Universality of Human Microbial Dynamics: Supplementary Information

Amir Bashan, Travis E. Gibson, Jonathan Friedman, Vincent J. Carey, Scott T. Weiss, Elizabeth L. Hohmann, Yang-Yu Liu

May 2, 2016

Contents

1 Methodology

1.1 Dissimilarity and Overlap Measures

We want to detect the effect of inter-species ecological interactions on the abundance profiles of microbial communities by analyzing the interplay between two measures of diversity: Dissimilarity and Overlap. We define those two measures to be mathematically independent such that ecological inter-species interactions will lead to a characteristic pattern in the Dissimilarity-Overlap relation.

1.1.1 Dissimilarity

We quantified the dissimilarity $D(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ between the renormalized abundance profiles of the n shared species of microbial sample pair $\hat{\mathbf{x}} = \{\hat{x}_i\}_{i\in S}$ and $\hat{\mathbf{y}} = \{\hat{y}_i\}_{i\in S}$. Here $\hat{x}_i \equiv \frac{\tilde{x}_i}{\sum_{j\in S}\tilde{x}_j} = \frac{x_i}{\sum_{j\in S} x_j}$, x_i and $\tilde{x}_i$ are the absolute abundance and relative abundance of species i, respectively, and S ($|S| = n$) is the set of shared species (present in both samples). $\hat{\mathbf{y}}$ is defined similarly. There are many dissimilarity measures widely used in ecology and biology:

i. Jensen-Shannon Divergence (JSD) [4]

$$D_{\mathrm{JSD}}(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = \frac{1}{2}\left[\mathrm{KLD}(\hat{\mathbf{x}}, \mathbf{m}) + \mathrm{KLD}(\hat{\mathbf{y}}, \mathbf{m})\right], \tag{S1}$$

where $\mathbf{m} \equiv \frac{\hat{\mathbf{x}}+\hat{\mathbf{y}}}{2}$ and KLD is the Kullback-Leibler divergence between $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ defined as $\mathrm{KLD}(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = \sum_{i=1}^{n} \hat{x}_i \log \frac{\hat{x}_i}{\hat{y}_i}$.

ii. Root-JSD (rJSD) [7]

$$D_{\mathrm{rJSD}}(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \equiv \sqrt{D_{\mathrm{JSD}}(\hat{\mathbf{x}}, \hat{\mathbf{y}})}. \tag{S2}$$

iii. Bray-Curtis (BC) [3]

$$D_{\mathrm{BC}}(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \equiv \frac{\sum_{i=1}^{n} |\hat{x}_i - \hat{y}_i|}{\sum_{i=1}^{n} \hat{x}_i + \hat{y}_i}. \tag{S3}$$

iv. Yue-Clayton (YC) [14]

$$D_{\mathrm{YC}}(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \equiv 1 - \Theta_{\mathrm{YC}}(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = 1 - \frac{\sum_{i=1}^{n} \hat{x}_i \hat{y}_i}{\sum_{i=1}^{n} \hat{x}_i^2 + \hat{y}_i^2 - \hat{x}_i \hat{y}_i}. \tag{S4}$$

v. Negative Spearman Correlation (nSC) [5]

$$D_{\mathrm{nSC}}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) = D_{\mathrm{nSC}}(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \equiv 1 - \rho = \frac{6\sum d_i^2}{n(n^2-1)}. \tag{S5}$$

where ρ is the Spearman's rank correlation coefficient (also known as Spearman's rho), d_i is the difference between the ranks of x_i and y_i.

A dissimilarity measure $D(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ is a *distance metric* if it satisfies the following conditions:

i. $D(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \geq 0$ (non-negativity)

ii. $D(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = 0$ if and only if $\hat{\mathbf{x}} = \hat{\mathbf{y}}$ (identity)

iii. $D(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = D(\hat{\mathbf{y}}, \hat{\mathbf{x}})$ (symmetry)

iv. $D(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \leq D(\hat{\mathbf{x}}, \hat{\mathbf{z}}) + D(\hat{\mathbf{z}}, \hat{\mathbf{y}})$ (triangle inequality).

Among the five dissimilarity measures presented above, which are all widely used for comparison of microbial samples, only the rJSD is a distance metric. We show in the main text the results calculated using $D_{\mathrm{rJSD}}(\hat{\mathbf{x}}, \hat{\mathbf{y}})$.

1.1.2 Renormalization of shared species' abundance.

Consider two vectors $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$ of length N, representing the relative abundances of two communities X and Y, such that $\tilde{x}_i \equiv x_i / \sum_{j=1}^{N} x_j$ and $\tilde{y}_i \equiv y_i / \sum_{j=1}^{N} y_j$, where x_i and y_i are the absolute abundance of species i in the two communities, respectively. Note that the absolute abundances are typically not available in microbiome data. We denote as S the set of shared species exist in both communities ($\tilde{x}_i > 0$ and $\tilde{y}_i > 0$ for all $i \in S$). The set of unique (non-shared) species of $\tilde{\mathbf{x}}$, i.e. species that exist only in $\tilde{\mathbf{x}}$ (or $\tilde{\mathbf{y}}$), is denoted as U_x (or U_y, respectively).

Importantly, the relative abundances of the shared species, $\tilde{x}_i$ and $\tilde{y}_i$ ($i \in S$) depend, due to the compositionality, on the relative abundance of other species in U_x and U_y. In order to eliminate this spurious dependence, we renormalize the common parts of x and y

$$\hat{x}_i \equiv \frac{\tilde{x}_i}{\sum\limits_{j \in S} \tilde{x}_j} = \frac{x_i / \sum_{k \in S, U_x} x_k}{\sum\limits_{j \in S} x_j / \sum_{k \in S, U_x} x_k} = \frac{x_i}{\sum\limits_{j \in S} x_j}. \tag{S6}$$

The renormalized abundance profiles $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ depend only on species in S and are not affected by the species in U_x and U_y. Consequently, the dissimilarity calculated over the renormalized abundance profiles of shared species does not depend on the non-shared species in U_x and U_y.

Note that when using the Negative Spearman Correlation as a dissimilarity measure between the shared species, the renormalization is not required.

Table S1 demonstrates the renormalization of shared species' abundances. In this case, species 1 and 2 are shared in X and Y while species 3 and 4 appear only in Y. The relative abundances $\tilde{y}_1$ and $\tilde{y}_2$ depend on the absolute abundances of both the shared and the non-shared species of Y. However, the renormalized abundances $\hat{y}_1$ and $\hat{y}_2$ depend only on the shared species 1 and 2 and not on the unique species 3 and 4.

i	1	2	3	4
x	x_1	x_2	0	0
y	y_1	y_2	y_3	y_4
$\tilde{x}$	$\frac{x_1}{x_1+x_2}$	$\frac{x_2}{x_1+x_2}$	0	0
$\tilde{y}$	$\frac{y_1}{y_1+y_2+y_3+y_4}$	$\frac{y_2}{y_1+y_2+y_3+y_4}$	$\frac{y_3}{y_1+y_2+y_3+y_4}$	$\frac{y_4}{y_1+y_2+y_3+y_4}$
$\hat{x}$	$\frac{x_1}{x_1+x_2}$	$\frac{x_2}{x_1+x_2}$		
$\hat{y}$	$\frac{y_1}{y_1+y_2}$	$\frac{y_2}{y_1+y_2}$		

Table S1: The renormalized abundances of the shared species $\hat{y}_1$ and $\hat{y}_2$ do not depend on the non-shared species abundance y_3 and y_4 at all.

1.1.3 n-dependence of the dissimilarity measures

The dissimilarity measures (D_{JSD}, D_{rJSD}, D_{BC} and D_{YC}) are methods of measuring the dissimilarity between two probability distributions. As such, they are designed to be independent on n for large n ($n \gg 1$).

To demonstrate this effect of n-independence in the general context of normalized abundance distributions, we consider two vectors $\hat{\mathbf{x}}, \hat{\mathbf{y}} \in \mathbb{R}^n$ and $\sum_{i=1}^{n} \hat{x}_i = \sum_{i=1}^{n} \hat{y}_i = 1$ representing two probability distributions binned into n bins. We then define two new vectors, $\hat{\mathbf{x}}'$ and $\hat{\mathbf{y}}'$, each with $2n$ elements by splitting each element of $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ to two elements.

In the limit of large n we have

$$\hat{x}'_{2i-1} = \hat{x}'_{2i} \simeq \frac{\hat{x}_i}{2}, \text{ and } \hat{y}'_{2i-1} = \hat{y}'_{2i} \simeq \frac{\hat{y}_i}{2} \text{ for } i = 1, ..., n. \tag{S7}$$

For the vector $\mathbf{m}' \equiv \frac{\hat{\mathbf{x}}'+\hat{\mathbf{y}}'}{\mathbf{2}}$, we have

$$m'_{2i-1} = m'_{2i} \simeq \frac{m_i}{2} \text{ for } i = 1, ..., n. \tag{S8}$$

Hence,

$$\mathrm{KLD}(\hat{\mathbf{x}}', \mathbf{m}') = \sum_{j=1}^{2n} \hat{x}'_j \log \frac{\hat{x}'_j}{m'_j} \simeq \sum_{i=1}^{n} 2\hat{x}'_{2i} \log \frac{\hat{x}'_{2i}}{m'_{2i}} = \sum_{i=1}^{n} 2\frac{\hat{x}_i}{2} \log \frac{\hat{x}_i}{m_i} = \mathrm{KLD}(\hat{\mathbf{x}}, \mathbf{m}) \tag{S9}$$

thus, $D_{\mathrm{JSD}}(\hat{\mathbf{x}}', \hat{\mathbf{y}}') \simeq D_{\mathrm{JSD}}(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ and $D_{\mathrm{rJSD}}(\hat{\mathbf{x}}', \hat{\mathbf{y}}') \simeq D_{\mathrm{rJSD}}(\hat{\mathbf{x}}, \hat{\mathbf{y}})$, independent of n. Similarly,

$$D_{\mathrm{BC}}(\hat{\mathbf{x}}', \hat{\mathbf{y}}') = \frac{1}{2} \sum_{j=1}^{2n} |\hat{x}'_j - \hat{y}'_j| \simeq \frac{1}{2} \sum_{i=1}^{n} 2|\hat{x}'_{2i} - \hat{y}'_{2i}| = \frac{1}{2} \sum_{i=1}^{n} 2|\frac{\hat{x}_i}{2} - \frac{\hat{y}_i}{2}| = D_{\mathrm{BC}}(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \tag{S10}$$

and

$$D_{\mathrm{YC}}(\hat{\mathbf{x}}', \hat{\mathbf{y}}') = 1 - \frac{\sum_{j=1}^{2n} \hat{x}'_j \hat{y}'_j}{\sum_{j=1}^{2n} \hat{x}'^2_j + \hat{y}'^2_j - \hat{x}'_j \hat{y}'_j} \simeq 1 - \frac{2\sum_{i=1}^{n} \hat{x}'_{2i} \hat{y}'_{2i}}{2\sum_{i=1}^{n} \hat{x}'^2_{2i} + \hat{y}'^2_{2i} - \hat{x}'_{2i} \hat{y}'_{2i}}$$
$$= 1 - \frac{2\sum_{i=1}^{n} \hat{x}_i \hat{y}_i/4}{2\sum_{i=1}^{n} (\hat{x}_i^2 + \hat{y}_i^2 - \hat{x}_i \hat{y}_i)/4} = D_{\mathrm{YC}}(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \tag{S11}$$

This illustrates the fact that for large n the dissimilarity measures remain the same even though the number of bins is changed. However, for small n, the above approximation (S7) is invalid. In particular, for $n = 1$, by definition $\hat{x}_1 = \hat{y}_1 = 1$, and obviously, for any dissimilarity measure, $D(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = D(1, 1) = 0$.

To systematically study the effect of n on the dissimilarity measures numerically, we show in Extended Data Fig. 10 the average dissimilarity between two independent normalized random vectors of length n. We found that, indeed, for large n the mean dissimilarity $\langle D(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \rangle$ (in all dissimilarity measures) doesn't depend on n and the value depends on the distributions of the random elements. As n increases, $\langle D(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \rangle$ monotonically increases. As n reaches certain value, $\langle D(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \rangle$ saturates and displays no n-dependence. This behavior explains, for example, why in some body sites (e.g. right/left retroauricular crease) the dissimilarity measure decreases to zero for pairs with low overlap.

Note that for the case of negative Spearman Correlation, which considers only the rank of the elements, $\langle D_{\mathrm{nSC}}(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \rangle = 1$ independent on the length of the vectors n even for small n. This feature is reflected in the DOC results of gut microbial samples shown in Extended Data Fig. 5 comparing the different dissimilarity measures.

1.1.4 Overlap

The Overlap measure is calculated from the relative abundance of the shared species only. We can show that it actually represents the ratio between the absolute abundance of the shared (subset S) and the non-shared species (subsets U_x and U_y)

$$O(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) = \sum_{i \in S} \frac{\tilde{x}_i + \tilde{y}_i}{2} = \frac{1}{2} \left(\frac{1}{1 + \frac{\sum_{i \in U_x} x_i}{\sum_{i \in S} x_i}} \right) + \frac{1}{2} \left(\frac{1}{1 + \frac{\sum_{i \in U_y} y_i}{\sum_{i \in S} y_i}} \right). \tag{S12}$$

So $O(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ is strongly influenced by non-shared species. Specifically, adding a new species to one community, say Y, will increase the term $\sum_{i \in U_y} y_i$ and decrease the overlap.

In the extreme case when the relative abundance is the same for all species in the two communities X and Y, the overlap measure can be written as a function of the classical Jaccard index. Consider both X and Y contain N equally abundant species ($N \geq n$), and $n = |S|$ of them are shared, the Jaccard index is then

$$J(\mathrm{X}, \mathrm{Y}) \equiv \frac{|\mathrm{X} \cap \mathrm{Y}|}{|\mathrm{X} \cup \mathrm{Y}|} = \frac{n}{2N - n} \tag{S13}$$

and the overlap is

$$O(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) = \frac{n}{N} = \frac{2J}{1+J}. \tag{S14}$$

Note that $2J/(1+J)$ is nothing but the Sørensen's index[9].

There are three key reasons why we introduced the Overlap measure instead of adopting the classical Jaccard index in our DOC analysis:

i. Robust to noise: Real microbial samples are characterized by highly diverse abundance profiles. The relative abundances of some species are close to the detection limit, hence subject to high "noise" in the presence/absence signal. Unlike the classical Jaccard index, our overlap measure weights the contributions of species based on their relative abundances, naturally reducing the "noise" effect due to species with extremely low abundance.

ii. Robust to "OTU-splitting": True OTUs are often "split" artificially into multiple very closely related "sub-OTUs" due to sequencing error and sometimes to stochasticity in the heuristic alignment tools being used. We demonstrate this point in Table R2, where we show that the Jaccard index is very sensitive to this OTU-splitting issue, whereas our Overlap measure is quite robust to it.

iii. Consistent with population dynamics: In our DOC analysis, the negative slope of DOC in the high-overlap regime is a characteristic feature of universal dynamics with strong inter-species interactions. The key idea is that two samples with similar species collections should have similar abundance profiles, whereas the presence of non-shared species will change the abundances of the shared species via inter-species interactions. Importantly, the impact of a non-shared species (i) on any other species (j) it interacts with is a function of its own abundance x_i, rather than simply its presence. For example, in the canonical Generalized Lotka-Volterra (GLV) model, the impact that species i has on the population change of species j is simply given by $a_{ji}x_i$, where a_{ji} accounts for the interaction strength. Therefore, we expect that the impact of the non-shared species on the dissimilarity between the abundance profiles of shared species to be a function of their abundances rather than their presence only. This fact has been explicitly considered in the Overlap measure, but not in the Jaccard index.

Due to the above considerations, our Overlap measure is not only novel, but also more appropriate than the classical Jaccard index to explore the universality of microbial dynamics.

1.2 Dissimilarity-Overlap Curve Analysis

1.2.1 $O(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ has no mathematical constraints on $D(\hat{\mathbf{x}}, \hat{\mathbf{y}})$

For a given sample pair, the two quantities (overlap and dissimilarity) can be presented as a point in the dissimilarity-overlap plane. Importantly, $D(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ is well

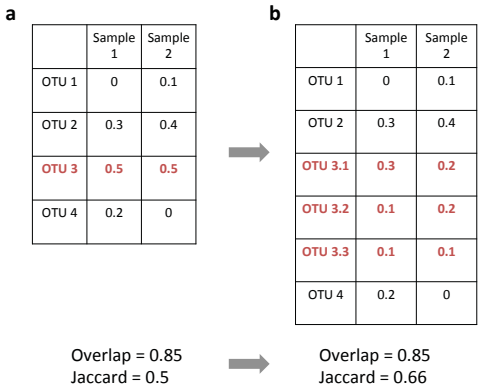

Table S2: The Overlap measure is fairly robust to the OTU-splitting issue, whereas Jaccard index is sensitive to it. a, example of two relative abundance profiles. 'OTU 2' and 'OTU 3' are shared in Sample 1 and Sample 2. **b**, 'OTU 3' is split to three sub-OTUs: 'OTU 3.1', 'OTU 3.2' and 'OTU 3.3'. The splitting changes the Jaccard index, but the Overlap measure remains the same in this scenario.

defined for any $O(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) > 0$ without any mathematical constraints. This is because $D(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ is a function of renormalized abundances of the shared species only, which are independent of the non-shared species, while $O(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ depends on the ratio between the total abundance of the shared and non-shared species (see Eq. S12). Therefore, any point over the dissimilarity-overlap plane can be horizontally shifted (i.e. change its overlap without changing its dissimilarity) by changing the ratio between the total abundance of the unique and that of the shared species while fixing the ratios within the shared species. Similarly, any point can be shifted vertically by doing the opposite.

1.2.2 Mathematical and ecological dependencies

The renormalization of the shared species abundances (S6) is performed to filter out any mathematical dependence between the Overlap and the Dissimilarity measures (in the high-overlap regime). This allows us to detect their ecological dependency and hence reject the null hypothesis of no-universal-interactions. In the absence of universal interactions (either individual dynamics or no inter-species interactions at all), a flat DOC is expected due to the mathematical independence between the Dissimilarity and the Overlap. However, if the species are ecologically interacting, e.g. due to the fact of finite energy resources and substrates in the gut, and if those inter-species interactions are universal (host-independent), then a negative slope in the high-overlap regime of the DOC is expected. (See Sec.3 for an alternative explanation of this phenomenon.)

(a)

No inter-species interactions

i	1	2	3	4	5	$O(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$	$D_{\mathrm{rJSD}}(\hat{\mathbf{x}}, \hat{\mathbf{y}})$
x	10	20	30				
$y^{(0)}$	10	20	30			1	0
$y^{(1)}$	10	20	30	5		0.96	0
$y^{(2)}$	10	20	30	5	15	0.88	0

(b)

inter-species interactions

i	1	2	3	4	5	$O(\tilde{\mathbf{x}}, \tilde{\mathbf{z}})$	$D_{\mathrm{rJSD}}(\hat{\mathbf{x}}, \hat{\mathbf{z}})$
x	10	20	30				
$z^{(0)}$	10	20	30			1	0
$z^{(1)}$	11	21	28	5		0.96	0.02
$z^{(2)}$	13	24	23	5	15	0.88	0.08

Table S3: The negative slope of the Dissimilarity-Overlap Curve (DOC). In (a), samples $\mathbf{x}$ and $\mathbf{y}^{(0)}$ have the same species assemblage and same abundances. At this stage, the overlap is $O(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}^{(0)}) = 1$ and the dissimilarity is $D(\hat{\mathbf{x}}, \hat{\mathbf{y}}^{(0)}) = 0$. In the next steps, new species $y_4^{(1)}$ and $y_5^{(2)}$ are introduced to $\mathbf{y}$, while $\mathbf{x}$ remains unchanged. The overlap decreases in each step ($O(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}^{(1)}) = 0.96$, $O(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}^{(2)}) = 0.88$), however, the dissimilarity (measured as $D_{\mathrm{rJSD}}(\{\hat{x}_j\}, \{\hat{y}_j\})_{j \in A}$) remains the same. This case represents an ecosystem without inter-species interactions. In this case the DOC is flat. In contrast, in (b), sample $\mathbf{z}$ represents an ecosystem with inter-species interactions. The invasions of species $z_4^{(1)}$ and $z_5^{(2)}$ change the abundance of the shared species (z_1, z_2 and z_3). Thus, the dissimilarity increases ($D(\hat{\mathbf{x}}, \hat{\mathbf{z}}^{(0)}) = 0$, $D(\hat{\mathbf{x}}, \hat{\mathbf{z}}^{(1)}) = 0.02$, $D(\hat{\mathbf{x}}, \hat{\mathbf{z}}^{(2)}) = 0.08$) as the overlap decreases ($O(\tilde{\mathbf{x}}, \tilde{\mathbf{z}}^{(0)}) = 1$, $O(\tilde{\mathbf{x}}, \tilde{\mathbf{z}}^{(1)}) = 0.96$, $O(\tilde{\mathbf{x}}, \tilde{\mathbf{z}}^{(2)}) = 0.88$). In this case the DOC has a negative slope.

1.2.3 The DOC displays a negative slope in the presence of inter-species interactions

In this sub-section we demonstrate the impact of adding a new species to one of the two communities (Table S3). We compare two cases: a) a community without inter-species interactions and b) a community with inter-species interactions. In the first case, the added species affects only the overlap but not the dissimilarity. In contrast, in the second case, both the overlap and the dissimilarity are affected. These examples shown in Table S3 demonstrate the relation between Overlap and Dissimilarity, that is the slope of DOC, in the high overlap region. Note that in Table S3 we begin with a case where two samples with the same collection also have the same abundance profile.

Let us consider two microbial communities X and Y that have the same underlying ecological dynamics and initially the same abundance profiles (i.e. $\tilde{x}_i = \tilde{y}_i$ for all i). In this case, the Overlap $O(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) = 1$ and the Dissimilarity $D(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = 0$. Now we consider a new species invades one of the communities, say Y, resulting in a smaller Overlap $O(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$. The newly added species to Y may change the abundances of other species in Y that it directly interacts with. The new abundance profile $\hat{\mathbf{y}}'$ will be less similar to that of the unchanged community $\hat{\mathbf{x}}$, i.e., $D(\hat{\mathbf{x}}, \hat{\mathbf{y}}') > D(\hat{\mathbf{x}}, \hat{\mathbf{y}})$. This invasion process can be repeated many times, and each newly added species reduces the Overlap and increases the Dissimilarity between the two communities, rendering a negative slope of the DOC (see Table S3 for a simple example). Importantly, the compositionality of the relative abundance data results in an upper bound of the dissimilarity measure (See Sec. 1.2.4). Therefore, the negative slope is expected only for high-Overlap values while below a certain critical Overlap value

the Dissimilarity level saturates.

In contrast, in the case of (i) individual dynamics; or (ii) universal dynamics but without inter-species interactions, a flat DOC is expected. In case (i), higher overlap of two communities, or even sharing exactly the same species assemblage, will not lead to more similar abundance profiles, due to the different underlying dynamics. In case (ii), the non-shared species have no effect on the abundances of the shared species, thus, the dissimilarity between the renormalized abundance profiles of the shared species is Overlap-independent. In both cases we expect a flat DOC.

1.2.4 Boundness of the dissimilarity measure

The relative abundance of each species is by definition bounded in $[0, 1]$. Moreover, due to the compositionality, a point in the N-dimensional state space representing the relative abundance of N species is bounded in the "simplex" of $(N-1)$ dimensions. As a dissimilarity measure between compositional samples, D_{JSD} is bounded in $[0, \log(2)]$ and hence D_{rJSD} is bounded in $[0, \sqrt{\log(2)}]$. The D_{BC} and D_{YC} measures are bounded in $[0, 1]$. Note that the upper bound of dissimilarity represents the dissimilarity between two *extremely different* compositions, e.g. $\mathbf{x} = \{1, 0, 0, ...\}$ and $\mathbf{y} = \{0, 1, 0, 0, ...\}$.

These constraints affect the pattern of the DOC. As demonstrated above in Table S3, the negative slope of DOC at high overlap range is explained as follows: the limit of $O(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) = 1$ represents the case of two communities with identical species assemblages. Assume that they have in addition identical species abundance profiles. Adding a new species to one of the communities i) decreases the overlap level (see Sec. 1.1.4); ii) may change the ecosystem's steady state and, thus, increase the dissimilarity. Representing the steady state of a system as a point in an N-dimensional space, changes of species abundance profiles are represented as shifts of the steady state to a new position. The direction of the shift and its magnitude are determined by the vectorial sum of the changes in all the coordinates (species abundances) which is a function of inter-species interactions. This process can be repeated many times and in some sense is similar to an N-dimensional random walk. Consider a point $\mathbf{x}^{(0)}$ in an N-dimensional space, representing a steady state of a microbial ecosystem. At each step, a random N-dimensional "walk" (or displacement) $\boldsymbol{\delta}^{(t)}$ is added with random direction and fixed length $|\boldsymbol{\delta}| = 1$. The location at step t is $\mathbf{x}^{(t)} = \mathbf{x}^{(t-1)} + \boldsymbol{\delta}^{(t)}$ with a constraint that the random walk is forced to stay in the positive orthant. Then, $\mathbf{x}^{(t)}$ is normalized such that the sum of all its coordinations is one (projection over the simplex plane), $\tilde{x}_i^{(t)} = x_i^{(t)} / \sum_j x_j^{(t)}$. The total displacement at step t is measured as D_{rJSD} distance between the normalized location at time t, i.e. $\tilde{\mathbf{x}}^{(t)}$ and the normalized initial state $\tilde{\mathbf{x}}^{(0)}$.

For small t, the Dissimilarity increases as t increases since the random walk is still close to its initial state and will not be affected by the boundness. When t is larger than a certain value, the Dissimilarity value saturates, because after so many changes the probability to find the random walk at certain state is approximately equal all over the bounded state space.

Similarly, in the case of universal dynamics, at the region of high overlap (anal-

ogous to small t) a negative slope is observed, i.e. Dissimilarity increases as the Overlap decreases (analogous to increasing t). Above a certain Overlap value the Dissimilarity value saturates.

1.3 Null models

1.3.1 The purpose of using null models in DOC analysis

In this paper, we use null models to demonstrate flat DOCs. The comparison between the DOCs of the real and the null model qualitatively demonstrates the effect of the real inter-species interactions. The DOCs of real microbial samples from certain body sites as well as synthetic samples calculated from the classical GLV model have a clear negative slope in the high-overlap region. This is in marked contrast with the DOCs of the samples from null models, which are always flat.

Nevertheless, the quantitative analyses, including the statistical tests, are independent of any null model (see Methods).

1.3.2 Two null models

Null model 1. We aim to study the universality feature in a cohort of samples as a result of possible significant inter-species interactions. To achieve that, we compared the results of the real data with a randomized model that removes the effect of true inter-species interactions but *preserves the species assemblages* and abundance distributions. Let $\tilde{x}_{ij}$ denote the relative abundance of species i in sample j, where $i = 1, ..., N$ and $j = 1, ..., M$. We generate Q sets, each of M randomized samples (Q is an integer, in our case $Q = 5$) so in total we have $Q \times M$ randomized samples. For each set

$$R_{ij} = \begin{cases} 0 & \text{if } \tilde{x}_{ij} = 0 \\ \tilde{x}_{ik} & \text{otherwise} \end{cases}$$

where k is a random index for which $\tilde{x}_{ik} > 0$. The new sample R_{ij} preserves the collection of species of the real sample j but the abundances are taken from different random samples. All the R_{ij}'s are then normalized to one.

Null model 2. Here the randomized samples are generated *without restrictions on the species assemblage*. In other words, not only the abundances are randomized but also the collections. The drawback of the null model is that non-realistic species assemblages may appear.

Both null models largely preserve the rank order abundance of the real data and the overlap distribution. However, the species richness (i.e. the number of species in a sample) distribution of the real data is perfectly preserved by null model 1, but not by null model 2. Therefore, we used null model 1 in the main text.

(a)

Original samples

Species	a	b	c	d	e	f
$\tilde{\mathbf{x}}^{(1)}$	0.33	0.33	0.34	0	0	0
$\tilde{\mathbf{x}}^{(2)}$	0.4	0.4	0.2	0	0	0
$\tilde{\mathbf{x}}^{(3)}$	0	0	0	0.33	0.33	0.34
$\tilde{\mathbf{x}}^{(4)}$	0	0	0	0.4	0.4	0.2

(b)

Null model 1

Species	a	b	c	d	e	f
$R^{(1)}$	0.33	0.4	0.34	0	0	0
$R^{(2)}$	0.4	0.33	0.2	0	0	0
$R^{(3)}$	0	0	0	0.33	0.4	0.34
$R^{(4)}$	0	0	0	0.4	0.33	0.2

(c)

Null model 2

Species	a	b	c	d	e	f
$R^{(1)}$	0.33	0	0.2	0	0.4	0
$R^{(2)}$	0	0.4	0	0.33	0	0.2
$R^{(3)}$	0.4	0.4	0.34	0.4	0.4	0.34
$R^{(4)}$	0	0	0	0	0	0

Table S4: Different randomizations of microbial samples yield different null models. The original samples **(a)** are divided to two clusters (in samples 1,2 only species "a", "b" and "c" exist and in samples 3,4 only species "d", "e" and "f"). Null model 1 **(b)** preserves this order of species assemblage but selects at random the abundances. Null model 2 **(c)** does not preserve this order and may introduce species assemblages that never appear in the real data.

2 DOC analysis: strengths and caveats

2.1 Core species with non-interacting periphery

In the presence of (1) core species (shared by all subjects) that are interacting among themselves; and (2) non-shared (peripheral) species that are not interacting with any other species, the DOC will be flat, even if core species interact in a host-independent (universal) manner. We confirmed this theoretical expectation with numerical simulations (see the blue curve in Extended Data Fig. 8 **c4**). Moreover, we find that as long as the interactions among the peripheral species themselves (periphery-periphery) and interactions between the peripheral species and the core species (periphery-core) have a non-zero characteristic strength ($\sigma_{\rm p} > 0$), the DOC has a pronounced negative slope (see the red and green curves in Extended Data Fig. 8 **c4**).

We point out that the negative slope of the DOC is not a property of the presence of core species. As shown in Extended Data Fig. 8 **c3**, in our simulations with the GLV model, we assumed that the species have very similar probability to be present

in any local community (subject). In other words, the core species that are shared by all subjects do not exist in our model. We find a clear negative slope in the high-overlap regime of the DOC, though. Real data analysis showed that a few species have higher presence probability than others (see Extended Data Fig. 8 **c2**), leading to a soften version of "core species". This also results in a negative slope in the high-overlap regime of the DOC.

2.2 Even and diverse abundance distributions

In order to rule out effects of the abundance distributions on our DOC analysis, we did the following:

i. When analyzing a given cohort of real samples we also analyzed the randomized samples generated from the real samples (see SI Sec.1.3). The randomized samples ("null model 1") largely preserve the abundance distribution and species richness of the real samples, but the effect of inter-species interactions, if exists, is completely removed. This way, a difference between the patterns of the DOC of the real samples and that of the randomized samples is solely due to the different feature (inter-species interactions) rather than the preserved one (abundance distribution).

ii. In Extended Data Fig. 8 **b**, we compared the DOCs in two cases: 1) homogeneous (even) abundance distribution and 2) heterogeneous (skewed) abundance distribution. In case 1), synthetic microbial samples were generated from the steady states of the GLV model with or without inter-species interactions (Extended Data Fig. 8 **b1, b2**). In the presence of inter-species interactions, a negative slope of the DOC is observed (Extended Data Fig. 8 **b5**). In contrast, in the absence of inter-species interactions, we observe a flat DOC (Extended Data Fig. 8 **b6**). Real gut microbiome samples (from the HMP study, at the genus level) exhibit a high level of alpha-diversity and a very skewed abundance distribution (Extended Data Fig. 8 **b3**). A negative slope in the DOC is observed (Extended Data Fig. 8 **b7**). The randomized samples preserve the abundance distribution of the real samples but the effect of inter-species interactions is completely removed (Extended Data Fig. 8 **b4**). In this case the DOC is flat (Extended Data Fig. 8 **b8**). This suggests that the role of inter-species interactions is the key feature captured by the DOC analysis rather than the abundance distributions.

2.3 Sequencing depth and rarefaction

The number of observed species in a microbial sample ("richness") is a function of the sequencing depth, i.e. the number of sequences collected in the genomic survey. In order to compare features such as richness, alpha-diversity etc., the samples have to be rarefied to standardize the effective sequencing depth. We tested the effect of sequencing depth and rarefaction on the DOC analysis.

We first confirmed in Extended Data Fig. 8 **d1** that the increase of species richness (or equivalently, the decrease of mean taxon prevalence) with increasing sequencing depth in 190 gut microbiome samples from the HMP study. In order to test how different sequencing depths and rarefaction techniques affect our DOC

analysis, we divided the HMP gut microbiome samples into two groups as follows. According to the standard rarefaction technique, 12 samples with less than 1,300 reads/sample were excluded. The remaining 178 samples were then assigned into two groups of equal size: m=89 samples). A sample is assigned to group-1 (or group-2) if its sequencing depth is smaller (or larger) than 4,300 reads, respectively (see Extended Data Fig. 8 **d1**). The sequencing depth does affect the average overlap between samples; the average overlap between samples of group-1 is slightly smaller than the average overlap between samples of group-2. However, the negative slope of DOC is clearly observed in both cases (Extended Data Fig. 8 **d2, d3**).

Next, we performed the following analysis. We first rarefied each group using the minimal sequencing depth as a threshold in each of the groups (see Extended Data Fig. 8 **d4**). Again we observed negative slope in the DOC for both groups (Extended Data Fig. 8 **d5, d6**). Then we rarefied the samples of both groups using the same threshold determined as the minimal sequencing depth (Extended Data Fig. 8 **d7**). This time, the average overlaps in the two groups are very similar. Again, we observed negative slope in the DOC for both groups (Extended Data Fig. 8 **d8, d9**).

The insensitivity of our DOC analysis to the sequencing depth actually demonstrates clearly the big advantage of our overlap measure over the classical Jaccard index. Deeper sequencing allows for the discovery of more low-abundance species. Yet, in contrast to the Jaccard index that considers only the presence/absence of species, our overlap measure also considers the species abundance. Hence, the effect of those low-abundance species is reduced, which explains why the overlap is more robust to the sequencing depths. The fact that the DOC analysis is insensitive to sequencing depth suggests that the signal of the universal dynamics is observed mostly in the highly abundant species (which are less affected by rarefaction) rather than the low abundance species (which are highly affected by rarefaction).

3 Analysis of host factors

An alternative explanation for the observed negative slope of the DOC calculated from human gut and mouth microbiomes could be that some host factors not only select for the presence of certain microbes but also drive their relative abundances by enforcing certain optimally adapted compositions. However, we systematically analyzed microbial samples while controlling for the effect of several leading candidates for potential confounding factors, e.g. body mass index, age, race, long-term dietary pattern and stool consistency, and show that as long as their values are in the normal range they cannot explain the observed DOC pattern.

3.1 Methodology

To test the alternative hypothesis, we systematically studied the impact of the following host factors on our DOC analysis: 1) Body Mass Index (BMI); 2) Diet; 3) Age; 4) Stool consistency; 5) Race, which have been previously shown to be

associated with the gut microbiome. To this aim, we analyzed the metadata of the Human Microbiome Project (HMP) as well as two additional datasets of healthy populations with host metadata of interest. For each dataset, we systematically performed the DOC analysis while controlling for host factors as follows. For a given host factor x, we quantified the difference between two subjects i and j as $\Delta x_{ij} \equiv |x_i - x_j|$. For examples, $\Delta BMI_{ij} \equiv |BMI_i - BMI_j|$ is the absolute BMI difference between subjects i and j. We then studied (1) the association between Δx and Overlap (O); and (2) the association between Δx and Dissimilarity (D), for each host factor x, in order to test whether subjects with similar host factor (small Δx) tend to have more similar microbiomes. In addition, we grouped the sample pairs according to their host-factor differences (Δx) and plotted the DOC for each group. For example, in the group of low ΔBMI we have only sample pairs with the same, or very similar, BMI values. This way we can filter the possible influence of the host factors.

3.2 The effect of BMI

Abnormal BMI values (obesity) have been shown to be associated with changes in the relative abundance of the two dominant bacterial divisions (*Bacteroidetes* and *Firmicutes*) in mouse gut microbiota [10]. These changes affect the metabolic potential and increase the capacity of gut microbiota to harvest energy from the diet. Hence, it is a very legitimate concern that BMI could be a confounding factor of our DOC analysis. To address this concern, we study gut microbiome samples from 190 healthy subjects (at the OTU level) from the HMP study, with BMI range 19-34 (mean 24). We first checked if samples from obese subjects (13 of the subjects have $BMI > 30$) have abnormal overlap-dissimilarity values. Extended Data Fig. 7 **a1** shows the overlap and dissimilarity values of all sample pairs, where the blue points represent pairs of normal-weight subjects (both with $BMI \leq 30$) and the red points represent sample pairs associated with obese subjects (each sample pair contains at least one subject with $BMI > 30$). The red points are uniformly scattered over the cloud with no special tendency, compared with the blue points. In addition, comparing the DOC of all subjects (Extended Data Fig. 7 **a1**) and the DOC of only non-obese subjects (Extended Data Fig. 7 **a2**) we found no effect on the DOC. Next, we tested the effect of BMI similarity on the overlap and dissimilarity. Extended Data Fig. 7 **a3, a4** show (**a3**) the overlap and (**a4**) dissimilarity of pairs versus their BMI difference, ΔBMI. Samples from subjects with similar BMI have the same average overlap and average dissimilarity as subjects with very different BMI values. Finally, we divided the sample pairs into four groups according to their ΔBMI values (Extended Data Fig. 7 **a5**). The DOCs calculated for different groups show qualitatively similar negative slope in their respective DOCs (Extended Data Fig. 7 **a6-a9**).

In sum, the DOC analysis on healthy population is driven neither by abnormal BMI values nor by BMI differences.

3.3 The effect of diet

The nutrition intake has been shown to have a large impact on the gut microbiome [1, 13]. This raises a natural concern: Different individuals that share similar diet may have similar microbiota (in terms of both species collection and abundance profile), which can be an alternative explanation of our observation, i.e. the negative slope of the DOC.

Studies on the effect of short-term diet on the human microbiome have shown that short-term extreme diet (e.g. animal-based diet) rapidly alters the gut microbiome of each individual compared to his or her personal baseline [1], but doesn't drive the entire microbial communities of different subjects to the same state, as found by Wu et al. [13]: "Over 10 days of controlled feeding, there was no reduction in UniFrac distances for stool or biopsy samples between individuals fed the same diet, demonstrating that a short-term identical diet does not overcome intersubject variation".

To study whether the same long-term diet leads to similar microbial communities, we analyzed the data published by Wu et al. [13]: A cross-sectional dataset of 97 healthy subjects and their habitual long-term diet information measured by food frequency questionnaire. We defined the diet difference between two subjects as the Euclidean distance between the projections of their diet profiles on the plane of the two leading principal components (PC1 and PC2), as illustrated in Extended Data Fig. 7 **b1**. We also tried the Euclidean distance between two diet profiles in the original space, finding very similar results. Note that the Euclidean distance has been frequently used in cluster analysis of dietary patterns [6, 8, 12]. Extended Data Fig. 7 **b2, b3** show that the average overlap and average dissimilarity are independent upon the diet difference. In other words, people who consume the same diet do not have more similar microbial communities than those who consume different diets. By splitting the sample pairs to four groups according with their diet-difference (Extended Data Fig. 7 **b4**) and analyzing the DOC of each group separately (Extended Data Fig. 7 **b5-b8**), we observed a negative slope in each group, ruling out diet as a potential confounding factor of our DOC analysis.

We emphasize that there is no contradiction between our results and Wu et al.'s work [13]. First of all, species-nutrient associations presented in [13] do not imply that two subjects with similar diet must have high overlap of their microbiome. Second, Wu et al. compared the average diet associated with enterotypes, while we compare the individual dietary patterns in a pair-wise manner. These are two fundamentally different measures.

3.4 The effect of age

To study the effect of age difference, we analyzed the HMP gut microbiome samples from subjects with ages between 18-40. We found no correlation between age difference and Overlap (or Dissimilarity). Subjects with similar (or even the same) age do not tend to have more similar microbial communities than subjects of different ages (Extended Data Fig. 7 **c1,c2**). The negative slope of DOC is observed in pairs

of same or similar age as well as in pairs of large age difference (Extended Data Fig. 7 **c3-c7**).

3.5 The effect of stool consistency

Recent studies reported that the stool consistency, quantified by the Bristol stool scale (BSS, a discrete value between 1-7), is strongly associated with the richness and composition of gut microbiota. A liquid stool (with high BSS value) may affect the ecology of the microbiome due to the large amount of water and the shorter transit time in the gut[2, 11].

To rule out this potential confounding factor of our DOC analysis we analyze the dataset from [2] of 53 subjects with BSS values between 1 and 6. We first checked whether high BSS values are associated with abnormal overlap and dissimilarity values. Extended Data Fig. 7 **d1** shows that many sample pairs associated with at least one subject with $BSS = 6$ (red symbols) tend to have high dissimilarity values. Indeed, $BSS = 6$ suggests a tendency towards diarrhea or even inflammation and thus we excluded 7 subjects with $BSS = 6$. The remaining 46 subjects with BSS between 1 and 5 display a clear negative slope (Extended Data Fig. 7 **d2**), even when divided into two groups according to their BSS differences (Extended Data Fig. 7 **d5-d7**) and the Overlap and Dissimilarity are independent on the BSS difference (Extended Data Fig. 7 **d3, d4**). We conclude that for the broad range of normal stool consistency ($1 \leq BSS < 6$) the DOC is not confounded by differences in the stool consistency. Extreme cases ($BSS \geq 6$) might lead to abnormal dissimilarity and/or overlap behavior, which may ruin the normal negative slope. The extremely liquid stool (with $BSS \geq 6$) of patients with recurrent *C. Diff.* infection may be a reason for not observing the normal negative slope in DOC, i.e. undetectable universal microbial dynamics.

3.6 The effect of race

In the HMP study, the majority of subjects (153 of 190) are white. In order to filter confounding variables due to race differences, we repeated the DOC analysis for the white people only. In this case the negative slope is clearly observed, thus, it is not driven by race differences (Extended Data Fig. 7 **e2**). The second largest group consists of 25 Asian people. The DOC of those subjects is flat, that is, no universality was detected Extended Data Fig. 7 **e3**. Even though the possible effect of race on the universality of microbial dynamics is intriguing, the current result is still inconclusive since the size of this group is too small and we also lack a larger dataset from African Americans and Hispanics. Therefore, after ruling out the possible confounding effect due to mixing different races, we limit our conclusion to the Caucasian population. The question of possible race effects remains open.

3.7 Summary

To summarize, we tested the five leading host factors as possible confounders on our DOC analysis. We found that none of them alone could explain the negative slope in the DOC analysis. This could be due to the fact that the assembly processes are highly host-specific with complex contribution of several factors including immigration and stochasticity. Consequently, similar host factors do not imply similar microbial communities of different subjects. While those host factors were found to be associated with the presence/absence or abundance level of species in a microbial community, here we show that they are not associated with the overlap and dissimilarity of two microbial communities.

Though the alternative hypothesis is intriguing (because it does not require any dynamics model or inter-species interactions), we admit that with currently available datasets we cannot possibly account for all other potential confounders, e.g. drugs, genetics, inflammation, or combinations of them. More datasets will be needed to test their effects on the DOC analysis and hence *directly* verify the alternative hypothesis.

Here we mention an *indirect* way to disprove this alternative hypothesis. Indeed, if this hypothesis were true, the overall $> 90\%$ cure rate of fecal microbiota transplantation (FMT) in treating recurrent *C. difficile* Infection (rCDI) patients will be very questionable. Indeed, those recipients and donors have different host factors before the FMT (e.g. the donors ages vary between 18-50, and the recipients ages independently vary between 7-90), and the recipients were not asked to do anything to mimic the lifestyle of their donors after the FMT.

We argue that universal microbial dynamics (which can be parameterized by meaningful ecological parameters) is so far the simplest or most parsimonious model to explain the observed negative slope in the DOC. Other models or explanations will not be superior to this model when cost and complexity are taken into account. Indeed, allowing for non-universal dynamics may fit data at least as well as the universal dynamics model, but non-universal models typically require more parameters. And the cost of the additional parameters or model complexity is not adequately compensated by improvement to the likelihood criterion.

References

[1] Lawrence A. David, Corinne F Maurice, Rachel N Carmody, David B Gootenberg, Julie E Button, Benjamin E Wolfe, Alisha V Ling, a Sloan Devlin, Yug Varma, Michael a Fischbach, Sudha B Biddinger, Rachel J Dutton, and Peter J Turnbaugh, *Diet rapidly and reproducibly alters the human gut microbiome.*, Nature **505** (2014), 559–63.

[2] Jack A Gilbert and John Alverdy, *Stool consistency as a major confounding factor affecting microbiota composition: an ignored variable?*, Gut (2015), available at `http://gut.bmj.com/content/early/2015/07/17/gutjnl-2015-310043.full.pdf+html`.

[3] J. T. Curtis J. Roger Bray, *An ordination of the upland forest communities of southern wisconsin*, Ecological Monographs **27** (1957), no. 4, 326–349.

[4] J. Lin, *Divergence measures based on the shannon entropy*, IEEE Trans. Inf. Theor. **37** (September 1991), no. 1, 145–151.

[5] Jerome L. Myers, Arnold D. Well, and Robert F. Lorch Jr, *Research design and statistical analysis*, 3rd ed., Routledge, 2010.

[6] P. K. Newby and Katherine L. Tucker, *Empirically derived eating patterns using factor or cluster analysis: A review*, Nutrition Reviews **62** (2004), no. 5, 177–203, available at `http://nutritionreviews.oxfordjournals.org/content/62/5/177.full.pdf`.

[7] Ferdinand Österreicher and Igor Vajda, *A new class of metric divergences on probability spaces and its applicability in statistics*, Annals of the Institute of Statistical Mathematics **55** (2003), no. 3, 639–653.

[8] Jill Reedy, Elisabet Wirflt, Andrew Flood, Panagiota N. Mitrou, Susan M. Krebs-Smith, Victor Kipnis, Douglas Midthune, Michael Leitzmann, Albert Hollenbeck, Arthur Schatzkin, and Amy F. Subar, *Comparing 3 dietary pattern methodscluster analysis, factor analysis, and index analysiswith colorectal cancer risk: The nihaarp diet and health study*, American Journal of Epidemiology **171** (2010), no. 4, 479–487, available at `http://aje.oxfordjournals.org/content/171/4/479.full.pdf+html`.

[9] Thorvald Julius Sørensen, *A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons*, I kommission hos E. Munksgaard, København, 1948.

[10] Peter J Turnbaugh, Ruth E Ley, Michael a Mahowald, Vincent Magrini, Elaine R Mardis, and Jeffrey I Gordon, *An obesity-associated gut microbiome with increased capacity for energy harvest.*, Nature **444** (2006), 1027–31.

[11] Doris Vandeputte, Gwen Falony, Sara Vieira-Silva, Raul Y Tito, Marie Joossens, and Jeroen Raes, *Stool consistency is strongly associated with gut microbiota richness and composition, enterotypes and bacterial growth rates*, Gut (2015), available at `http://gut.bmj.com/content/early/2015/06/11/gutjnl-2015-309618.full.pdf+html`.

[12] A. K. Elisabet Wirfält and Robert W. Jeffery, *Using cluster analysis to examine dietary patterns: Nutrient intakes, gender, and weight status differ across food pattern clusters*, Journal of the American Dietetic Association **97** (1997), no. 3, 272 –279.

[13] Gary D Wu, Jun Chen, Christian Hoffmann, Kyle Bittinger, Ying-Yu Chen, Sue a Keilbaugh, Meenakshi Bewtra, Dan Knights, William a Walters, Rob Knight, Rohini Sinha, Erin Gilroy, Kernika Gupta, Robert Baldassano, Lisa Nessel, Hongzhe Li, Frederic D Bushman, and James D Lewis, *Linking long-term dietary patterns with gut microbial enterotypes.*, Science (New York, N.Y.) **334** (2011), 105–8.

[14] Jack C. Yue and Murray K. Clayton, *A similarity measure based on species proportions*, Communications in Statistics - Theory and Methods **34** (2005), no. 11, 2123–2131, available at `http://dx.doi.org/10.1080/STA-200066418`.