**Material and Methods**

We have analysed more than 35000 time series of taxa from the gut microbiome 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 span1, people with various degrees of obesity2, twin pairs discordant for kwashiorkor3, response to diet changes4 or to antibiotic perturbation5, 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 supplemental material). The bacteria and archaea taxonomic assignations were obtained by analyzing 16S rRNA sequences, which were clustered into operational taxonomic units (OTUs) sharing 97 % sequence identity using QIIME7. WGS data3 were analyzed 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).

### WC = 192

### merge whole point 2 in supplementary to the main text

### añadir toda la parte de complexCruncher

### 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)

**1. Model**

We model the microbial abundances across time along the lines of Blumm et al. (1). The dynamics of taxon relative abundances is described by the Langevin equation:

x ̇i=Fi·xαi +V·xβiξi(t)−φ(t)·xi, (1)

where Fi captures the fitness of the taxon i, V corresponds to the noise amplitude and ξi(t) is a Gaussian random noise with zero mean < ξi(t) > = 0 and variance uncorrelated in time, < ξi(t)ξi(t′) > = δ(t′ −t), . The function φ(t) ensures the normalization at all times, xi(t)=1,andcorrespondstoφ(t)= Fixαi + Vxβiξi(t). Thetemporalevolution of the probability that a taxon i has a relative abundance xi(t), P(xi,t), is determined by the Fokker-Planck equation: ∂P ∂ α 1∂2 2 2β

∂t =−∂x[(Fi·xi −φ(t)·xi)·P]+2∂x2(V ·xi ·P). (2) ii

   

The microbiota evolves towards a steady-state with a time-independent probability de- pending on the values of α, β, Fi and V. For α < 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 Fi dominates or is overwhelmed by the noise amplitude V. The steady-state solution of the Fokker-Planck equation is given by:

2F x1+α−2β P0(xi) = Cne(α,β,Fi,V)·x−2β ·exp i i

φ x2−2β − 0 i

if 2β̸=1+α,

i V21−β where φ0 = ( F1/(1−α))1−α and Cne and Ce are integrals that should be solved numer-

ically for the parameters of interest. The ordered phase happens when the solution has a maximum in the physical interval (0 < xi < 1). For larger V, the transition to a dis- ordered phase happens when the maximum shifts to the unphysical region xi < 0, which sets the phase transition region V(α, β, Fi). The phase transition region can be calculated analytically in particular cases:

Fi2 = 4βφ0V2 if β=α̸=1, Fi = βV2 if 2β=1+α,

where the first case, simplifies to F = 3V 2 if β = 0.75 and the fitness of this taxon dominates in φ0. 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 (2). 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).

**2.1 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:

**2.2 16rRNA sequences processing**

Reads from the selected studies were first quality filtered using the FastX toolkit (4), 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 (5) (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 (6) (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.

**2.3 Metagenomic sequences processing**

Metagenomic shotgun (and 16S too) sequences were analyzed with LMAT (Livermore Metagenomics Analysis Toolkit) software package (7) (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” databaseI, 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.

**2.4 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 12), we tested two different data sets. In the former, the antibiotics study (8) 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 1 (overview) and 2 (detail) plot the comparison between studies (and so, 16S and SMS) and between adjacent taxonomical levels.

12 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 (V ) for the healthy subpopula- tion, composed of h individuals, is:

1hh

[equation]

as [equation], since ωi are normalized weights calculated as: 1

being σVi the estimation of the uncertainty in Vi obtained together with Vi from the X-weighted power-law fit described in Section 8, for healthy individuals.

Likewise, the estimation of the standard deviation for the healthy population (σ V ) is: [equation]

[equation]

being W2 = hiωi2, which finally yields to:

[equation]