Ecology meets human health: studies on human gut microbiota in health and disease

Pinto, S.

Chapter 3

Wavelet clustering analysis for characterizing community structure

Elisa Benincà[1], Susanne Pinto[2], Bernard Cazelles[3, 4, 5], Susana Fuentes[1], Sudarshan Shetty[1, 6], Johannes A. Bogaards[7, 8]

1 Centre for Infectious Disease Control, National Institute for Public Health and the Environment (RIVM), Bilthoven, the Netherlands
2 Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, the Netherlands
3 UMMISCO, Sorbonne Université, Paris, France
4 INRAE, Université Paris-Saclay, MaIAGE, Jouy-en-Josas, France
5 Eco-Evolution Mathématique, IBENS, UMR 8197, CNRS, Ecole Normale Supérieure, Paris, France
6 Department of Medical Microbiology and Infection prevention, Virology and Immunology research Group, University Medical Center Groningen, the Netherlands
7 Department of Epidemiology and Data Science, Amsterdam UMC location Vrije Universiteit Amsterdam, Amsterdam, the Netherlands
8 Amsterdam Institute for Infection and Immunity (AI&I), Amsterdam UMC, Amsterdam, the Netherlands

Sci Rep. 2023 May 17;13(1):8042. doi: 10.1038/s41598-023-34713-8.

Wavelet clustering analysis as a tool for characterizing community structure in the human microbiota

Abstract

Human microbiota research is helped by the characterization of microbial networks, as these may reveal key microbes that can be targeted for beneficial health effects. Prevailing methods of microbial network characterization are based on measures of association, often applied to limited sampling points in time. Here, we demonstrate the potential of wavelet clustering, a technique that clusters time series based on similarities in their spectral characteristics. We illustrate this technique with synthetic time series and apply wavelet clustering to densely sampled human gut microbiota time series. We compare our results with hierarchical clustering based on temporal correlations in abundance, within and across individuals, and show that the cluster trees obtained using either method are significantly different in terms of elements clustered together, branching structure, and total branch length. By capitalizing on the dynamic nature of the human microbiota, wavelet clustering reveals community structures that remain obscured in correlation-based methods.

Introduction

The human microbiota is the collective of microbial communities living on the various surfaces of the human body. These communities consist of microorganisms which do not live in isolation but interact with each other and with their human host.[252, 284] In the past decade, thanks to advances in sequencing techniques and data analyses, an increasing number of studies have attempted to gain ecological insights from microbiota abundance data, e.g., by reconstructing networks of interacting species with the nodes representing the microorganisms and the edges representing the dependencies between them.[285]

Most of the studies that aim to reconstruct the network of interacting species are based on measures of co-occurrence, e.g., using correlations between pairs of species as proxies of between-species dependencies.[8, 250, 286] Despite the popularity of such methods in microbiota studies,[230, 250, 251] their usefulness in describing community structure is still a matter of debate.[114, 229, 287] While these co-occurrence studies are often performed on a relatively large number of individuals, they are limited to one or a few sampling points in time, presenting a mere snapshot of the dynamic microbiota. Other methods infer the ecological network by fitting an a priori chosen population-dynamic model to time series data of the microbial community.[100, 254, 288] These methods have the limitation that the inferred community structures strongly rely upon the assumptions that are intrinsic to the chosen model, and require considerable prior knowledge of the community of interest. There are also examples where the ecological interactions are inferred from repeated measurements around steady states.[255] This circumvents the need for a priori specification of a population dynamic model but makes the implicit assumption that the microbial system tends towards a stable equilibrium.

However, many experimental and field studies have shown the presence of complex dynamics in ecological communities, such as alternative stable states,[2, 97, 105] oscillations, and chaos,[5, 6, 281, 282] questioning the steady states assumptions for the human microbiota. These dynamics are driven by a complex interplay between intrinsic factors (e.g., interaction mechanisms between organisms such as competition, mutualism, and parasitism) and external perturbations (e.g., environmental conditions and interventions).[6, 289, 290] Complex dynamics are also likely to occur in the human microbiota, because the bacterial communities living in our body are characterized by a plethora of interactions[291] and are also affected by external perturbations (e.g., diet, use of antibiotics, and travel patterns).[59, 292, 293] A study with a thousand healthy western individuals suggested the existence of tipping elements in the intestinal microbiota,[102] reflecting the presence of alternative attractors and the possibility of more complex microbiota dynamics. The presence of complex dynamics in the human microbiota has not yet been demonstrated, probably due to the paucity of long and dense time series of the human microbiota. However, the study with one of the longest time series of human microbiota measurements available shows strong variability in the abundance of the bacteria over time, indicating that the human microbiota might not be at the presumed steady state.[46]

To advance our ecological understanding of the human microbiota, methodology is needed that can exploit the temporal information in microbiota time series data without a priori knowledge of data generating mechanisms or steady-state assumptions. In the last decade, many methods have been developed to model the abundances of compositionally sampled data with the purpose of either fitting or predicting the temporal dynamics of the microbiota communities.[294-296] Here, we perform wavelet clustering analysis, a technique that clusters time series based on similarities in their periodical patterns.[297] This technique, which is commonly applied in climate and engineering studies,[298] more recently gained popularity in ecological,[290] and epidemiological studies.[299-301] Wavelet clustering analysis has only recently been applied to time series derived from 16S rRNA gene amplicon data to reveal coastal plankton community structure,[302] but, to our knowledge, our study is the first application to human gut microbiota data. The novelty of the wavelet clustering approach, relative to prevailing co-occurrence or time series methodologies in human microbiota research, is that it is able to characterize community structure on the basis of collective temporal behaviour of the microbiota, without directly fitting a dynamic model or reconstructing the network of interacting species.

We illustrate wavelet clustering first with synthetic time series and then with densely sampled time series of human gut microbiota data from a male and female subject.[46] For both examples, we compare our results with clustering obtained on the basis of correlations in bacterial abundances over time. Our results show that correlation-based clustering is significantly different from clustering using wavelets. Wavelet clustering uncovered more diverse community structures and retained more of the differences between the male and the female subject compared to methods using temporal correlation. The results of this work highlight how the choice of method determines the type of communities found in microbiota data analysis. This is particularly important, considering that most of the putative microbiota communities, and their associations with a particular disease state or physical host condition, strongly rely on prevailing correlation-based methods or steady-state assumptions. Our results suggest that wavelet clustering readily capitalizes on the dynamic nature of the human microbiota and reveals more diverse community structures than those based on temporal correlations or associations.

Methods

Wavelet analysis

Wavelet analysis makes use of a periodic function (the mother-wavelet). The relative importance of periodicities (wavelet power) is then plotted in contour plots as a function of time (wavelet power spectra). Here, we use as mother-wavelet the Morlet wavelet, which is particularly suited for detecting periodicities.[298, 303] Significance of the detected periodicities is assessed using a Markov surrogate significance test.[304] Statistical significance is assessed by testing against the null hypothesis that observed periodicities are identical to those generated by a stochastic Markov process, characterized by the same mean, the same variance, the same distribution of values and the same short-term autocorrelation structure. More detailed information on wavelet analysis is provided elsewhere.[6, 305-307]

Wavelet clustering

The wavelet spectra are compared using a procedure based on the maximum covariance analysis.[297] To be more precise, as described in Rouyer, Fromentin et al. (2008), the distance matrix is computed based on leading patterns and singular vectors obtained using matrix decomposition analysis.[297] Matrix decomposition analysis relies on a singular value decomposition performed on the covariance matrix between two wavelet power spectra. This enables construction of a distance matrix based on the wavelet power spectra. Only periodicities with a confidence higher than 90% have been considered in the computation of the dissimilarity matrix. Wavelet analysis and wavelet clustering were performed using wavelet software written in Matlab which is available at Bernard Cazelles' research page (www.biologie.ens.fr/~cazelles/bernard/Research.html).[297]

Comparison among cluster trees

We quantified similarities between cluster trees using the B_k statistic (i.e., Fowlkes-Mallows index).[308] The B_k statistic measures the degree of similarity between two hierarchical clusters. Consider two hierarchical trees C_1 and C_2, each with the same number of elements n and partition each tree to produce $k = 2, \ldots, n-1$ subclusters for each tree. For each value of k we can compute the quantity $m_{i,j}$ which quantifies the number of objects in common between the i^{th} cluster of C_1, and the j^{th} cluster of C_2. The statistic B_k is then defined:

$$B_k = \frac{T_k}{\sqrt{P_k Q_k}} \qquad \textbf{Eq. 3.1}$$

where:

$$T_k = \sum_{i=1}^{k}\sum_{j=1}^{k} m_{i,j}^2 - n \qquad \textbf{Eq. 3.2}$$

$$P_k = \sum_{i=1}^{k} \left(\sum_{j=1}^{k} m_{i,j}\right)^2 - n \qquad \textbf{Eq. 3.3}$$

$$Q_k = \sum_{j=1}^{k} \left(\sum_{i=1}^{k} m_{i,j}\right)^2 - n \qquad \textbf{Eq. 3.4}$$

B_k is calculated for all the k partitions and B_k takes values between 0 and 1; $B_k = 1$ indicates that k subclusters in each tree correspond completely whereas $B_k = 0$ indicates that the subclusters in each tree don't correspond at all.

Details on the B_k statistic are described in Fowlkes et al. (1983).[308] The B_k statistic has been calculated using the 'dendextend' R package.[309] The computed values of B_k are then plotted as a function of k. The significance of the B_k values is tested against the null hypothesis that the two cluster trees are not related. A one-sided rejection line (with significance level of 5%) is drawn based on the asymptotic distribution of B_k values, for each k, under the null hypothesis of no relation between the clusters.

Calculation of total branch length

The total branch length was calculated by summing the lengths of connecting segments in the tree using the 'treeheight' function of the 'vegan' R package.[310]

Microbiota data

In our analysis, we used previously published time series of the gut microbiota of two healthy subjects, one male and one female, on which fecal samples have been taken for 15 and 6 months, respectively.[46] The V4 variable region of the 16S rRNA gene was amplified by PCR and sequenced on an Illumina Genome Analyzer IIx. In the original paper of Caporaso et al. (2011)[46] the raw sequences were clustered in Operational Taxonomic Units (OTU) using the Quantitative Insights Into Microbial Ecology (QIIME) pipeline. However, recent studies have shown that the use of OTUs is more prone to produce noisy features which are artifacts of sequencing errors.[208] Nowadays, the use of Amplicon Sequence Variants (ASV) data has been shown to be more reliable than OTU's.[208]

Following the same line, here we used the ASV gut microbiota data of Caporaso et al. (2011) which is available at the Earth Microbiota Project (EMP) platform (earthmicrobiome.org).[46] The ASV data provided at the EMP platform have been generated from the raw sequence data with the Deblur pipeline[311] and the detailed protocol is provided in Thompson et al. (2017).[312] The data for human microbiota time series was obtained from 'emp_deblur_150bp.release1.biom' by filtering to keep only samples from the Qiita study ID 5501.

We removed singletons and ASV sequences assigned to mitochondria and chloroplasts. We assembled the taxa at the genus level and this yielded 578 unique genera. For both the male and female subject, we first removed samples with less than 500 reads, then we transformed the time series to relative abundances and then we made a selection of genera, using a bootstrapping method[313] with a prevalence value of 25% and a relative abundance threshold value of 0.005 (i.e., select the genera in which the relative abundance has a value higher than 0.005 in at least 25% of the samples). We disregarded the taxa that were not identified as uniquely defined genera. This yielded a total of 19 genera for the male subject and of 12 genera for the female subject. The aim of our analysis is to compare clusters (and techniques to obtain these clusters) among the two different subjects. Therefore, we considered in our analysis the genera that were present in at least one subject, yielding a total of 19 genera for each subject. Processing of the data from ASV to the core-microbiota taxa was done using the 'phyloseq'[314] and 'microbiota'[313] R packages. Subsequently, we applied a centered log-ratio (CLR) transformation to the relative abundance time series using the 'compositions' R package.[315] The CLR transformed time series of the selected genera are shown in Figure 3.2. Wavelet analysis requires equidistance between subsequent datapoints, therefore we interpolated the time series of both subjects using cubic Hermite interpolation to obtain data with equidistant time intervals of 1.6 days (the mean time interval of the original data of the male subject is 1.6 days and the female subject is 1.5 days), yielding a total of 336 data points for the male subject and of 131 data points for the female subject.

Before performing wavelet analysis to the data, the microbiota CLR transformed time series were rescaled using a Box-Cox transformation to suppress sharp peaks, homogenize the variance and approximate a normal distribution. For each time series the optimal parameter of the Box-Cox transformation has been estimated by optimizing the normal probability plot correlation coefficient using the 'EnvStats' R package (see Appendix Figure 3.1).[316]

Results

Wavelet cluster analysis

Wavelet analysis enables investigation of time series characterized by different periodicities and is particularly suited for time series which are not stationary, as applies to many biological systems. We first illustrate this technique using synthetic time series (Figure 3.1A left hand side). Consider for instance time series 1 and 2: they are stationary and oscillate at the same periodicity of eight days, but in antiphase. They are therefore characterized by the same wavelet spectrum: a significant period of eight days (orange area inside the black dotted line) occurring along the entire time span of 100 days. The average wavelet spectrum, which is an estimation of the classical Fourier spectrum, is also identical among the two time series (see plot at the far most right-hand side). If one considers time series 7 and 8, one may see that they are showing opposite patterns. Time series 7 oscillates fast at a periodicity of about four days in the first 50 days and then slows down and oscillates at a periodicity of about 20 days in the second half of the time series. Time series 8 is doing exactly the opposite, it oscillates slowly with a periodicity of about 20 days in the first half of the time series and then oscillates with a periodicity of about four days in the second half of the time series. While the average wavelet spectrum is identical for both time series, the wavelet spectra are showing opposite patterns and are therefore able to depict the differences between the temporal behaviour in the oscillations of the two time series (Figure 3.1A).

The wavelet spectra are then compared using a procedure based on maximum covariance analysis which enables construction of a distance matrix based on the wavelet power spectra.[297] The constructed distance matrix is used to build a cluster tree based on the WARD agglomeration criterion (Figure 3.1B).[317] For comparison, we also constructed a Spearman dissimilarity matrix calculated as $d = 1 - \rho$ (where ρ is the correlation coefficient), using all data points in the time series pairs. The Spearman dissimilarity matrix is also used to construct a cluster tree based on the WARD agglomeration criterion (Figure 3.1C). We compare the wavelet clustering with a clustering based on Spearman's correlation, because the latter is a common method used in microbiota studies to infer relationships between microorganisms.[253] One may immediately observe substantial differences between the trees obtained with the two different methods (Figure 3.1B and 3.1C). The time series are clustered differently within the trees according to the two methods, but also branching structure and the total length of the branches is noticeably different.

Time series 1 and 2 are close together in the wavelet cluster tree (Figure 3.1B), but they fall apart in the Spearman cluster tree (Figure 3.1C). The first results from the fact that the two time series have identical wavelet spectra, which indicates that the time series oscillate at the same periodicity. However, they are considered dissimilar in correlation analysis, because the time series are in antiphase (i.e., the peaks of one time series coincide with the troughs of the other time series and vice versa).

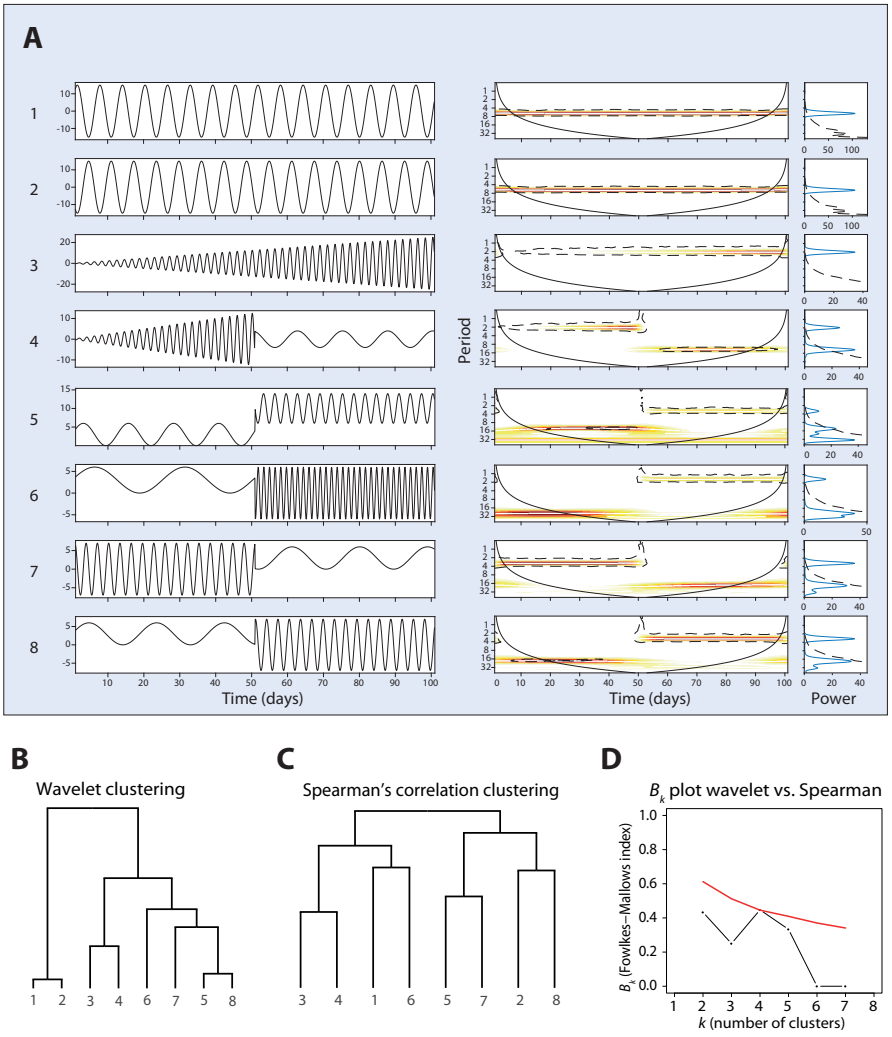


Figure 3.1 - Illustration of wavelet clustering analysis with synthetic time series. A) Wavelet analysis of synthetic time series: synthetic time series (left hand side) characterized by different periodicities; wavelet spectra (right hand side) and average wavelet spectra (far right) of the synthetic time series. Color codes represent wavelet power and range from low (white) to high (red). Black dotted lines enclose the 5% significance areas computed using a Markov surrogate significance test. The solid black line delimits the cone of influence, where edge effects become important. Clustering of the synthetic time series based on two methods. In B), clustering is based on the wavelet spectra. The cluster tree is constructed by grouping the time-frequency patterns of the time series using maximum covariance analysis. In C), clustering is based on Spearman's correlations calculated for each pair of time series. The correlations are used to compute the dissimilarity matrix which is used to cluster the data. For both methods the hierarchical clustering of the time series is performed using the WARD agglomeration criterion. D) Comparison of the hierarchical clusters obtained using the B_k statistic.[308] Black dots represent the B_k values plotted against the k number of clusters in which each tree has been partitioned. Red line represents the one-sided rejection region based on the asymptotic distribution of B_k values, for each k, under the null hypothesis of no relation between the clusters (significance $\alpha = 5\%$).

Similarly, time series 5 and 8 cluster together in the wavelet tree but they fall apart in the Spearman cluster tree. Both time series 5 and 8 oscillate slowly at a periodicity of about 13 and 20 days, respectively, in the first part of the time series but then oscillate faster (at a periodicity of about four days) in the second part of the time series. Therefore, their wavelet spectra are very similar.

If the synthetic time series would represent the dynamical behaviour of microorganisms, one would conclude from the Spearman cluster tree that microorganisms 1 and 2 (or 5 and 8) are not or only weakly related, because when one microorganism is highly abundant then the other one has very low abundance (and the other way around). The wavelet clustering instead shows that these microorganisms are strongly connected because they oscillate with similar periodicities and therefore share the same dynamical properties, which may point to ecological interdependence e.g., through parasitic interactions or neutral niche competition.

In addition to visual inspection, we used the B_k statistic to quantify the similarity in cluster trees constructed with the two methods.[308] The B_k statistic assesses the chance-corrected proportion of items that two cluster trees have in common, as a function of the number of subclusters k that the two trees are partitioned into. Plotting B_k versus k gives a quantitative representation of the similarity between two cluster trees (black dots in Figure 3.1D). The red line represents the 95% rejection region under the null hypothesis of no relation between the trees. For all partitions k, the blacks dots fall below the red line, hence we cannot conclude that the trees calculated with the wavelets and the Spearman's correlations for the synthetic time series are significantly related.

In Box 3.1 we give an additional demonstration of wavelet clustering analysis applied to the outputs of an ecological model of four consumers and four resources. In this case, wavelet clustering accurately captures the competitive coupled dynamics between consumers and resources, whereas clustering based on Spearman's correlation does not (Figure 3.A - panels D and E in Box 3.1).

Application to human microbiota data

We tested our approach, as illustrated for the synthetic time series, on real data of microbiota communities. We used previously published gut microbiota time series of two healthy subjects, one male and one female, from whom fecal samples had been collected for 15 and 6 months, respectively.[46] We considered the data at genus level and we selected the same 19 genera for the male and the female subject. A detailed description of the data and of the selection criterion is provided in the methods.

Time series of the selected genera for the male and the female subject are shown in Figure 3.2. CLR transformed relative abundances over time show remarkable fluctuations. Some genera (e.g., *Lachnospira* and *Roseburia* in the male subject; *Bacteroides* in both subjects) show a clear wax and wane in their dynamical pattern. There are other genera (e.g., *Campylobacter* and *Finegoldia* in the female subject) that show more spiky dynamics, dominated by low CLR transformed relative abundances, but with few very high peaks.

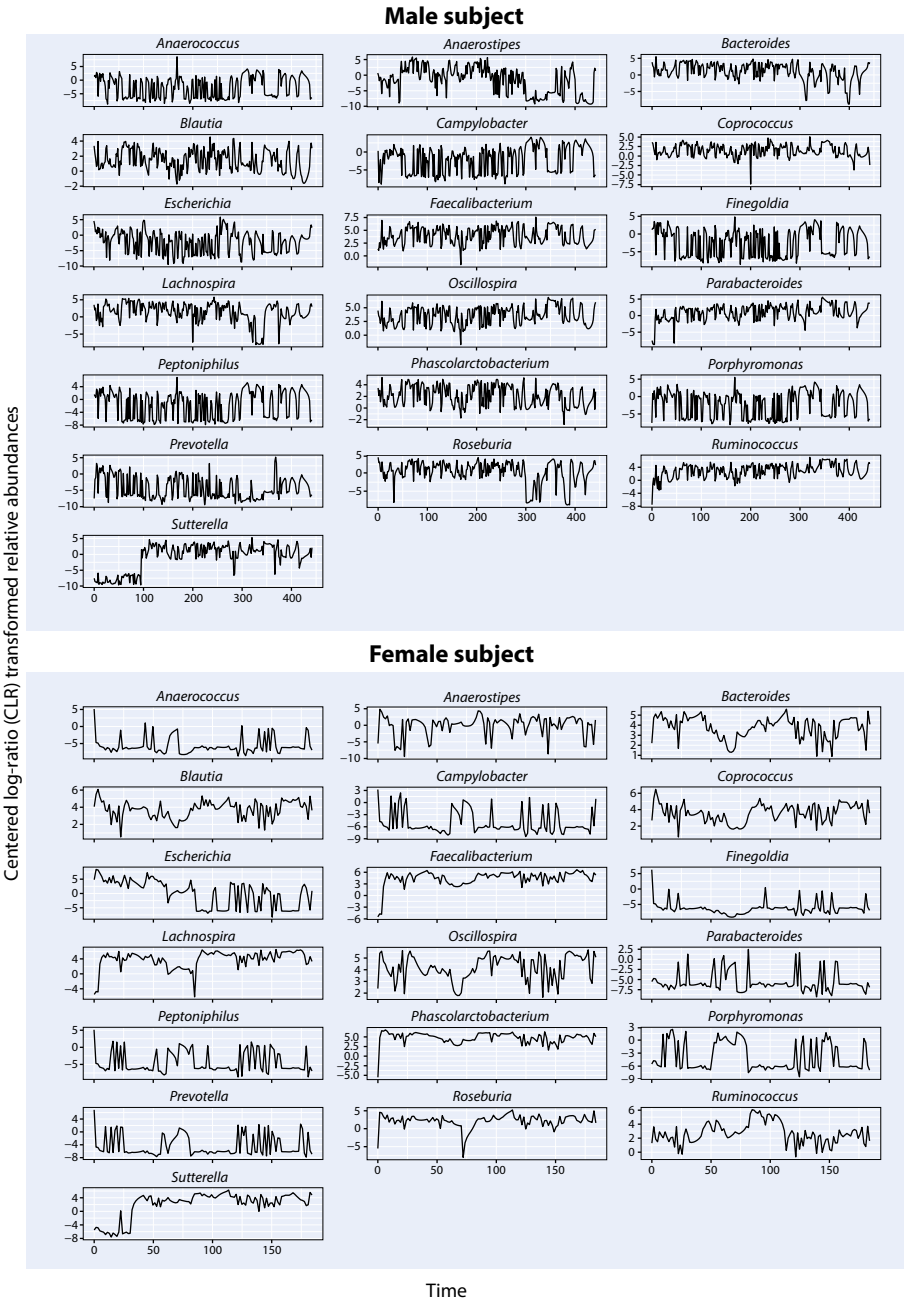


Figure 3.2 - Gut microbiota time series of CLR transformed relative abundances for selected genera. Male (upper graphs) and female (lower graphs) subject. The time series show clear fluctuations. Note the distinct time axes in the male and the female subject.

Box 3.1 - Wavelet clustering applied to the dynamics of four consumers feeding on four resources. In this section we give an extra demonstration of the potential of wavelet clustering by performing the analysis on the outputs of a simplified ecological model describing the dynamics of four consumers and four resources. The model is a modified version of the previously published model of Vandermeer of two species feeding on two resources.[318-320]

The model reads as follows:

$$\frac{dC_i}{dt} = \frac{aR_iC_i}{1+bR_i} - mC_i \qquad \textbf{Eq. 3.A}$$

$$\frac{dR_i}{dt} = r_iR_i\left(\frac{K - R_i - \alpha_{ij}R_j}{K}\right) - aR_i\left(\frac{C_i}{1+bR_i}\right) \qquad \textbf{Eq. 3.B}$$

$$\frac{dC_k}{dt} = \frac{aR_kC_k}{1+bR_k} - mC_k \qquad \textbf{Eq. 3.C}$$

$$\frac{dR_k}{dt} = r_kR_k\left(\frac{K - R_k - \alpha_{kl}R_l}{K}\right) - aR_k\left(\frac{C_k}{1+bR_k}\right) \qquad \textbf{Eq. 3.D}$$

for $i = 1, 2$ and $k = 3, 4$ and $i \neq j$ and $k \neq l$, where C_i and C_k are the abundances or densities of the i^{th} and the k^{th} consumers, respectively, and R_i and R_k denote those of the i^{th} and the k^{th} resources. The parameters r_i and r_k represent the intrinsic growth rates of the i^{th} and the k^{th} resource, respectively. m is the mortality rate of the consumers, α_{ij} is the competition coefficient between resource 1 and 2, α_{kl} is the competition coefficient between resource 3 and 4, a is the resource consumption rate, b is the functional response parameter (with higher values denoting diminished response in consumer growth at a given resource abundance), and K is the carrying capacity of each resource, which we assume for simplicity to be the same for all four resources.

The model consists of two separated food webs of two consumers each feeding on one resource (Figure 3.A - panel A). Consumer C_1 feeds on resource R_1, consumer C_2 feeds on resource R_2 and the two resources R_1 and R_2 negatively interact with a parameter α_{12}. Similarly, consumer C_3 feeds on resource R_3, consumer C_4 feeds on resource R_4, and the two resources R_3 and R_4 negatively interact with a parameter α_{34}. In Figure 3.A (panel B left hand site) are shown the temporal dynamics of the four consumers and the four resources. We applied wavelet analysis to all eight of the time series (Figure 3.A - panel B right hand side) and we used this information to build the cluster tree (Figure 3.A - panel C). Wavelet clustering identifies two big subclusters: subcluster 1 with consumers C_3 and C_4 and resources R_3 and R_4, and subcluster 2 with consumers C_1 and C_2 and resources R_1 and R_2. Wavelet clustering successfully identifies the two separated food webs. In addition, inside each cluster we observe that each consumer is clustered together with its own resource (C_1 with R_1, C_2 with R_2, C_3 with R_3, and C_4 with R_4). For comparison we build a tree based on Spearman's correlation (Figure 3.A - panel D). In contrast to wavelet clustering, clustering based on Spearman's correlation is not able to identify neither the two distinct food webs, neither the pairs of consumers-resources. Clustering based on Spearman's correlation is substantially different from clustering based on wavelets as it is shown by the corresponding B_k plot (Figure 3.A - panel E).

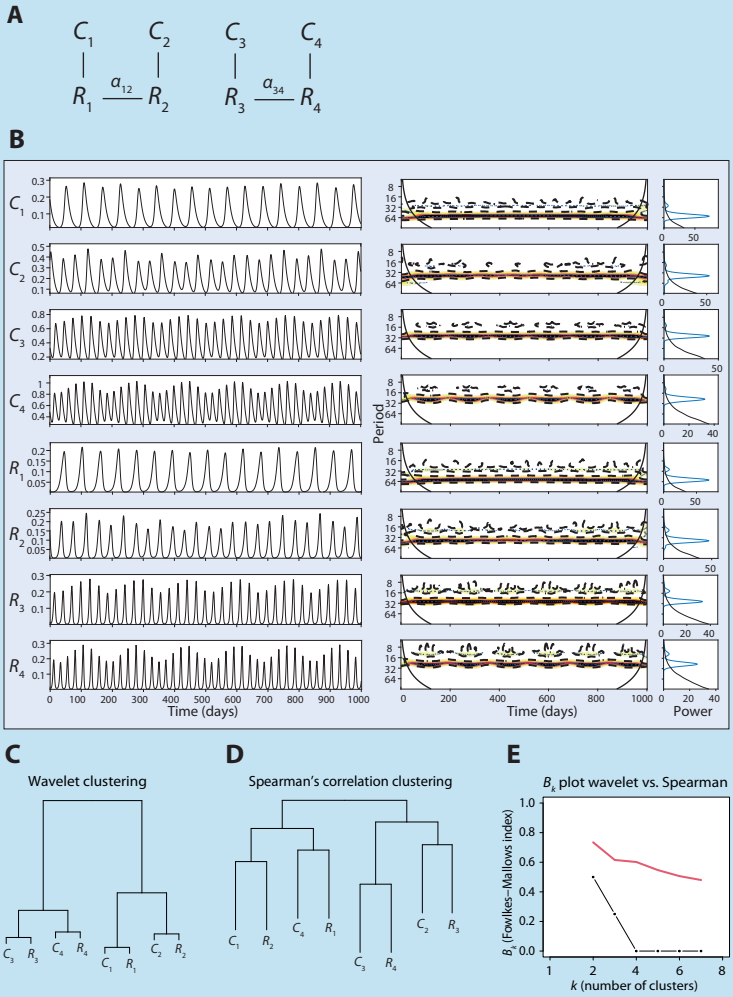


Figure 3.A - Application of wavelet clustering to the outputs of a model with four consumers feeding on four resources. A) The model consists of two separated food webs of two consumers on two resources. The two resources within each food web negatively interact with a competition coefficient a. B) (Left) Outputs of the resources-consumers model. Simulations have been run for 2000 time units. The plots shown here covers the last 1000 time units of the simulation. Parameters: $K = 1$; $a = 2$; $b = 1.3$; $m = 0.1$; $r_1 = 0.2$; $r_2 = 0.4$; $r_3 = 0.8$; $r_4 = 1.2$; $a_{12} = 0.8$; $a_{34} = 0.4$; (Right) Wavelet spectra and average wavelet spectra (far right) of the model outputs. Color codes represent wavelet power and range from low (white) to high (red). Black dotted lines enclose the 5% significance areas computed using a Markov surrogate significance test. The solid black line delimits the cone of influence, where edge effects become important. C) Clustering based on the wavelet spectra. The cluster tree is constructed by grouping the time-frequency patterns of the time series using maximum covariance analysis. D) Clustering based on Spearman's correlations calculated for each pair of time series. The correlations are used to compute the dissimilarity matrix which is used to cluster the data. For both methods the hierarchical clustering of the time series is performed using the WARD agglomeration criterion. E) Comparison of the hierarchical clusters obtained using the B_k statistics. Black dots represent the B_k values plotted against the k number of clusters in which each tree has been partitioned. Red line represents the one-sided rejection region based on the asymptotic distribution of B_k values, for each k, under the null hypothesis of no relation between the clusters (significance $\alpha = 5\%$).

To capture possible similarities in the dynamical patterns of the bacteria, we applied wavelet analysis to each of the bacterial time series in both subjects. Wavelet spectra detected several significant periodicities in the fluctuations of bacteria both for the male (Figure 3.3) and the female subject (Figure 3.4). A first visual inspection of the spectra already reveals similarities between the dynamical patterns of the bacteria. For instance, in the male subject (Figure 3.3), *Porphyromonas*, *Phascolarctobacterium*, and *Peptoniphilus* show common periodicities of about 30–40 days co-occurring for approximately 100 days at the end of the time series. In addition, *Campylobacter* and *Roseburia* clearly show common periodicities of 64 days occurring approximately in the last 150 days of the time series, whereas *Blautia* and *Coprococcus* share this periodicity at the beginning of the time series. Common patterns are less clear in the female subject (Figure 3.4), though some similar periodicities can be identified. For instance, many genera show the same periodicity of about 60 days occurring along the entire length of the time series.

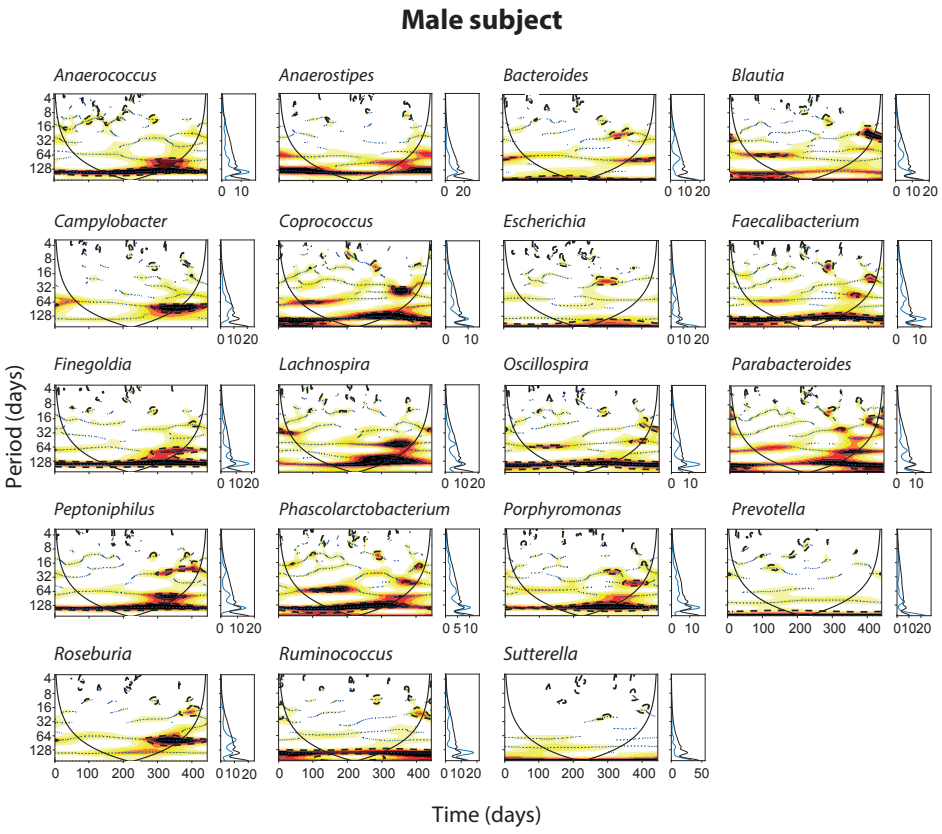


Figure 3.3 - Wavelet analysis of time series for selected genera in the male subject. For each genus the wavelet spectrum (left) and the average wavelet spectrum (right) are computed. Color codes represent wavelet power and range from low (white) to high (red). Black dotted lines enclose the 5% significance areas computed using a Markov surrogate significance test. The solid black line delimits the cone of influence, where edge effects become important.

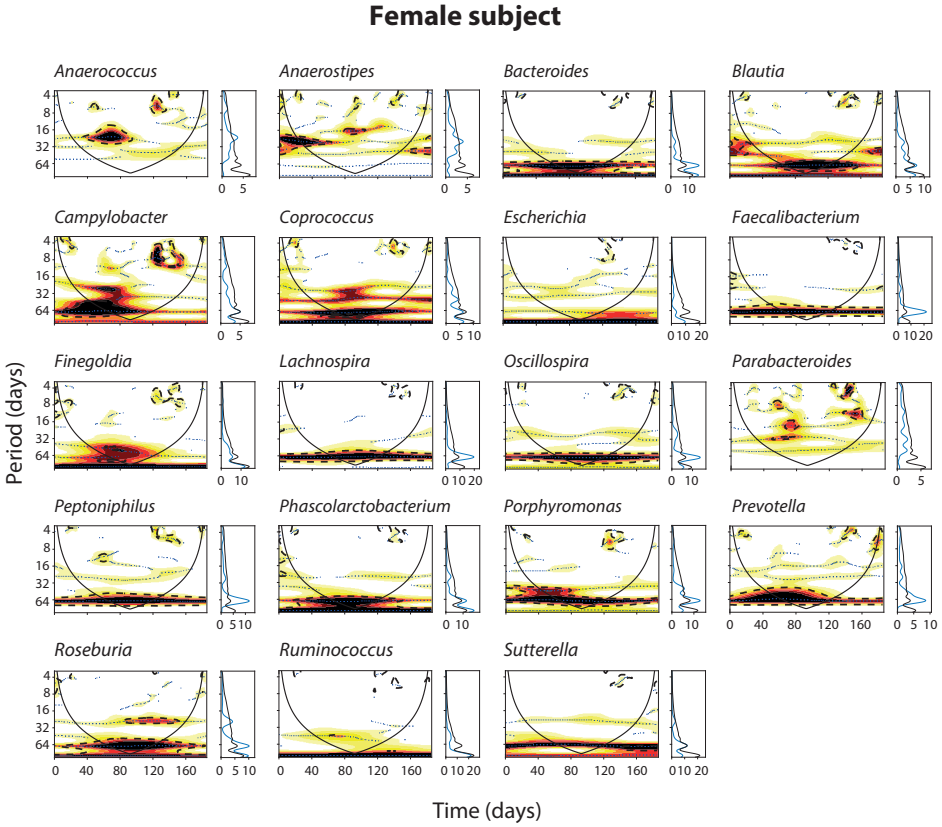


Figure 3.4 - Wavelet analysis of time series for selected genera in the female subject. For each genus the wavelet spectrum (left) and the average wavelet spectrum (right) are computed. Color codes represent wavelet power and range from low (white) to high (red). Black dotted lines enclose the 5% significance areas computed using a Markov surrogate significance test. The solid black line delimits the cone of influence, where edge effects become important.

With the wavelet spectra at hand, we built trees based on the wavelet distance matrix as described for the synthetic time series. Both the clusters based on wavelet spectra for the male and the female subject show a clear partition in two subgroups (Figure 3.5A and 3.5B). The clusters based on Spearman's correlations for the male and the female subjects are also characterized by two main subclusters (Figure 3.5C and 3.5D). Although there are few bacteria that are clustered together with both methods (i.e., *Peptoniphilus*, *Finegoldia*, *Porphyromonas*, and *Anaerococcus* in the male subject), the two methods yield very different clusters. For instance, *Bacteroides* and *Prevotella* are clustered together in the male subject with the wavelet method, but they are in two different clusters in the male subject with the correlation method. The case of *Prevotella* and *Bacteroides* resembles the example of signals 1 and 2 (or 5 and 8) illustrated before: two time series with similar dynamical properties are clustered together based on wavelets but are considered not related by the correlation method.

Also, visual comparison of the clusters obtained using wavelets (Figure 3.5A and 3.5B) with the clusters obtained by pairwise correlations (Figure 3.5C and 3.5D) reveals substantial differences between the two methods in the positioning of branches within the two subclusters and in the total length of the branches.

Of note, total branch length (see 'Methods') is substantially higher in the wavelet cluster tree as compared to the tree based on Spearman's correlations (male subject: 80.9 vs. 27.6; female subject: 70.0 vs. 21.9). Further visual comparison of the trees based on wavelets among the two subjects also reveals that the members of each subcluster are substantially different between the male and the female subject (compare Figure 3.5A with Figure 3.5B). In contrast, comparison of the cluster trees based on correlations shows that many bacteria that are clustered together in the male subject are also clustered together in the female subject (compare Figure 3.5C with Figure 3.5D).

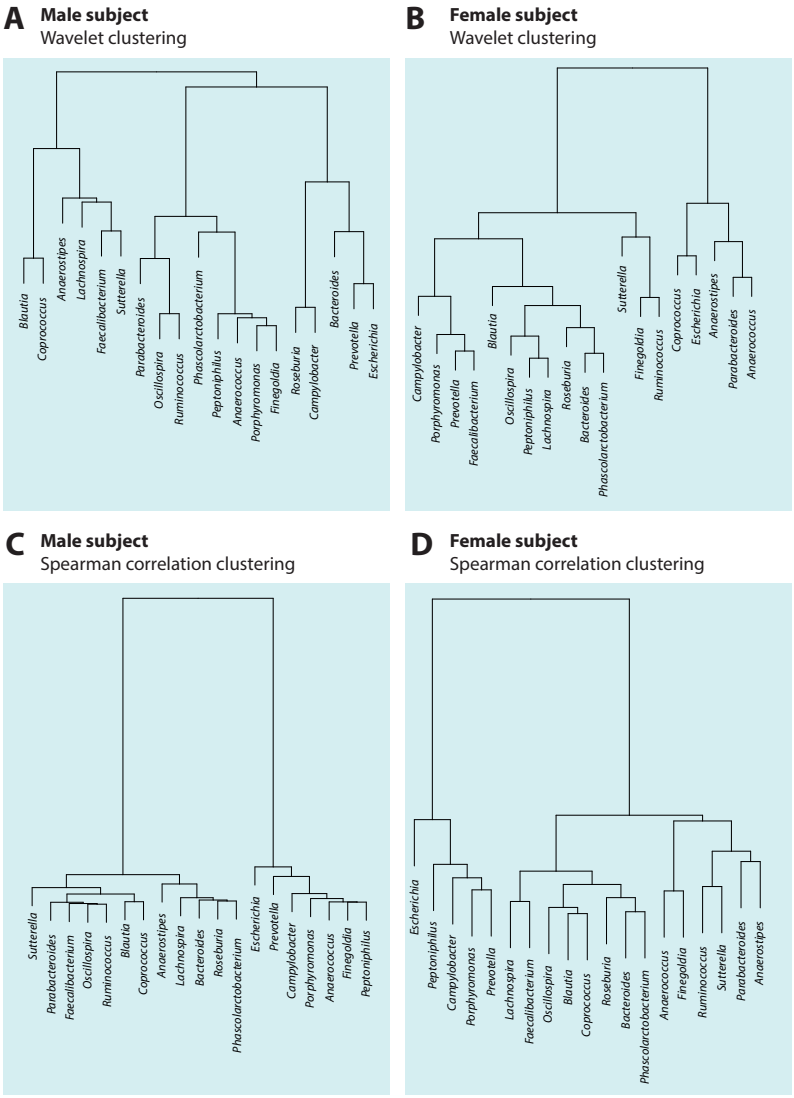


Figure 3.5 - Clustering for the male and female subjects based on different methods. Cluster tree obtained using the dissimilarity matrix obtained from the wavelet clustering analysis for A) the male subject and B) the female subject. Cluster tree obtained using the dissimilarity matrix obtained from the Spearman's correlation matrix for C) the male subject and D) the female subject.

To further quantify the similarities between subjects and methods we calculated the B_k statistic as we did for the synthetic time series. For low values of k, the dots in Figure 3.6A and 3.6C fall below the 95% rejection line. Thus, wavelet clustering and Spearman's clustering are not significantly related when the community is partitioned into a limited number of subclusters, and this holds for both the male and female subject. This is likely because the wavelet clustering method accounts for other features (i.e., the spectral characteristics of the bacterial dynamics and their time evolution) than the correlation-based methods, which only consider quantities averaged over the whole series. For higher values of k, the dots sometimes fall above the rejection line (Figure 3.6A and 3.6C), meaning that wavelet clustering and Spearman's clustering get significantly related at some higher resolution when certain subclusters become apparent. For comparison (Figure 3.6B and 3.6D) we also applied the B_k statistic to correlation-based trees constructed with the Spearman's correlation and with the Pearson's correlation coefficient (trees not shown). For all k partitions (except the maximum partition for the male subject), the trees calculated with these two correlation methods are instead, as it could be expected, significantly related.

Finally, we also assessed the similarity between the two subjects. Interestingly, we found no evidence for related wavelet clusters between the male and female subjects, as all dots fall below the 95% rejection line irrespective the number of k partitions (Figure 3.6E). In contrast, in the B_k plot of the Spearman's correlation-based clustering, the majority of dots fall above the 95% rejection line (Figure 3.6F), indicating significantly related clusters for almost all subpartitions between the male and female subject. This suggests that wavelet clustering not only uncovers more diverse community structures within individuals, but might also be more sensitive towards subtle differences in community structures across individuals.

Discussion

Developments in high-throughput sequencing have improved our ability to track the temporal variability of microbial communities. This has led to an increase in longitudinal data from a variety of different microbiota ranging from wastewater,[321] marine,[322-324] freshwater,[325] and terrestrial[326, 327] environments. These time series offer unprecedented opportunities to gain ecological insights into microbial community dynamics and the mechanisms governing them, and to track the response of the microbial systems to external perturbations.

Ideally, long time series are required to capture the periodic patterns of microbial dynamics and reveal community structures. Unfortunately, only few of such datasets exist in human microbiota studies.[46, 59, 328, 329] This probably reflects the relative difficulty to repeatedly sample the human microbiota in comparison to a natural field habitat (e.g., sampling strongly relies on the consent of the host to provide sampling material at a regular basis). As a result, the majority of studies on human microbial community structures have relied on sparse data and methods based on co-occurrence, which may have produced biased associations, e.g., towards positive correlations.[253, 268, 269] Clearly, there is a need to shift from a static to a dynamical approach, that takes into account the temporal development of bacterial communities and can shed new light on microbial community structure.[50] This also has bearing on the ability to employ microbiota data for clinical practice, as more and more studies move from association to prediction of disease course, e.g., exacerbation of inflammatory bowel disease (IBD),[330] and treatment response in *Clostridioides difficile* infection.[190]

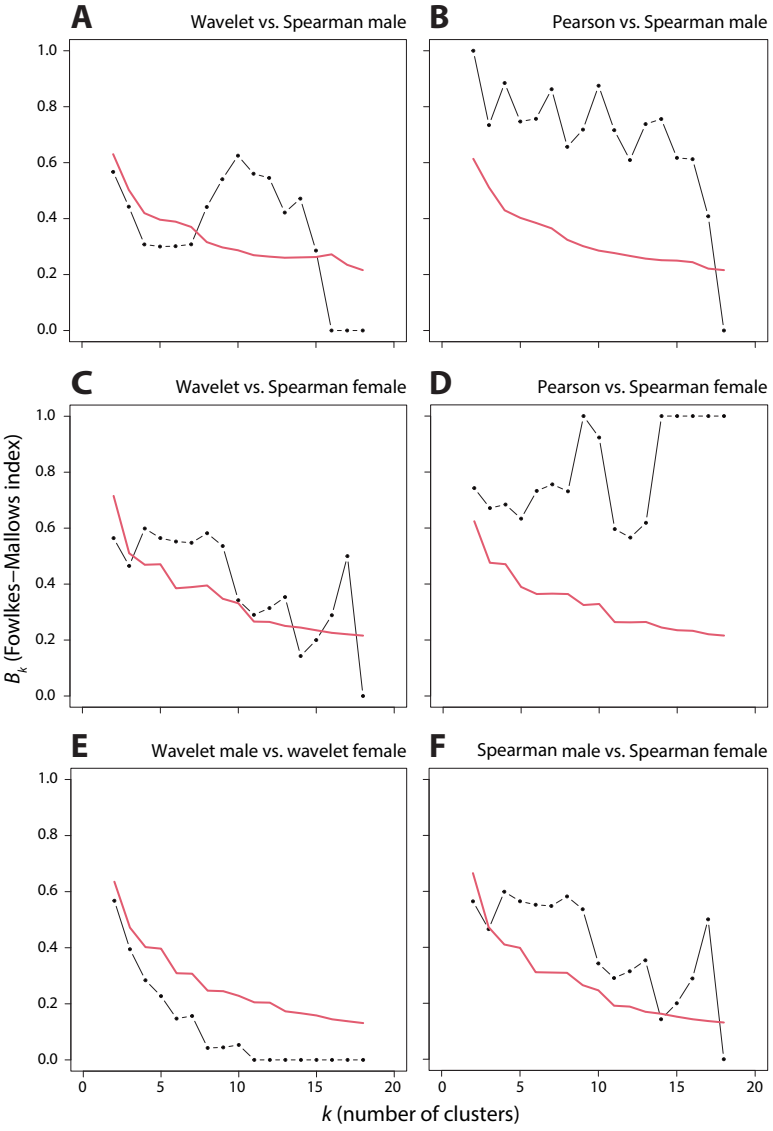


Figure 3.6 - Comparison of hierarchical clusters using the B_k statistic.[308]
Black dots represent the B_k values plotted against the k number of clusters in which the tree has been partitioned. Red line represents the one-sided rejection region based on the asymptotic distribution of B_k values, for each k, under the null hypothesis of no relation between the clusters (significance $\alpha = 5\%$). A) Comparison of the tree based on wavelets and the tree based on Spearman's correlations for the male subject. B) Comparison of the tree based on Pearson's correlations and the tree based on Spearman's correlations for the male subject. C) Comparison of the tree based on wavelets and the tree based on Spearman's correlations for the female subject. D) Comparison of the tree based on Pearson's correlations and the tree based on Spearman's correlations for the female subject. E) Comparison of the trees based on wavelets for the male and female subject. F) Comparison of the trees based on Spearman's correlations for the male and female subject.

Interestingly, our reanalysis of the widely used Caporaso et al. (2011) data reveals some novel important patterns.[46] The trees obtained with the two different methods show significant differences in the way microbial genera are clustered together. For instance, there are cases where pairs of bacteria are clustered together in the male and female subject when using correlations, but not when using wavelets. For example, according to wavelet analysis *Blautia* and *Coprococcus* only cluster together in the male subject, and *Phascolarctobacterium*, *Roseburia*, and *Bacteroides* only in the female subject, whereas these genera are clustered together in both subjects with the correlation-based method. In general, similarity of the cluster trees between subjects seems to be stronger with the correlation-based method than with wavelet clustering, for which we found no evidence for significant relations between the male and female trees. Tree correspondence according to clustering method within subjects was more ambiguous, as similarity also depends on tree resolution. This emphasizes how sensitive the clustering is to the type of method chosen.

In addition, we also note differences in the pattern of branching and in the total branch length of the cluster trees. Studies have shown that the total length of the branches in a traits tree is indicative of the functional diversity in ecosystems.[331] Analogously, total branch length can here be considered as an indicator of the diversity of community structure. While we are not considering functional traits here, we could speculate that the higher total length observed in the wavelet clustering of the microbiota time series is indicative of a higher diversity in community structure as compared to the correlation-based method. A likely explanation is that wavelet analysis is able to detect dependencies that are not apparent in correlations, whereas the reverse is not the case: highly correlated time series are still detectable in wavelet spectra. Thus, wavelet clustering can extract more information on the dependencies within microbial communities than is reflected in mere correlations.

Looking at the clusters identified by the wavelet method one can speculate about possible interaction mechanisms between the bacteria. For instance, in the male subject, two genera are observed together, *Blautia* and *Coprococcus*. Members of genus *Blautia* are known to produce acetate and lactate which is shown to support improved growth of *Coprococcus in vitro*.[332] *Coprococcus* bacteria can convert lactate and acetate to butyrate, a short chain fatty acid that is associated with a healthy microbiota.[333] This mutualistic mechanism could potentially lead to similar dynamical patterns and explain why these bacteria co-occur in the same cluster. Although these 'potential' interaction mechanisms are based on associative dynamical patterns of 16S rRNA gene sequence data they may provide ground for further investigation of these interactions *in vitro* and *in vivo*. In addition wavelet cluster analysis can be used as a starting point for investigation for time series causality inference methods such as Granger causality[334, 335] or convergence-cross mapping.[336, 337] For instance, there are methods that are able to estimate Granger's causality from wavelet spectra of time series data.[338, 339] Application to a complex system such as the microbiota has not yet been done and can be subject of investigation in future studies.

In ecological and epidemiological studies, wavelet analysis is often used to evaluate the effect of external factors, such as climatic or meteorological variables, on species or disease dynamics. Examples include studies which evaluate the effect of external factors on the spread of dengue fever,[340] malaria,[341] and cholera,[342] or on the dynamics of communities of benthic organisms,[6] marine[343] and freshwater plankton,[344] or fish.[290, 345]

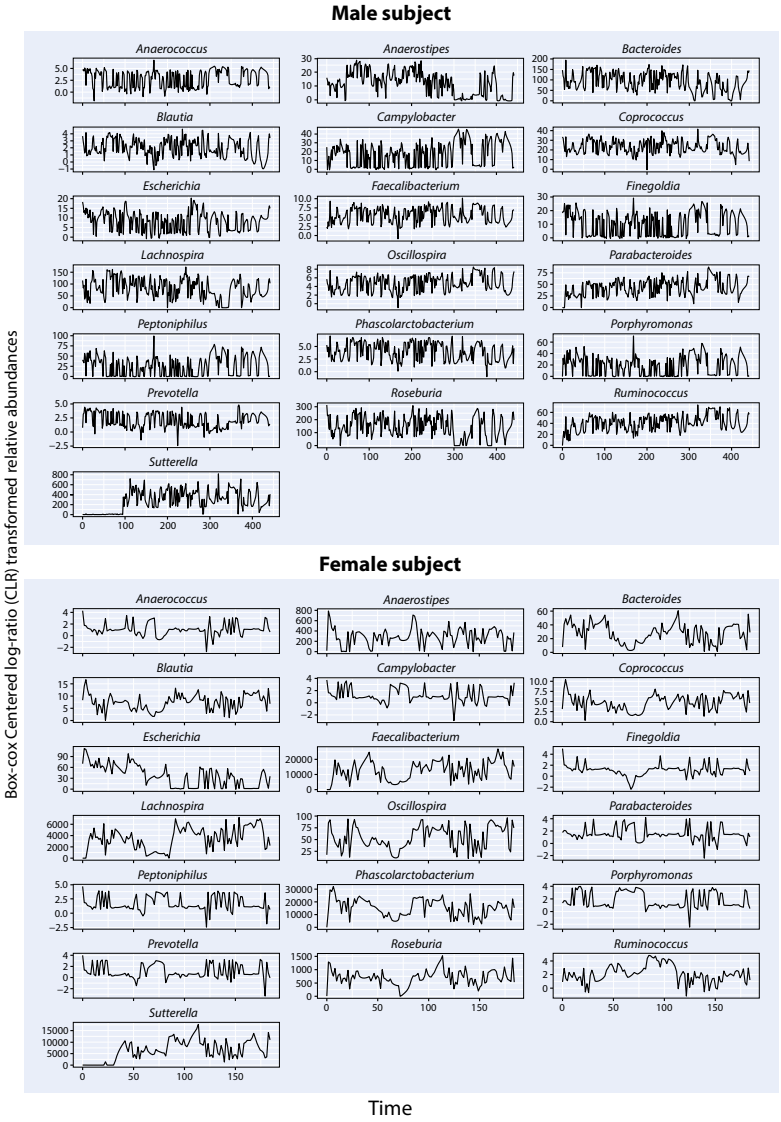
In an analogous way, when longitudinal studies on human microbiota dynamics become more widely available, metadata can be exploited using wavelet analysis to evaluate the effect of interventions, as for instance vaccination, the use of antimicrobials or probiotics, fecal microbiota transplantation, and cancer treatment.

The reader interested in using the wavelet clustering approach might wonder how many points are needed for applying such an analysis. The limits in the number of data points for wavelet analysis are similar to those of Fourier analysis and depend on the periodic components that one wants to highlight. For instance, Murdoch et al. (2002)[346] suggest that with a minimum time series length of 25 time units one can identify periodicities between two time units (the Nyquist frequency) and 8–10 time units. Cazelles et al. (2012)[300] are more conservative and they suggest time series with a minimum length of 30–40 time units which allows detection of a maximum periodicity equal to 20–25% of the total length of the time series. Another practical aspect is that wavelet analysis requires equidistant data. Although this might appear as a limiting factor, this requirement can easily be addressed. For instance, when possible, an experiment or a sampling strategy could be designed in such a way to obtain equidistant sampling points. If this is not possible, there are interpolation methods that can be used to obtain equidistant data. Different interpolation methods should be tested, and the interpolated data should be checked against the original data to see if the general dynamical behaviour is unaffected by the interpolation. This is the approach taken in this study. In addition, as for Fourier analysis, there are extensions of wavelet analysis that can be applied to non-equidistant data.[347-351]

In our study we analysed the time series of two individuals, and we compared the wavelet dendrograms of the two subjects using a pairwise metric. Ideally, new longitudinal human microbiota studies will track the joint dynamics of much more than two individuals. When time series of multiple subjects become available, one might want to compare dendrograms among classes of individuals (e.g., individuals of the same gender or patients versus healthy controls). Instead of a pairwise metric between individuals, our analysis could then be applied to consensus dendrograms between classes of individuals to assess how communities differ with respect to the condition of interest.[352]

To summarize, wavelet cluster analysis has the big advantage of accounting for non-stationary dynamics which are often preponderant in biological systems. In addition, we show that it appears to be a sensitive method for recovering microbial community structure from densely sampled microbiota time series. By taking into account the spectral features of bacterial abundance and their time evolution that are ignored in methods focusing on co-occurrence at any one time point, wavelet clustering analysis is able to extract more information on the dependencies within microbial communities, and to uncover more diverse communities within and across individuals than conventional methods. The results show that interpretation of microbial networks and communities, inferred on the basis of only a few sampling points in time, should be done with care, and be compared to alternatives.

Appendix of Chapter 3



Appendix Figure 3.1. Box-cox transformed CLR time series of selected genera in the male (upper graphs) and the female (lower graphs) subject. The relative abundance time series of both subjects have been interpolated using cubic Hermite interpolation to obtain data with equidistant time intervals of 1.6 days (the mean time interval of the original data of the male subject is 1.6 days and the female subject is 1.5 days), yielding a total of 336 data points for the male subject and of 131 data points for the female subject. Subsequently, we applied a CLR transformation to the relative abundance time series using the 'compositions' R package.[315] Before performing wavelet analysis on the data, the microbiota CLR transformed time series were rescaled using a Box-Cox transformation to suppress sharp peaks, homogenize the variance and approximate a normal distribution. For each time series the optimal parameter of the Box-Cox transformation has been estimated by optimizing the normal probability plot correlation coefficient using the 'EnvStats' R package.[316]