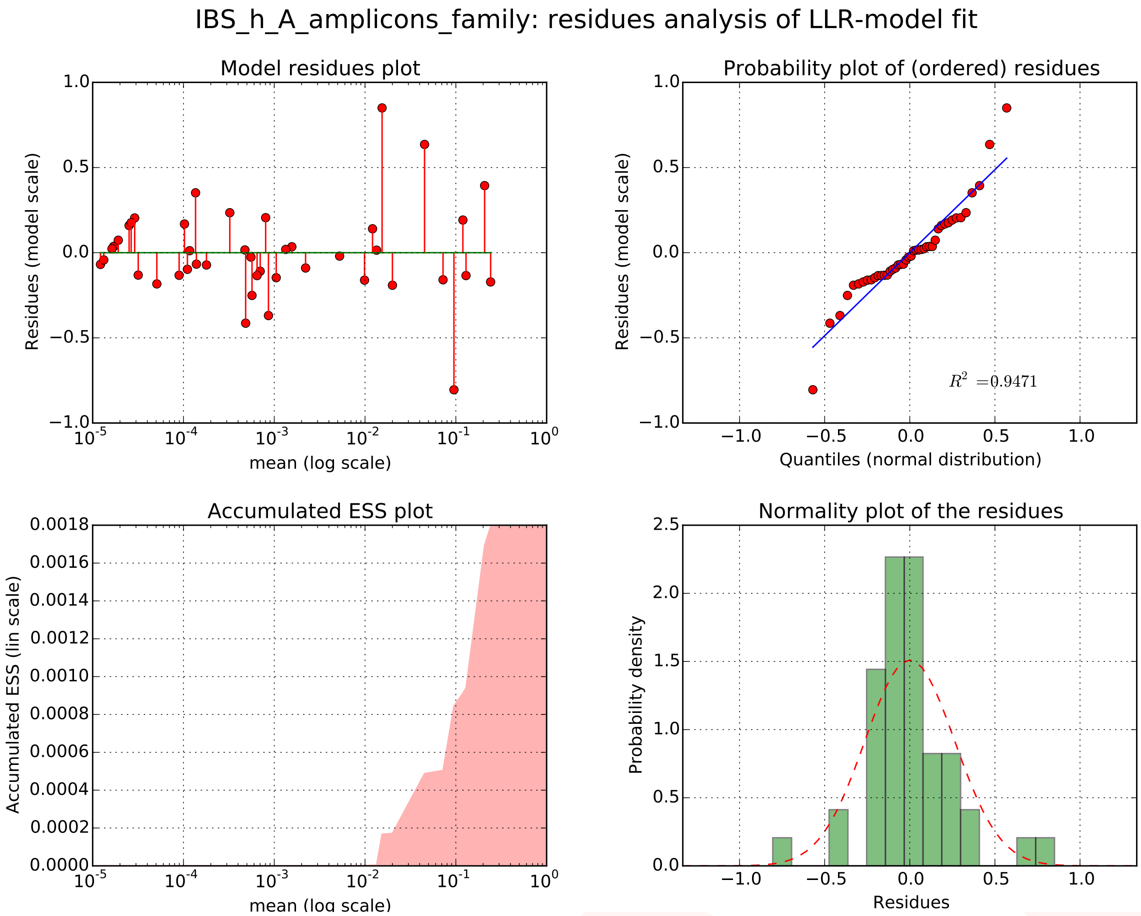


Figure 5. Residues analysis plot corresponding to the unweighted fit for patient A of the IBS study (*3*). The top-left subplot is a simple residues plot. The top-right subplot is a Normal quantiles plot with linear fitting (value of coefficient of determination is provided). The bottom-left subplot 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.

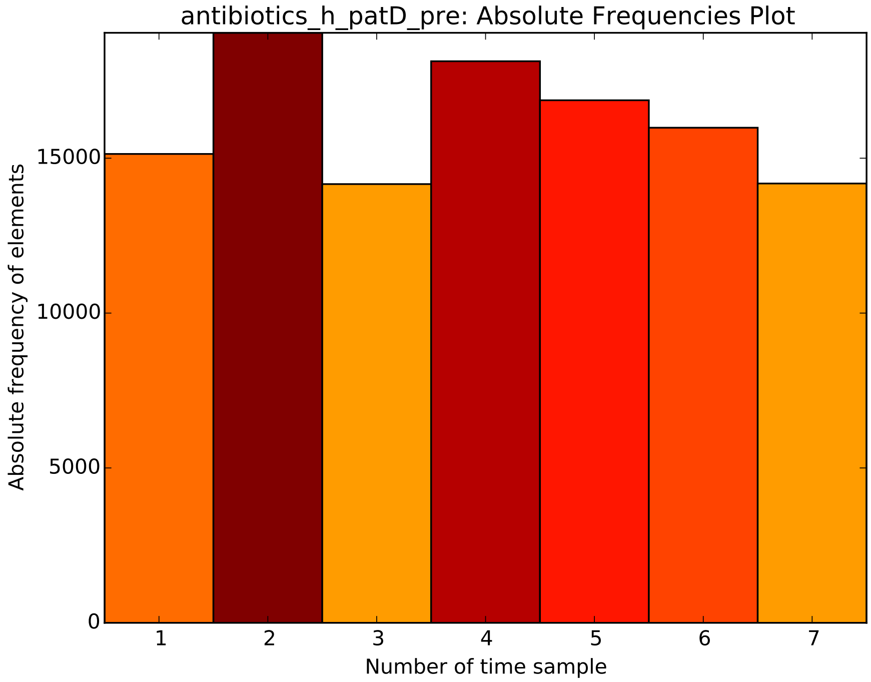


Figure 6. Histogram with the absolute frequencies of the pre-treatment data (7 first times) of patient "D" in the antibiotics study (*12*). We can see that there are no *outlayers* among the total taxa sum for the time series points, so all of them were considered for further analysis by *cmplxcruncher* software

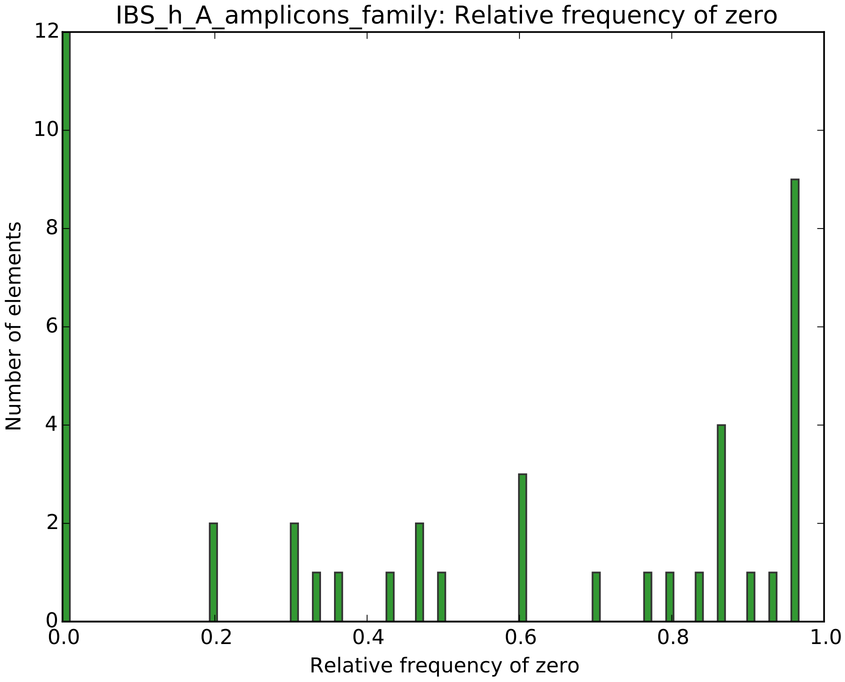


Figure 7. Histogram with the relative frequency of zero for the elements (taxa) in the data for patient A of the IBS study (*3*). In this particular case, the bar over 0.0 counts the quantity of taxa that are present at every time point of the data set ("core"), while the bar over 1.0 sums the total of taxa that are never found (this bar never appears because all these "null" elements are automatically filtered by *cmplxcruncher*): here, 12 taxa are present at all the time points of the time series while 9 taxa basically appear only once.

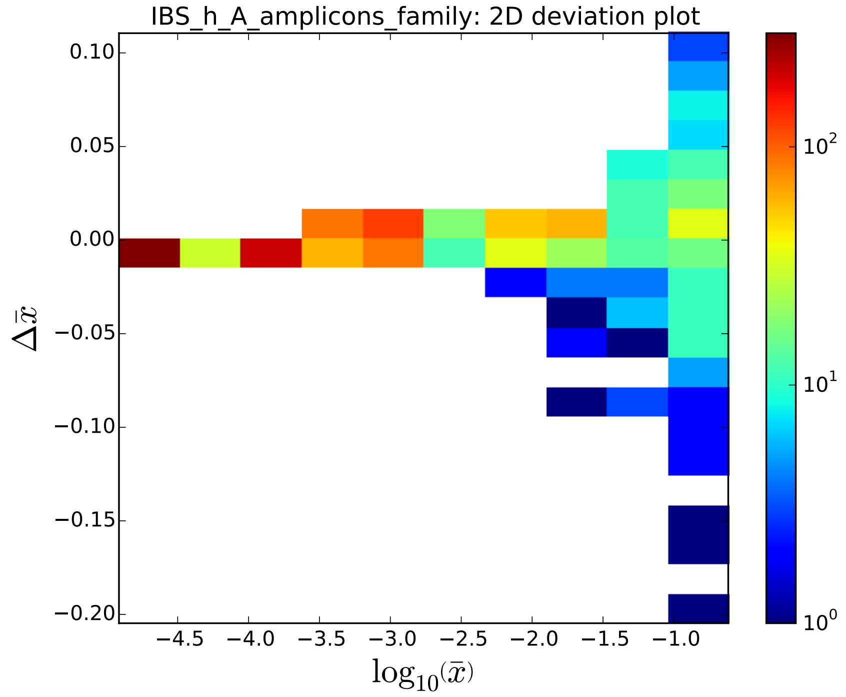
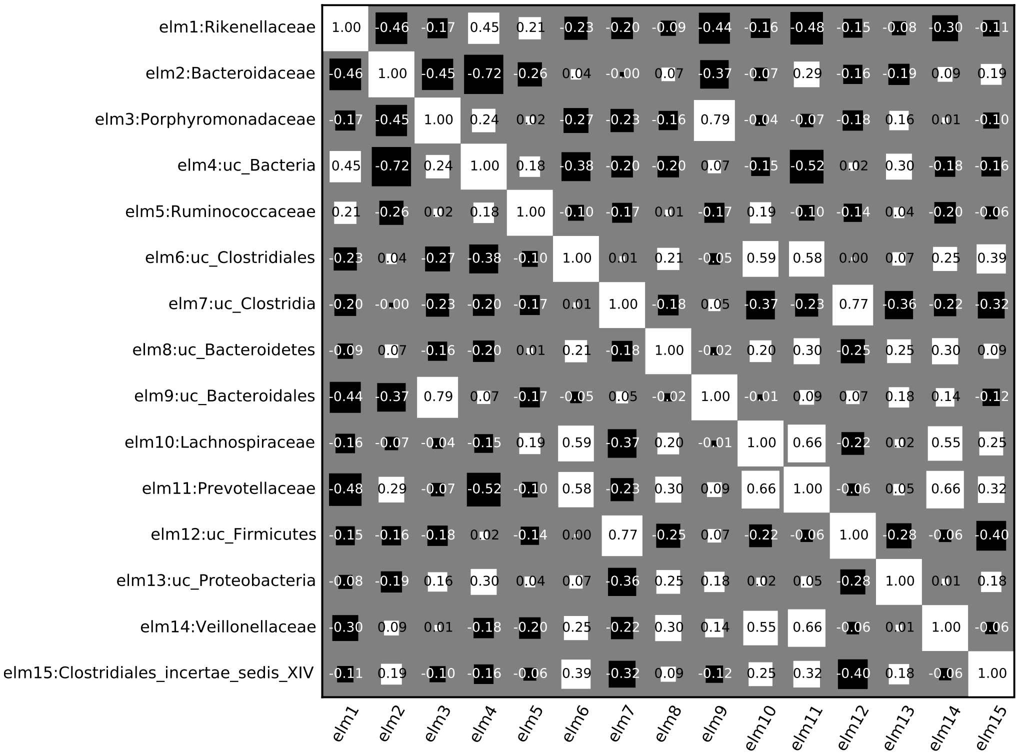


Figure 8. 2D histogram deviation plot of the data for patient A of the IBS study (*3*)

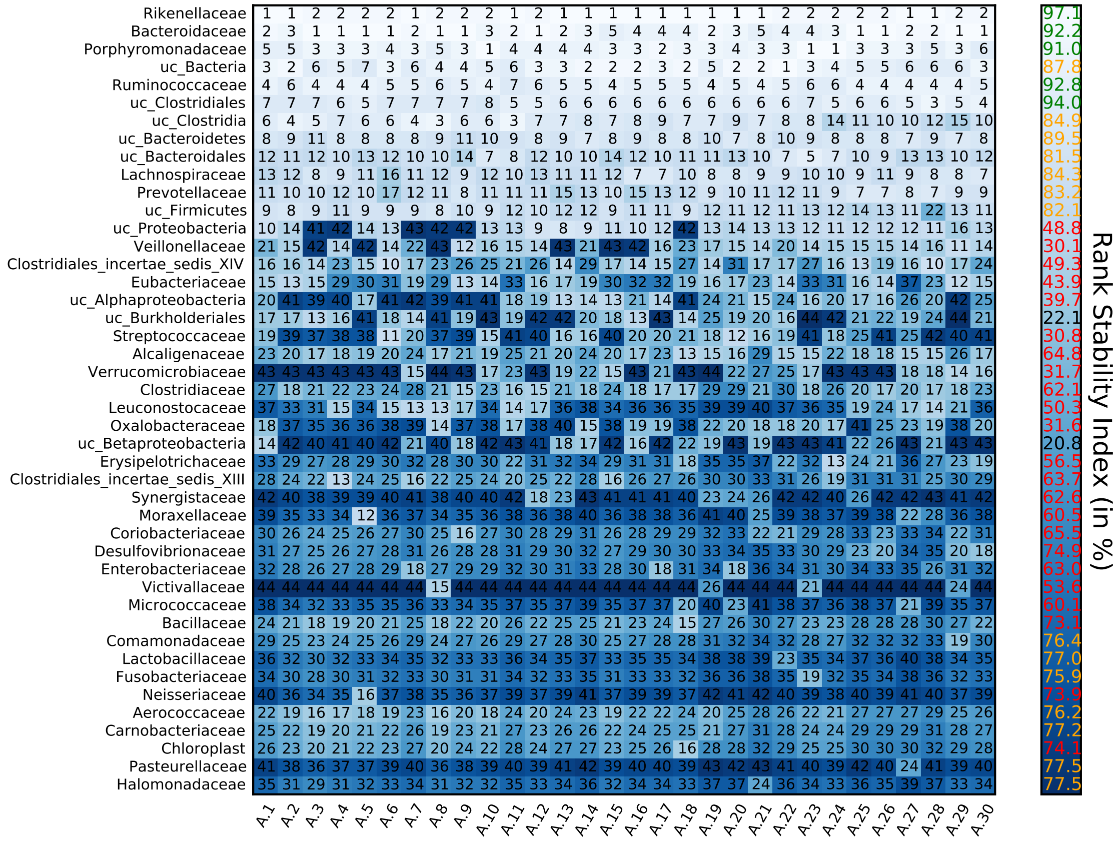
Correlation and Rank Plots

cmplxcruncher generates two different plots falling under this category, as well as Excel files with the resulting matrices. On the one hand, the elements correlation matrix plot shows a correlation matrix among the taxa, calculated with the time as independent variable. For these calculations, the data set is not normalized to avoid entering an additional constraint. Figure 9 shows this matrix for the most dominant taxa present in the data of the patient “A” of the IBS study (*3*). On the other hand, the rank dynamics and stability plot shows the variation in the rank with time for the most dominant taxa and their calculated RSI, as discussed in Material and Methods. Figure 10 shows this plot for the taxa of the aforementioned patient A.



	elm1	elm2	elm3	elm4	elm5	elm6	elm7	elm8	elm9	elm10	elm11	elm12	elm13	elm14	elm15
elm1:Rikenellaceae	1.00	-0.46	-0.17	0.45	0.21	-0.23	-0.20	-0.09	-0.44	-0.16	-0.48	-0.15	-0.08	-0.30	-0.11
elm2:Bacteroidaceae	-0.46	1.00	-0.45	-0.72	-0.26	0.04	-0.00	0.07	-0.37	-0.07	0.29	-0.16	-0.19	0.09	0.19
elm3:Porphyromonadaceae	-0.17	-0.45	1.00	0.24	0.02	-0.27	-0.23	-0.16	0.79	-0.04	-0.07	-0.18	0.16	0.01	-0.10
elm4:uc_Bacteria	0.45	-0.72	0.24	1.00	0.18	-0.38	-0.20	-0.20	0.07	-0.15	-0.52	0.02	0.30	-0.18	-0.16
elm5:Ruminococcaceae	0.21	-0.26	0.02	0.18	1.00	-0.10	-0.17	0.01	-0.17	0.19	-0.10	-0.14	0.04	-0.20	-0.06
elm6:uc_Clostridiales	-0.23	0.04	-0.27	-0.38	-0.10	1.00	0.01	0.21	-0.05	0.59	0.58	0.00	0.07	0.25	0.39
elm7:uc_Clostridia	-0.20	-0.00	-0.23	-0.20	-0.17	0.01	1.00	-0.18	0.05	-0.37	-0.23	0.77	-0.36	-0.22	-0.32
elm8:uc_Bacteroidetes	-0.09	0.07	-0.16	-0.20	0.01	0.21	-0.18	1.00	-0.02	0.20	0.30	-0.25	0.25	0.30	0.09
elm9:uc_Bacteroidales	-0.44	-0.37	0.79	0.07	-0.17	-0.05	0.05	-0.02	1.00	-0.01	0.09	0.07	0.18	0.14	-0.12
elm10:Lachnospiraceae	-0.16	-0.07	-0.04	-0.15	0.19	0.59	-0.37	0.20	-0.01	1.00	0.66	-0.22	0.02	0.55	0.25
elm11:Prevotellaceae	-0.48	0.29	-0.07	-0.52	-0.10	0.58	-0.23	0.30	0.09	0.66	1.00	-0.06	0.05	0.66	0.32
elm12:uc_Firmicutes	-0.15	-0.16	-0.18	0.02	-0.14	0.00	0.77	-0.25	0.07	-0.22	-0.06	1.00	-0.28	-0.06	-0.40
elm13:uc_Proteobacteria	-0.08	-0.19	0.16	0.30	0.04	0.07	-0.36	0.25	0.18	0.02	0.05	-0.28	1.00	0.01	0.18
elm14:Veillonellaceae	-0.30	0.09	0.01	-0.18	-0.20	0.25	-0.22	0.30	0.14	0.55	0.66	-0.06	0.01	1.00	-0.06
elm15:Clostridiales_incertae_sedis_XIV	-0.11	0.19	-0.10	-0.16	-0.06	0.39	-0.32	0.09	-0.12	0.25	0.32	-0.40	0.18	-0.06	1.00

Figure 9. Element correlation plot of the for the most dominant taxa in the data for patient A of the IBS study (*3*)



	A.1	A.2	A.3	A.4	A.5	A.6	A.7	A.8	A.9	A.10	A.11	A.12	A.13	A.14	A.15	A.16	A.17	A.18	A.19	A.20	A.21	A.22	A.23	A.24	A.25	A.26	A.27	A.28	A.29	A.30	Rank Stability Index (in %)
Rikenellaceae	1	1	2	2	2	2	1	2	2	2	1	2	1	1	1	1	1	1	1	1	1	2	2	2	2	2	1	1	2	2	97.1
Bacteroidaceae	2	3	1	1	1	1	2	1	1	3	2	1	2	3	5	4	4	4	2	3	5	4	4	3	1	1	2	2	1	1	92.2
Porphyromonadaceae	5	5	3	3	3	4	3	5	3	1	4	4	4	4	3	3	2	3	3	4	3	3	1	1	3	3	3	5	3	6	91.0
uc_Bacteria	3	2	6	5	7	3	6	4	4	5	6	3	3	2	2	2	3	2	5	2	2	1	3	4	5	5	6	6	6	3	87.8
Ruminococcaceae	4	6	4	4	4	5	5	6	5	4	7	6	5	5	4	5	5	5	4	5	4	5	6	6	4	4	4	4	4	5	92.8
uc_Clostridiales	7	7	7	6	5	7	7	7	7	8	5	5	6	6	6	6	6	6	6	6	6	6	7	5	6	6	5	3	5	4	94.0
uc_Clostridia	6	4	5	7	6	6	4	3	6	6	3	7	7	8	7	8	9	7	7	9	7	8	8	14	11	10	10	12	15	10	84.9
uc_Bacteroidetes	8	9	11	8	8	8	8	9	11	10	9	8	9	7	8	9	8	8	10	7	8	10	9	8	8	8	7	9	7	8	89.5
uc_Bacteroidales	12	11	12	10	13	12	10	10	14	7	8	12	10	10	14	12	10	11	11	13	10	7	5	7	10	9	13	13	10	12	81.5
Lachnospiraceae	13	12	8	9	11	16	11	12	9	12	10	13	11	11	12	7	7	10	8	8	9	9	10	10	9	11	9	8	8	7	84.3
Prevotellaceae	11	10	10	12	10	17	12	11	8	11	11	11	15	13	10	15	13	12	9	10	11	12	11	9	7	7	8	7	9	9	83.2
uc_Firmicutes	9	8	9	11	9	9	9	8	10	9	12	10	12	12	9	11	11	9	12	11	12	11	13	12	14	13	11	22	13	11	82.1
uc_Proteobacteria	10	14	41	42	14	13	43	42	42	13	13	9	8	9	11	10	12	42	13	14	13	13	12	11	12	12	12	11	16	13	48.8
Veillonellaceae	21	15	42	14	42	14	22	43	12	16	15	14	43	21	43	42	16	23	17	15	14	20	14	15	15	15	14	16	11	14	30.1
Clostridiales_incertae_sedis_XIV	16	16	14	23	15	10	17	23	26	25	21	26	14	29	17	14	15	27	14	31	17	17	27	16	13	19	16	10	17	24	49.3
Eubacteriaceae	15	13	15	29	30	31	19	29	13	14	33	16	17	19	30	32	32	19	16	17	23	14	33	31	16	14	37	23	12	15	43.9
uc_Alphaproteobacteria	20	41	39	40	17	41	42	39	41	41	18	19	13	14	13	21	14	41	24	21	15	24	16	20	17	16	26	20	42	25	39.7
uc_Burkholderiales	17	17	13	16	41	18	14	41	19	43	19	42	42	20	18	13	43	14	25	19	20	16	44	42	21	22	19	24	44	21	22.1
Streptococcaceae	19	39	37	38	38	11	20	37	39	15	41	40	16	16	40	20	20	21	18	12	16	19	41	18	25	41	25	42	40	41	30.8
Alcaligenaceae	23	20	17	18	19	20	24	17	21	19	25	21	20	24	20	17	23	13	15	16	29	15	15	22	18	18	15	15	26	17	64.8
Verrucomicrobiaceae	43	43	43	43	43	43	15	44	43	17	23	43	19	22	15	43	21	43	44	22	27	25	17	43	43	43	18	18	14	16	31.7
Clostridiaceae	27	18	21	22	23	24	28	21	15	23	16	15	21	18	24	18	17	17	29	29	21	30	18	26	20	17	20	17	18	23	62.1
Leuconostocaceae	37	33	31	15	34	15	13	13	17	34	14	17	36	38	34	36	36	35	39	39	40	37	36	35	19	24	17	14	21	36	50.3
Oxalobacteraceae	18	37	35	36	36	38	39	14	37	38	17	38	40	15	38	19	19	38	22	20	18	18	20	17	41	25	23	19	38	20	31.6
uc_Betaproteobacteria	14	42	40	41	40	42	21	40	18	42	43	41	18	17	42	16	42	22	19	43	19	43	43	41	22	26	43	21	43	43	20.8
Erysipelotrichaceae	33	29	27	28	29	30	32	28	30	30	22	31	32	34	29	31	31	18	35	35	37	22	32	13	24	21	36	27	23	19	56.5
Clostridiales_incertae_sedis_XIII	28	24	22	13	24	25	16	22	25	24	20	25	22	28	16	26	27	26	30	30	33	31	26	19	31	31	31	25	30	29	63.7
Synergistaceae	42	40	38	39	39	40	41	38	40	40	42	18	23	43	41	41	41	40	23	24	26	42	42	40	26	42	42	43	41	42	62.6
Moraxellaceae	39	35	33	34	12	36	37	34	35	36	38	36	38	40	36	38	38	36	41	40	25	39	38	37	39	38	22	28	36	38	60.5
Coriobacteriaceae	30	26	24	25	26	27	30	25	16	27	30	28	29	31	26	28	29	29	32	33	22	21	29	28	33	23	33	34	22	31	65.5
Desulfovibrionaceae	31	27	25	26	27	28	31	26	28	28	31	29	30	32	27	29	30	30	33	34	35	33	30	29	23	20	34	35	20	18	74.9
Enterobacteriaceae	32	28	26	27	28	29	18	27	29	29	32	30	31	33	28	30	18	31	34	18	36	34	31	30	34	33	35	26	31	32	63.0
Victivallaceae	44	44	44	44	44	44	44	15	44	44	44	44	44	44	44	44	44	44	26	44	44	44	21	44	44	44	44	44	24	44	53.6
Micrococcaceae	38	34	32	33	35	35	36	33	34	35	37	35	37	39	35	37	37	20	40	23	41	38	37	36	38	37	21	39	35	37	60.1
Bacillaceae	24	21	18	19	20	21	25	18	22	20	26	22	25	25	21	23	24	15	27	26	30	27	23	23	28	28	28	30	27	22	73.1
Comamonadaceae	29	25	23	24	25	26	29	24	27	26	29	27	28	30	25	27	28	28	31	32	34	32	28	27	32	32	32	33	19	30	76.4
Lactobacillaceae	36	32	30	32	33	34	35	32	33	33	36	34	35	37	33	35	35	34	38	38	39	23	35	34	37	36	40	38	34	35	77.0
Fusobacteriaceae	34	30	28	30	31	32	33	30	31	31	34	32	33	35	31	33	33	32	36	36	38	35	19	32	35	34	38	36	32	33	75.9
Neisseriaceae	40	36	34	35	16	37	38	35	36	37	39	37	39	41	37	39	39	37	42	41	42	40	39	38	40	39	41	40	37	39	73.9
Aerococcaceae	22	19	16	17	18	19	23	16	20	18	24	20	24	23	19	22	22	24	20	25	28	26	22	21	27	27	27	29	25	26	76.2
Carnobacteriaceae	25	22	19	20	21	22	26	19	23	21	27	23	26	26	22	24	25	25	21	27	31	28	24	24	29	29	29	31	28	27	77.2
Chloroplast	26	23	20	21	22	23	27	20	24	22	28	24	27	27	23	25	26	16	28	28	32	29	25	25	30	30	30	32	29	28	74.1
Pasteurellaceae	41	38	36	37	37	39	40	36	38	39	40	39	41	42	39	40	40	39	43	42	43	41	40	39	42	40	24	41	39	40	77.5
Halomonadaceae	35	31	29	31	32	33	34	31	32	32	35	33	34	36	32	34	34	33	37	37	24	36	34	33	36	35	39	37	33	34	77.5

Figure 10. Matrix showing the rank variation throughout time for the most dominant elements (taxa) and their calculated Rank Stability Index (as discussed in Material and Methods) in the data for patient A of the IBS study (*3*)

Time dependence of model parameters

Finally, we have studied the time dependence of the variability V and power law index β (see Model under Material and Methods) 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 11 shows the variability V as a function of time for the largest sampling: two individuals in the Caporaso's study (*8*) corresponding to the gut microbiota of a male (upper plot) and a female (lower plot). Figure 12 shows the time evolution of V for patient P2 of the IBS study (*3*) (upper plot) and patient D in the antibiotics study (*12*) (lower plot).

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.

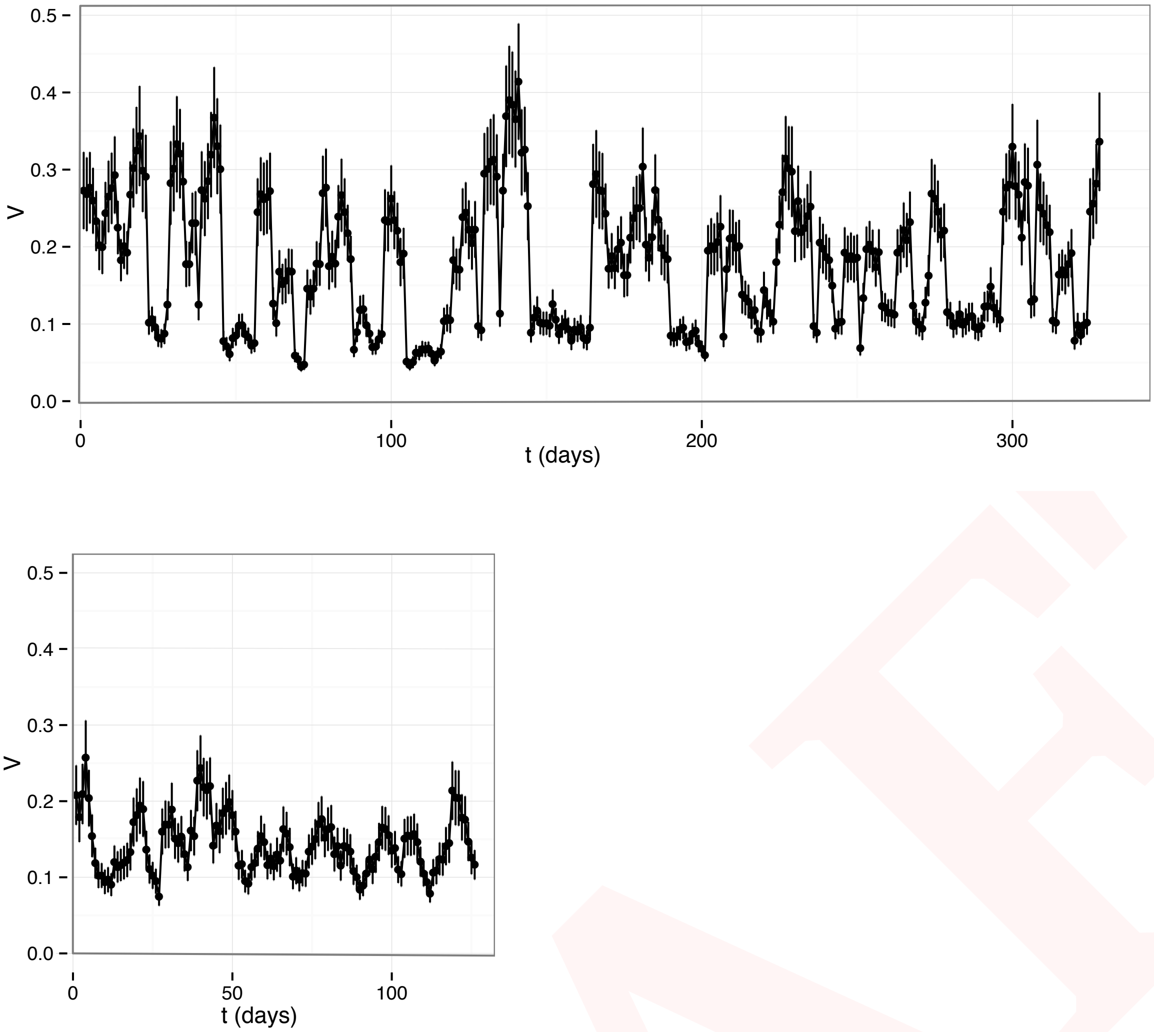
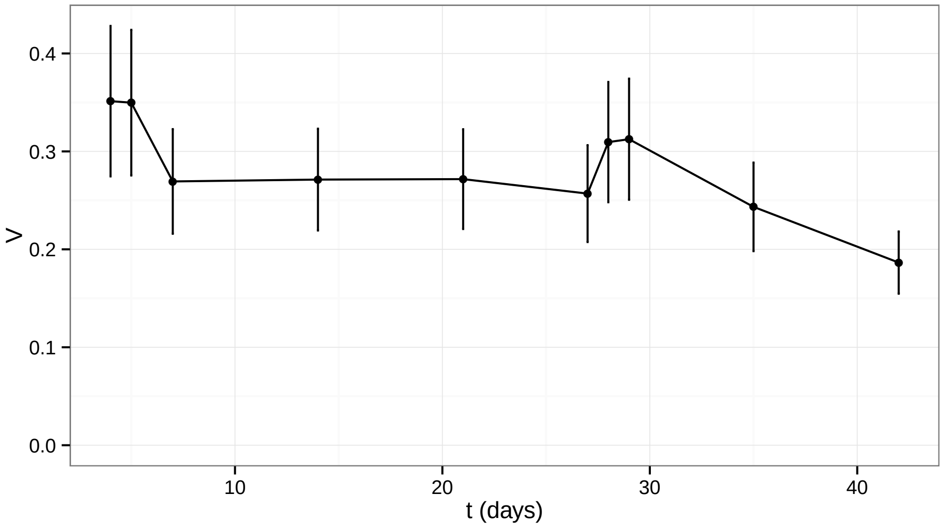


Figure 11. *V* as a function of time for the two individuals in the Caporaso's study (*8*): samples of gut microbiome of a male (upper plot) and a female (lower plot). Both samples show changes in the variability V with quasi–periodic behavior peaked at about 10 days. Variability grows more for the gut microbiota of the male and share a minimal value around 0.1 with the gut microbiota of the female.



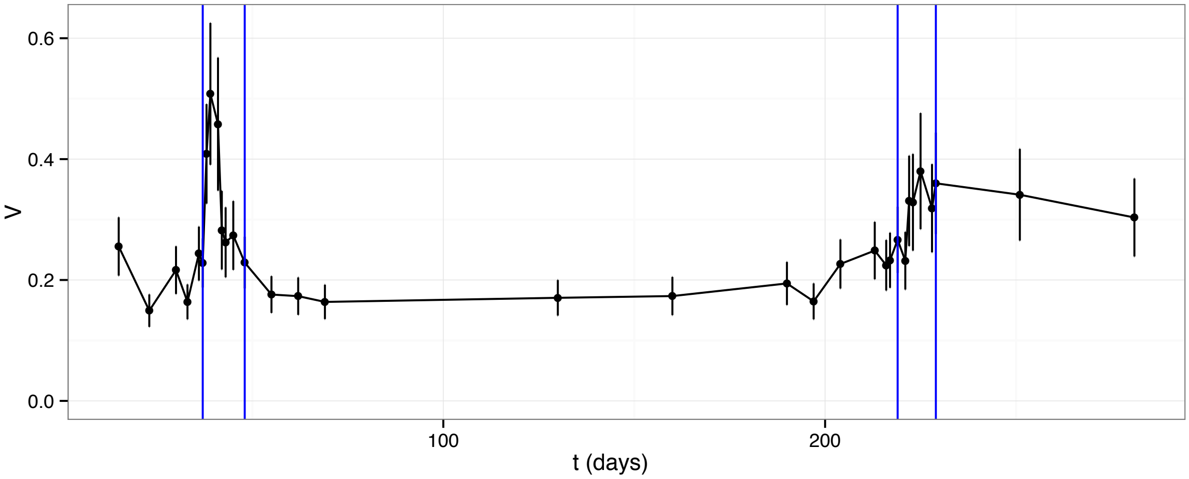


Figure 12. *V* as a function of time for patient P2 of the IBS study (*3*) (upper plot) and patient D in the antibiotics study (*12*) (lower plot). The variability of the gut microbiota of P2 decreases from above 0.3 to below 0.2, showing a slow tendency to increase the order of the system. Antibiotic 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.

Materials and Methods

Model

We model the microbial abundances across time along the lines of Blumm *et al.* (*21*). 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 unstable), 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:

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

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 $\mathrm{V}(\alpha, \beta, F_i)$. The phase transition region can be calculated analytically in particular cases:

$$\begin{aligned} F_i^2 &= 4\beta\phi_0 V^2 \quad \text{if} \quad \beta = \alpha \neq 1, \\ F_i &= \beta V^2 \quad \text{if} \quad 2\beta = 1 + \alpha, \end{aligned}$$

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 (*22*). 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 fluctuation-dissipation theorem

Selection and Methods

The bacteria and archaea taxonomic assignations were obtained by analysing 16S rRNA sequences, which were clustered into operational taxonomic units (OTUs) sharing 97 % sequence identity using QIIME (*13*). WGS data (*10*) were analysed and assigned at strain level by the Livermore Metagenomic Analysis Toolkit (LMAT) (*14*), 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 13).

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)

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 Tables 1 to 6. All used 16S rRNA gene sequencing except for the study of the discordant kwashiorkor twins (*10*) (see 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 (*23*), 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 (*13*) (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 (*24*) (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 (*14*) (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

[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.

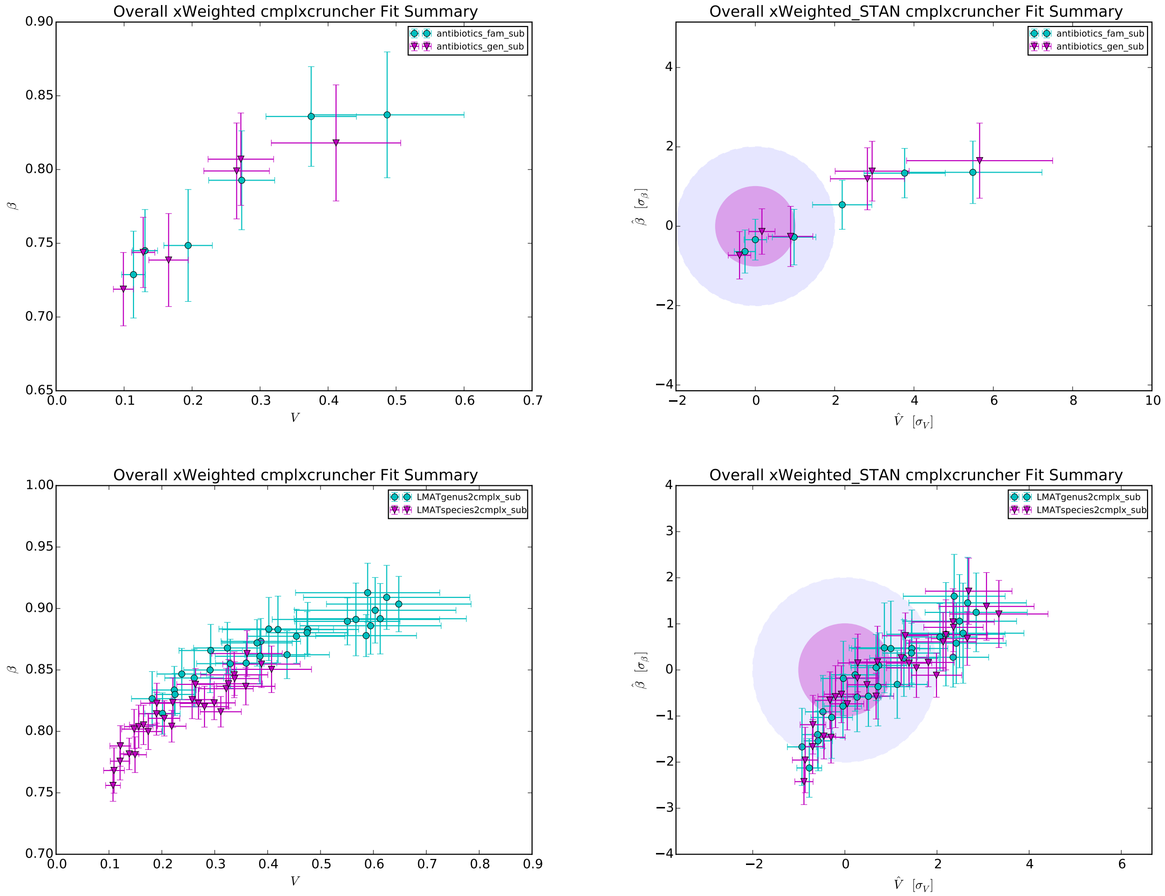


Figure 13. Overview of comparison of different approaches based on adjacent taxonomic levels using plots in the Taylor-parameters space. For 16S (former row of subfigures), the levels are family vs. genus, whereas for SMS (latter row of subfigures) levels are genus vs. species. The left column shows the raw results and the right column plots the standardized results (see Section)

former, the antibiotics study (*12*) with 16S data, we tested the differences between genus and family levels. The latter dataset tested was the kwashiorkor discordant twins study (*10*) for both genus and species taxonomic levels. The Figures 13 (overview) and 14 (detail) plot the comparison between studies (and so, 16S and SMS) and between adjacent taxonomic levels.

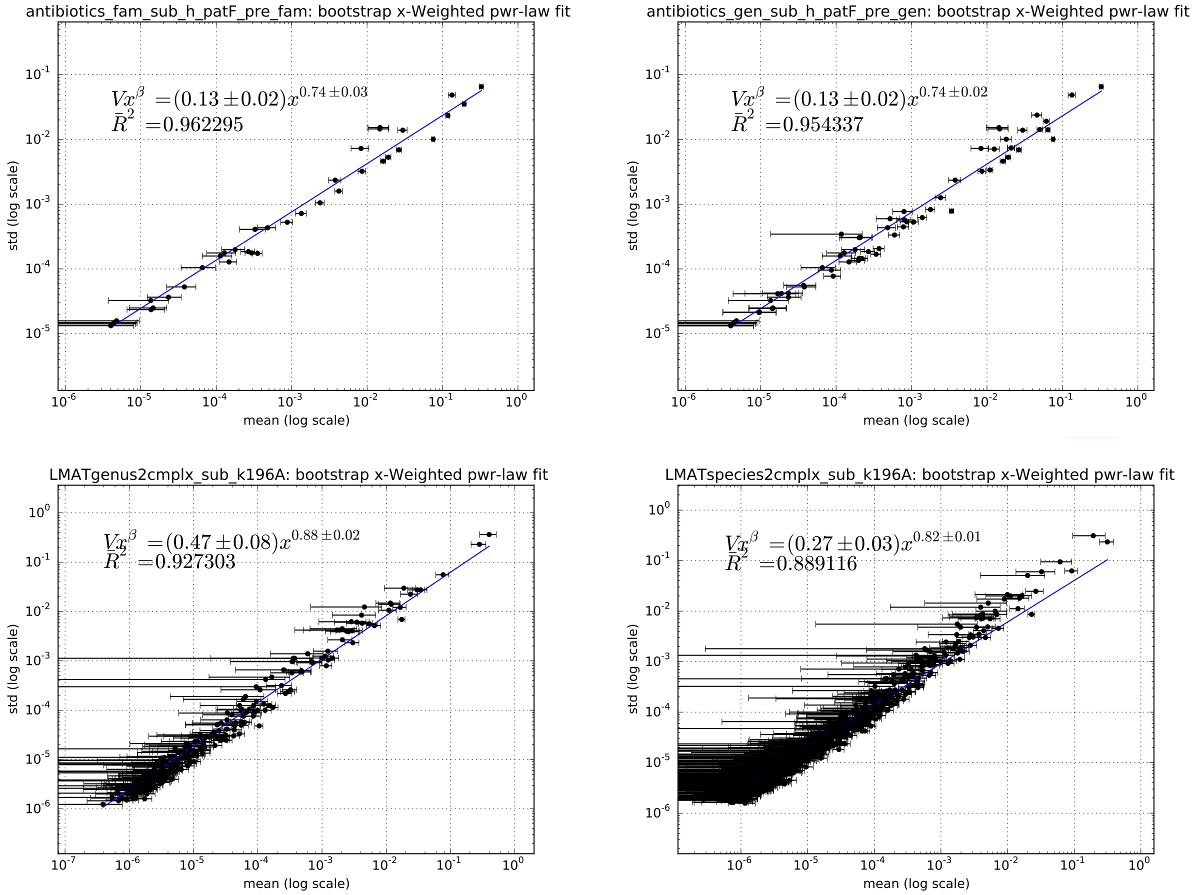


Figure 14. Detail of comparison of different approaches based on adjacent taxonomic levels using plots of X-weighted power-law fits (see Material and Methods). The former row of subfigures shows examples for 16S, whereas the latter row of subfigures plots examples for SMS. The left column shows results 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).

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 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.

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 (*25*). In the power-law fit plots we also show the generalized coefficient of determination computed for continuous models (*26, 27*).

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 (25). 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:

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

   - So, the loglikelihood of a normal distribution is:

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

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

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

   (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:

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

- So, the loglikelihood of a lognormal distribution is:

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

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

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

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:

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

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 (*26*):

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

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 Gaussian

[II] *cmplxcruncher* has available the next bootstrapping alternatives (*28*): ordinary, “Resampling Residuals” method, “Wild” method, and “Monte-Carlo” method.

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:

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

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 $\sigma_i = \text{SEM}_i$, as defined previously. 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

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.

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 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 40).

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 subpopulation, 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. Modificar becas de JMM y DMM ¿poner algún grant más?

References

1. **Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, Liang S, Zhang W, Guan Y, Shen D, Peng Y, Zhang D, Jie Z, Wu W, Qin Y, Xue W, Li J, Han L, Lu D, Wu P, Dai Y, Sun X, Li Z, Tang A, Zhong S, Li X, Chen W, Xu R, Wang M, Feng Q, Gong M, Yu J, Zhang Y, Zhang M, Hansen T, Sanchez G, Raes J, Falony G, Okuda S, Almeida M, LeChatelier E, Renault P, Pons N, Batto J-M, Zhang Z, Chen H, Yang R, Zheng W, Li S, Yang H, Wang J, Ehrlich SD, Nielsen R, Pedersen O, Kristiansen K, Wang J.** 2012. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature **490**:55–60.

2. **Brown JM, Hazen SL.** 2015. The Gut Microbial Endocrine Organ: Bacterially Derived Signals Driving Cardiometabolic Diseases. Annu Rev Med **66**:343–359.

3. **Durbán A, Abellán JJ, Jiménez-Hernández N, Artacho A, Garrigues V, Ortiz V, Ponce J, Latorre A, Moya A.** 2013. Instability of the faecal microbiota in diarrhoea-predominant irritable bowel syndrome. FEMS Microbiol Ecol **86**:581–589.

4. **Gevers D, Kugathasan S, Denson LA, Vázquez-Baeza Y, Van Treuren W, Ren B, Schwager E, Knights D, Song SJ, Yassour M, Morgan XC, Kostic AD, Luo C, González A, McDonald D, Haberman Y, Walters T, Baker S, Rosh J, Stephens M, Heyman M, Markowitz J, Baldassano R, Griffiths A, Sylvester F, Mack D, Kim S, Crandall W, Hyams J, Huttenhower C, Knight R, Xavier RJ.** 2014. The treatment-naive microbiome in new-onset Crohn's disease. Cell Host Microbe **15**:382–392.

5. **Ridaura VK, Faith JJ, Rey FE, Cheng J, Duncan AE, Kau L, Griffi NW, Lombard V, Henrissat B, Bain JR, Michael J, Ilkayeva O, Semenkovich CF, Funai K, Hayashi DK, Lyle J, Martini MC, Ursell LK, Clemente JC, Treuren W Van, William A, Knight R, Newgard CB, Heath AC, Gordon JI, Kau AL, Griffin NW, Muehlbauer MJ.** 2013. Gut Microbiota from Twins Discordant for Obesity Modulate Metabolism in Mice Gut Microbiota from Twins Metabolism in Mice. Science **341**:1241214.

6. **Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, Egholm M, Henrissat B, Heath AC, Knight R, Gordon JI.** 2009. LETTERS A core gut microbiome in obese and lean twins. Nature **457**:480–484.

7. **Subramanian S, Huq S, Yatsunenko T, Haque R, Mahfuz M, Alam MA, Benezra A, DeStefano J, Meier MF, Muegge BD, Barratt MJ, VanArendonk LG, Zhang Q, Province MA, Petri WA, Ahmed T, Gordon JI.** 2014. Persistent gut microbiota immaturity in malnourished Bangladeshi children. Nature **510**:417–21.

8. Caporaso, J.G. et al. Moving pictures of the human microbiome. *Genome Biol.* **12,** R50 (2011).

9. Faith, J.J. et al. The long-term stability of the human gut microbiota. *Science* **341,** 1237439 (2013).

10. Smith M.I. et al. Gut microbiomes of Malawian twin pairs discordant for kwashiorkor. *Science* **339,** 548-54 (2013).

11. David, L.A. et al. Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505,** 559-63 (2014).

12. 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).

13. Caporaso, J.G. et al. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7,** 335-6 (2010).

14. Ames, S.K. et al. Scalable metagenomic taxonomy classification usng a reference genome database. *Bioinformatics* **29,** 2253-2260 (2013).

15. Eisler,Z., Bartos,I., Kertesz,J. Fluctuation scaling in complex systems: Taylor's law and beyond. *Adv. Phys.* **57,** 85 (2008).

16. Taylor, L.R. Aggregation, Variance and the mean. *Nature* **189,** 732-35 (1961).

17. Jorgensen,B., Martinez,J.R., Tsao,M. Asymptotic behaviour of the variance function. *Scand. J. Statist.*. **21,** 223-243 (1994).

18. Fronczak,A., Fronczak,P. Origins of Taylor's power law for fluctuation scaling in complex systems. *Phys. Rev. E* **81,** 066112 (2010).

19. 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).

20. Kendal, W.S., Jorgensen,B. Tweedie convergence: A mathematical basis for Taylor's power law. *Phys. Rev. E* **84,** 066120 (2011).

21. Blumm, N. et al. Dynamics of ranking processes in complex systems. *Phys. Rev. Lett.* **109,** 128701 (2012).

22. Weber, J. *et al.* Fluctuation dissipation theorem. *Phys. Rev.* **101**, 1620-6 (1956).

23. Gordon, A., Hannon, G.J. FASTX-Toolkit. FASTQ/A shortreads pre-processing tools (2010). http://hannonlab.cshl.edu/fastx_toolkit/ (accessed 23 Feb 2015).

24. Quast C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools (2013)

25. 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).

26. Magee L., R^2 measures based on wald and likelihood ratio joint significance tests. *The American Statistician* **44**, 3, 250-253 (1990).

27. Nagelkerke N.J.D., A note on a general definition of the coefficient of determination. *Biometrika* **78**, 3, 691-692 (1991).

28. 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...