Microb Ecol (2011) 61:123–133
DOI 10.1007/s00248-010-9738-y

GENES AND GENOMES

Assessing Gut Microbial Diversity from Feces and Rectal Mucosa

Ana Durbán · Juan J. Abellán · Nuria Jiménez-Hernández · Marta Ponce · Julio Ponce · Teresa Sala · Giuseppe D'Auria · Amparo Latorre · Andrés Moya

Received: 22 April 2010 / Accepted: 5 August 2010 / Published online: 24 August 2010

Abstract Gut microbiota is the most complex bacterial community in the human body and its study may give important clues to the etiology of different intestinal diseases. Most studies carried out so far have used fecal samples, assuming that these samples have a similar distribution to the communities present throughout the colon. The present study was designed to test this assumption by comparing samples from the rectal mucosa and feces of nine healthy volunteers by sequencing libraries of 16S rRNA genes. At the family taxonomic level, where rarefaction curves indicate that the observed number of taxa is close to the expected one, we observe under different statistical analyses that fecal and mucosal samples cluster separately. The same is found at the level of species considering phylogenetic information. Consequently, it cannot be stated that both samples from a given individual are of similar composition. We believe that the evidence in support of this statement is strong and that it would not change by increasing the number of individuals and/or performing massive sequencing. We do not expect clinicians to stop using feces for research, but we think it is important to caution them on their potential lack of representativeness with respect to the bacterial biofilm on the rectal mucosa.

Electronic supplementary material The online version of this article (doi:10.1007/s00248-010-9738-y) contains supplementary material, which is available to authorized users.

A. Durbán · J. J. Abellán · N. Jiménez-Hernández · G. D'Auria · A. Latorre · A. Moya
Centro Superior de Investigación en Salud Pública (CSISP),
Avenida de Cataluña 21,
46020 Valencia, Spain

A. Durbán · J. J. Abellán · N. Jiménez-Hernández · G. D'Auria · A. Latorre · A. Moya (✉)
Instituto Cavanilles de Biodiversidad y Biología Evolutiva,
Universitat de València,
Apartado Postal 22085,
46071 Valencia, Spain
e-mail: andres.moya@uv.es

A. Durbán · J. J. Abellán · N. Jiménez-Hernández · G. D'Auria · A. Latorre · A. Moya
CIBER en Epidemiología y Salud Pública (CIBEResp),
Barcelona, Spain

M. Ponce · J. Ponce · T. Sala
Servicio de Medicina Digestiva, Hospital Universitario La Fe,
Avenida de Campanar 21,
46009 Valencia, Spain

M. Ponce · J. Ponce · T. Sala
Centro de Investigación Biomédica en Enfermedades Hepáticas y Digestivas (CIBER-EHD),
Barcelona, Spain

Introduction

The study of bacterial communities is currently an active topic in biomedical research. Historically, most human microbial research has focused on studying around 100 single pathogens, whereas we have limited knowledge of the 2,000 or so bacterial species that are beneficial to the human host [33]. This picture is changing rapidly with the advent of metagenomics, which has enabled the analysis of distribution and gene composition of bacterial communities [33].

Gut microbiota is arguably the most complex bacterial community in the human body [17] and probably one of the most complex natural environments [21]. It plays an important role in human well-being because of its contribution to nutrition, development of immune system and colonization resistance, to name a few [10, 11, 13, 17, 23]. Our knowledge of bacterial diversity in the human gastrointestinal (GI) tract has mainly been obtained by studying the sequence variability of the 16S rRNA genes, principally employing fingerprinting

techniques, fluorescent in situ hybridization, quantitative PCR, microarrays and sequencing of 16S rRNA gene amplicons. Such studies have been carried out on both healthy people [8] and patients affected by different disorders for the purpose of assessing the implication of imbalances in gut microbial composition in the etiology of diseases such as inflammatory bowel disease [27] or obesity [30, 31], as well as changes in composition in response to therapeutic treatment [22]. In summary, studies on this subject have shown that, firstly, there are two predominant bacterial phyla in the human GI tract, *Bacteroidetes* and *Firmicutes*; secondly, that there is substantial variation in the species composition and distribution between individuals [8, 31]; and finally, in spite of this variation, the composition of gut microbiota seems to correlate with specific disorders.

Most studies of gut microbiota are based on the analysis of fecal samples because they are easily collected in a non-invasive manner. However, fecal communities may not accurately represent the bacterial communities living in the GI tract, which nevertheless seem to be similar in the mucosal fraction along the colon [8, 14, 32].

Fecal bacteria could be a mixture of luminal and shed or poorly adhered mucosal bacteria. Moreover, inadequate storage of fecal samples can lead to alterations in fecal microbial composition [19]. For instance, a delay of several hours between the collection of fecal samples and their adequate storage is quite common. This may affect fecal microbial composition due to its dynamic nature, which depends on growing conditions such as nutrient availability and oxygen concentration. These two conditions, in particular, change dramatically after evacuation, potentially leading to alterations in community composition and function due to differential bacterial death or growth.

Unlike feces, colon mucosal biopsies provide samples collected directly from the GI tract and thus seem a more suitable option for the study of microbiota-related gut pathologies or treatments. Furthermore, biopsies can be extracted under controlled conditions and preserved immediately by freezing. On the other hand, there are some important drawbacks to using biopsy samples. The main one is that they must be collected by endoscopy, an invasive procedure that cannot be used routinely. Moreover, the endoscopic procedure is usually carried out after a bowel cleansing, which can have an impact on the mucosal bacterial community. Finally, biopsy samples of an individual may pose some methodological problems in techniques such as metagenomics or metatranscriptomics because there might not be enough material to work with.

Several studies have dealt with differences between fecal and colon mucosal samples [2, 4, 8, 14, 19, 20, 32]. Most of them used fingerprinting techniques and showed differences between the two types of sample. Despite providing a rapid method for the comparison and monitoring of microbial ecosystems, diversity profiles generated by fingerprinting techniques only recover the most dominant bacteria in the sampled communities, and sequencing is still necessary for identification of the community members. Only Eckburg and co-workers employed massive sequencing of 16S rRNA gene amplicons for the comparison of mucosal and fecal samples of three individuals [8]. They detected differences between the two sample types, but feces were collected 1 month after the intestinal biopsies, a lag that might have introduced changes in the composition of the microbiota.

From a clinical point of view, it is necessary to more accurately determine to what extent fecal microbial communities actually represent the bacterial communities in the gut. There is a critical question in this respect: how can we assess whether the gut microbiota is involved in the etiology of a particular disease when the mucosal microbiota has not actually been observed? Or put it in other terms, how reliable are the results obtained based on fecal samples? A first step is to assess how well fecal microbiota represents the intestinal one. To this end, it is essential that both samples are collected at the same time. Also, the bowel cleansing prior to colonoscopy may introduce perturbations in the composition of the mucosal community. We address both issues in the present study.

The objectives of the present work are: (1) to analyze the variability in the composition of bacterial communities between healthy individuals; (2) to analyze the within-subject variability of the bacterial composition of feces and colon mucosal biopsies; and finally (3) to measure the extent to which fecal microbial composition serves as a predictor of gut microbial composition. To this end, wide PCR-amplified 16S rRNA gene libraries were obtained from rectal biopsies and fecal samples of nine healthy volunteers and analyzed by different statistical and phylogenetic methods.

Methods

Sample Collection

Samples were rectal biopsies and fecal samples of nine healthy volunteers (subjects without intestinal organic disorders or systemic comorbidities). All subjects gave prior informed written consent to the study protocol, which was approved by the Ethics Committee of the Hospital Universitario La Fe (Valencia, Spain). Volunteers were administered a questionnaire face to face about lifestyle and relevant clinical features. None had a history of gastrointestinal disease, recent (in the last 3 months) treatment with antibiotics (except one, who had taken antibiotics the previous month), immunomodulating therapy, anti-diarrheal medication or laxatives. Relevant volunteer details are summarized in Table 1.

Four random biopsies were obtained from rectal mucosa by rectoscopy using a standard colonoscopy (Olympus) and single-use biopsy forceps (Radial Jaw™ 4, Boston Scientific). Neither laxatives nor enema were administered prior to endoscopy to avoid the potential disturbance of mucosal microbiota associated with this procedure. Biopsies were recovered in dry tubes, preserved on ice and immediately frozen at −80°C. Endoscopically, the rectal mucosa appeared normal in all volunteers.

Feces were self collected by the volunteers with the shortest possible time lapse to the biopsies in order to minimize potential temporal changes in community composition. Initially all fecal samples were collected the same day as rectoscopy (prior to rectoscopy). However, for four of the volunteers, the fecal sample was so tiny that it did not provide enough DNA for the study and eventually a second sample was obtained between 2 and 8 weeks after rectoscopy (see Table 1 for details). Fecal samples were recovered in tubes containing 10 mL of phosphate-buffered saline (PBS; containing, per liter, 8 g of NaCl, 0.2 g of KCl, 1.44 g of Na_2HPO_4, and 0.24 g of KH_2PO_4 [pH 7.2]) and stored in the volunteers' home freezers until its release to health service staff.

All samples were stored at −80°C until further processing.

DNA Extraction

The four biopsies of each individual were pooled together. DNA was extracted from biopsies using the QIAamp DNA Mini Kit (QIAGEN) and its protocol for DNA purification from tissues. The standard protocol was modified to maintain incubation at 56°C in buffer ATL and proteinase K overnight and to extend the incubation from 10 to 30 min at 70°C with RNase A (100 mg/mL).

DNA was extracted from fecal samples using the QIAamp DNA Stool Mini Kit (QIAGEN) and its protocol for isolation of DNA for pathogen detection. Before DNA extraction, fecal samples were resuspended in PBS and centrifuged at 4,000 rpm for 8 min to remove fecal debris as far as possible. Between 1 and 4 mL of the supernatants were centrifuged at 14,000 rpm for 5 min and pellets were resuspended in 2 mL of buffer ASL. Then, we went on to step 3 of protocol.

DNA extractions were stored at −20°C.

Bacterial 16S rRNA Gene Amplification

The 16S rRNA genes were amplified by polymerase chain reaction (PCR) using the universal primers 8F (5′-AGAGTTTGATCMTG GCTCAG-3′) and 1510R (5′-TACG-GYTACCTTGTTAC GACTT-3′) [1]. Each PCR mixture was composed of 25 μL GoTag Green Master Mix (Promega), 1 μL 8F (20 μM), 1 μL 1510R (20 μM) and 1 μL template DNA in a total volume of 50 μL. The PCR conditions were 5 min of initial denaturation at 95°C followed by 25 cycles of denaturation (30 s at 95°C), annealing (30 s at 56°C) and elongation (90 s at 72°C), with a final extension at 72°C for 8 min. The PCR products were purified by ethanol precipitation.

Cloning and Sequencing

The PCR products were ligated to pCR-XL-TOPO vectors using the TOPO XL PCR Cloning kit (Invitrogen) and One-Shot TOP10 electrocompetent *E. coli* cells (Invitrogen) were transformed, according to the manufacturer's instructions. Approximately 800 transformant colonies from each library were picked randomly and plasmid extraction was performed using the Montage Plasmid MiniPrep96 Kit (Millipore) and a MULTIPROBE II-Robotic Liquid Handling System.

The 5′ half of the cloned 16S rRNA genes was determined by cycle sequencing using BigDye Terminator v3.1 Cycle

Table 1 Characteristics of the volunteers and sample collection date

Volunteer	Age	Sex	Nationality	BMI	Smoker	Antibiotics	Collection date	
							Biopsies	Feces
1	29	F	Spain	21.5	no	–	27-11-07	08-02-08
2	26	F	Spain	20.2	no	–	27-11-07	11-12-07
3	36	F	Spaina	23.7	no	–	27-11-07	02-01-08
4	61	F	Spain	22.6	yes	–	27-11-07	08-02-08
5	42	F	Spain	32.0	no	–	11-12-07	11-12-07
6	33	M	Italy (48 months in Spain)	25.4	no	Amoxicilin 1 month earlier	02-06-08	02-06-08
7	37	M	Spain	31.3	no	Ampicilin 4 months earlier	03-06-08	03-06-08
8	40	M	Spain	24.4	ex	–	03-06-08	03-06-08
9	36	M	Mexico (8 months in Spain)	24.4	no	–	03-06-08	03-06-08

BMI body mass index, weight/(height^2)

Sequencing kit (Applied Biosystems) and 0.625 μM of 8F primer. Sequences were analyzed on ABI 3730 sequencers (Applied Biosystems).

Sequence Analysis and Taxonomic Affiliation

Base-calling of each sequence was performed by Pregap4; sequences were then revised by using the Trev and Gap4 programs, all in the Staden package [28]. After adjusting for quality values, the average read length was around 700 nucleotides.

The taxonomic affiliation of sequences was performed by similarity searches against a taxonomically curated dataset from the Ribosomal Database Project [5, 6] made as follows: from an original set of about 350,000 sequences, we obtained a non-redundant dataset of about 65,000 sequences with known taxonomic affiliation after performing a clustering at 99% of similarity using the cd-hit-est program [15]. A local BLAST search was performed against this dataset. Best-hit sequences were used to assign a minimal but confident taxonomic position to each sequence. When the taxonomic position was not clear, we stopped the assignation at the last clear phylogenetic level, leaving successive levels as "unidentified".

Estimation of Bacterial Diversity

Rarefaction curves were calculated using PAST (PAlaeontological STatistics) ver. 1.67 [12]. The Shannon diversity index [26] and the Chao1 richness estimator [3] were also calculated.

Comparing Sample Bacterial Composition

1. *UniFrac analysis*. Representative sequences of the clusters at 98% of similarity obtained with the cd-hit-est program [15] were aligned using mothur [25] and the aligned sequences of the Greengenes 'Core Set' as template alignment [7]. The closest template for each sequence was found using 9-mer searching and the pairwise alignment between the sequences and the templates was made using the Gotoh algorithm. A neighbor-joining tree was obtained with the programs DNADIST (by using the F84 model of nucleotide substitution) and NEIGHBOR from the PHYLIP package [9]. The derived tree was used as input for UniFrac together with taxa abundance in the different communities [16]. The UniFrac metric measures the difference between two communities in terms of the branch length that is unique to one community or the other. We employed weighted UniFrac, which weights the branches based on the relative abundance of a given sequence for each particular community. To compare multiple communities, we used principal coordinate analysis (PCoA).
2. *Analyzing variability in sample composition*. We first used detrended correspondence analysis (DCA) to explore patterns of variation in the taxonomic distributions found in our samples. We applied CA at several taxonomic levels. We then applied a Bayesian hierarchical model to analyze the variability of the bacterial composition between samples. For the sample of type k=1, 2 (representing feces and biopsies, respectively) from individual i=1,…, 9, the model assumes that the vector $\mathbf{Y}_{ik}=(Y_{ik1},\ldots,Y_{ikJ})^t$ with the number of sequences found in each taxon j=1, …, J is distributed according to a multinomial distribution with parameters equal to the total number of sequences n_{ik} of that sample and unknown proportions $\boldsymbol{\pi}_{ik}=(\pi_{ik1}, \ldots\pi_{ikJ})^t$. Then, the variation of the π_{ikj} (on the log-odds scale) is decomposed into individual, type and taxon random effects plus interactions according to logit $\pi_{ikj} = a + l_i + q_j + d_k + j_{ij} + g_{kj} + e_{ikj}$. All these parameters are assigned normal prior distributions with unknown variances that, in turn, are assigned prior distributions. Inference is made through Markov chain Monte Carlo methods, which provide a sample from the posterior distribution of the model parameters. See the appendix for further details. This model allows to estimate the unknown proportions π_{ikj} as well as to decompose their variation into different sources while taking into account the sampling variation due to the different number of sequences in each sample. The estimated proportions π_{ikj} (or their log-odds transformation) are then used for further statistical analyses.
3. *Evaluating closeness*. In order to assess the similarity between samples according to their bacterial composition, we computed Euclidean distances between samples based on their compositions (on the logit scale) estimated with the above Bayesian model.
4. *Predicting sample type from composition*. We applied linear discriminant analysis (LDA) and classification and regression trees (CART) to assess whether community composition could characterize sample type.

The statistical analyses were carried out using the free-license R package [24]. The biodiversity and richness indices were computed with the vegan R package [18].

Nucleotide Sequence Accession Numbers

The non-redundant sequences from this study have been deposited in the GenBank database under accession numbers GU097883–GU108023.

Results

A total of 13,368 sequences were obtained with an average length of 710 bp, distributed in 18 libraries with around 740 clones per library. Clones with ≥98% of sequence similarity

Table 2 Abundance (Obs), Chao1 richness estimator (Chao1) and associated standard error (SE), and Shannon biodiversity index calculated for each sample (*B* biopsy samples, *F* fecal samples) at the levels of family and clusters at 98% of identity

		Families				Clusters 98% identity			
Sample	Seq	Obs	Chao1	Chao 1 SE	Shannon	Obs	Chao1	Chao 1 SE	Shannon
B1	763	29	51	33.41	1.91	296	545.11	55.79	5.24
F1	705	16	16.75	2.29	1.69	199	405.72	59.84	4.64
B2	742	29	30.5	3.49	2.58	173	344.73	53.44	4.17
F2	867	18	23	17.14	1.92	281	680.03	96.4	4.77
B3	751	28	32.67	5.92	1.71	281	746	108.9	4.87
F3	772	14	17	11.66	1.41	219	490.55	75.53	4.63
B4	681	20	20.75	2.29	1.94	166	287.54	38.92	4.18
F4	708	11	11.5	3.74	1.26	126	241.56	44.53	3.74
B5	863	30	48.33	28.64	1.99	260	518.78	66.49	4.97
F5	740	16	18	5.29	1.67	206	399.89	52.71	4.52
B6	617	14	14.75	2.29	1.48	171	365	61.23	4.27
F6	749	10	16	[a]	1.01	182	345.9	49.22	4.14
B7	623	15	20	10.17	1.37	106	191.56	34.65	3.36
F7	701	14	17	[a]	1.74	216	448.92	60.83	4.54
B8	832	25	26.5	3.49	1.81	273	481.56	49.9	5.03
F8	678	15	15.25	1.31	1.86	226	679.25	128.31	4.69
B9	754	23	30	10.27	1.76	249	500.67	59.69	4.58
F9	822	17	17.5	3.74	1.56	258	663.03	101.23	4.56

[a] Not computable

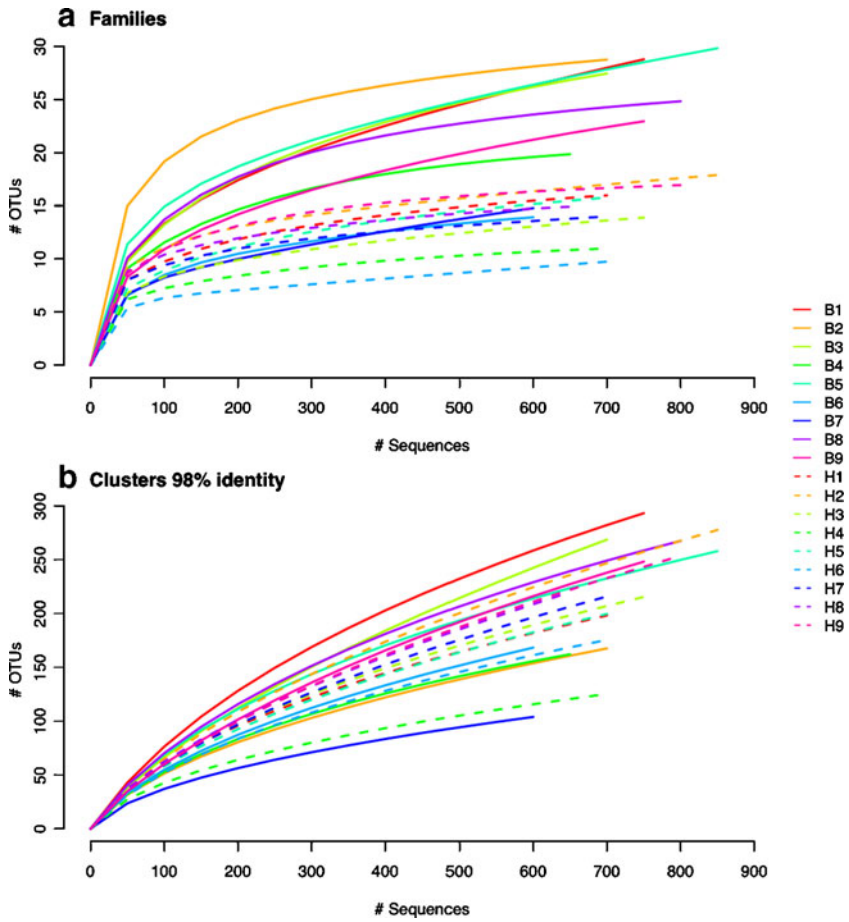

Figure 1 Rarefaction curves for each sample calculated at family level (**a**) and at 98% of sequence similarity clustering (**b**). *B* biopsy samples, *F* fecal samples

were grouped. In what follows, we will refer to these groups as phylotypes. In total 1,793 phylotypes were identified with 1,086 and 1,186 detected in the rectal mucosa and feces, respectively. Table 2 summarizes sequence counts, abundance and biodiversity at family and phylotype levels for each library.

Rarefaction Analysis

To determine the fraction of operational taxonomical units (OTUs, considering as such any of the extant taxonomic units under study) present in the samples that had been recovered, we carried out a rarefaction analysis (Fig. 1). Rarefaction curves were obtained by plotting the number of observed OTUs against the number of cloned sequences. At family level, curves reach or nearly reach a plateau for most samples, whereas for curves calculated using phylotypes the upward phase seems to be ongoing. This means that we have observed most of the families present in most samples, but it also indicates there are quite a few phylotypes that have been missed. The rarefaction curves also show that the sampled communities were less diverse in the fecal samples than in the respective biopsies.

Compared Richness and Diversity

We employed the Chao1 estimator of total richness to estimate the number of families and phylotypes present in the samples (Table 2). The comparison of the observed and estimated number of phylotypes indicates substantial

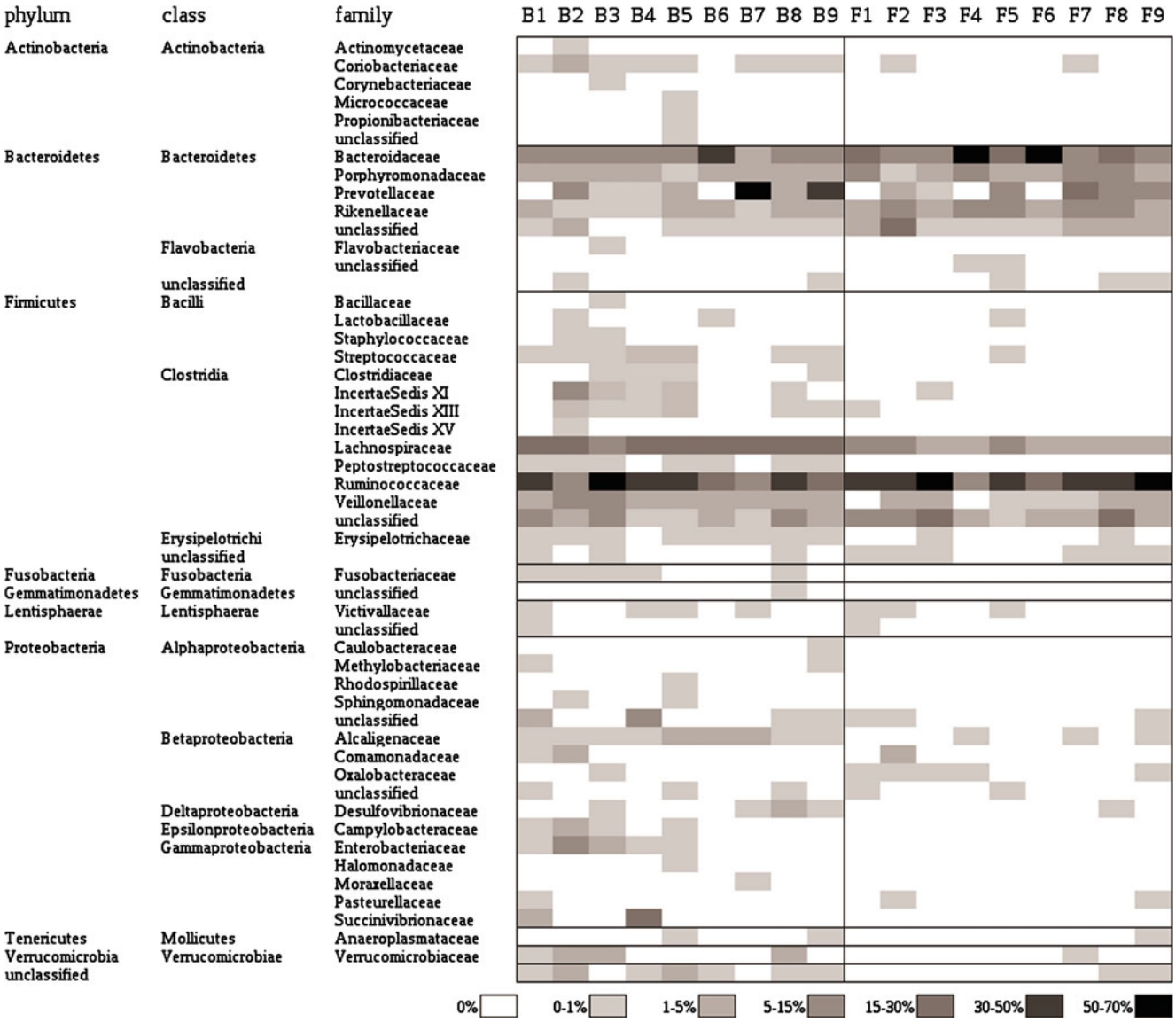

Figure 2 Percentage of sequences (in a gray scale) belonging to biopsies and fecal samples at phylum, class and family taxonomic levels. *B* biopsy samples, *F* fecal samples

numbers of unseen phylotypes in the samples, which could only be detected after sequencing many more clones (an average of 470 phylotypes are estimated to be present while only 215 were detected), which is in agreement with the rarefaction curves at this taxonomic level. This can be put down to the fact that many phylotypes appear at very low frequencies. The Shannon index (*H*), that correlates positively with species richness and evenness was also calculated at both family and phylotype levels. Overall, Chao1 estimates and Shannon diversity indices indicate great richness of the studied intestinal bacterial communities. Furthermore, there are few differences in the diversity found between fecal and biopsy samples. As expected, both richness and biodiversity decreased according to more inclusive taxa such as genus, family, and order (data shown only for family level in Table 2).

Compared Composition

Of the 13,368 clones analyzed, only 668 (5.0%) had a best hit in the database with a similarity lower than 97%, and 475 (3.6%) lower than 96%. The distribution of the 16S rRNA gene sequences at phylum, class and family levels is shown in Fig. 2. We observed that most sequences were assigned to the *Firmicutes* and *Bacteroidetes* phyla, which have repeatedly been described as major and functionally significant components of the human intestinal microbiota. *Proteobacteria* was the third most abundant phylum and its presence was lower in feces than in rectal biopsies. Low prevalence of other phyla was also found in the biopsy samples, such as *Actinobacteria*, *Fusobacteria*, *Gemmatimonadetes*, *Lentisphaerae*, *Tenericutes*, and *Verrucomicrobia*, which were lower or even absent in fecal samples. The relatively low abundance of *Actinobacteria* could be a result of an insufficiently rigorous cell lysis procedure (this phylum has been found as a major constituent of the GI tract microbiota by using other molecular approaches [33]).

The *Firmicutes* phylum covered 59.4% of the total number of sequences, and 67.9% of the phylotypes. Most (96.9%) of the *Firmicutes* sequences belonged to the *Clostridiales* order, being *Ruminococcaceae* and *Lachnospiraceae* the most abundant families. *Lachnospiraceae* was less abundant in feces than in rectal biopsies, while no trend was observed for *Ruminococcaceae*. The *Bacteroidetes* phylum included 36.1% of the sequences and 26.5% of the phylotypes. These were almost exclusively members of the *Bacteroidales* order (99.3%). Large variation between individuals was observed in the relative abundance of the families belonging to this order. The counts for the *Rikenellaceae* family were higher in feces than in rectal biopsies and the same occurred in general for *Bacteroidaceae*, again with high between-individual variability. All *Proteobacteria* classes were found in feces and biopsies, being *Alpha-*, *Beta-* and *Gammaproteobacteria* the most abundant. There was large variability in the abundances between samples where *Proteobacteria* were detected. *Betaproteobacteria* was the only class found in all individuals, at least in biopsies.

Sample composition was also studied at the species level by working with phylotypes defined using a threshold of 98% identity (Supplementary Figure 1). We observed a remarkable portion of species in each individual as sample type specific. An average of 17% of the phylotypes detected in one subject was present in both their feces and rectal biopsy, and an average of 52% of the cloned sequences belonged to clusters shared between the two types of sample. Species shared between feces and their respective biopsy also differed in their relative abundance, depending on the family. For instance, in *Lachnospiraceae*, most species found were not shared between the two paired samples and those that were had similar relative abundances in biopsies and feces. In contrast, many *Ruminococcaceae* species found in biopsies were also present in feces, but their relative abundance was different in the two types of sample.

Variation Between Individuals and Samples

Detrended correspondence analysis showed a great deal of variation in the sample community composition (Supplementary Figure 2). At coarse taxonomic levels such as phylum, no pattern was observed. However, when using intermediate levels such as family or genus, DCA plots separated fecal

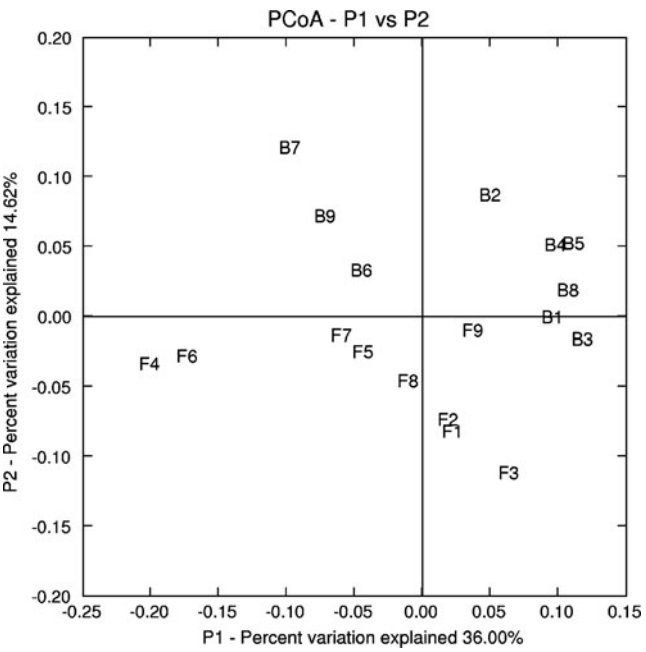

Figure 3 Principal coordinate analysis for biopsies and feces at 98% of sequence similarity clustering. Weighted UniFrac was used in the comparison. *B* biopsy samples, *F* fecal samples

and biopsy samples. At phylotype level a new pattern emerged showing fecal and biopsy samples to be closer to each other. We also used UniFrac to compare the composition of the sampled bacterial communities at this same level, taking into account phylogenetic distances between phylotypes besides their abundance. The UniFrac comparison showed some clustering by sample type similar to the DCA plots at genus or family level (Fig. 3).

The Bayesian model confirmed the high level of variation in the community composition between samples. It also revealed that most of the variation in the proportions π_{ikj} characterizing community composition was mainly due to differences between individuals (45%) and between sample types (45%), and to a lesser extent to differences between taxa (5%). These results were consistent across taxonomic levels.

A heatmap of the posterior medians of the π_{ikj} parameters (i.e., the median of the sample from the posterior distribution for each parameter provided by the MCMC methods) at family level (Fig. 4) highlighted, as stated above, that most communities were dominated by species from just a few taxa, though there was a low prevalence of many other families too. We also computed Euclidean distances between community distributions (on

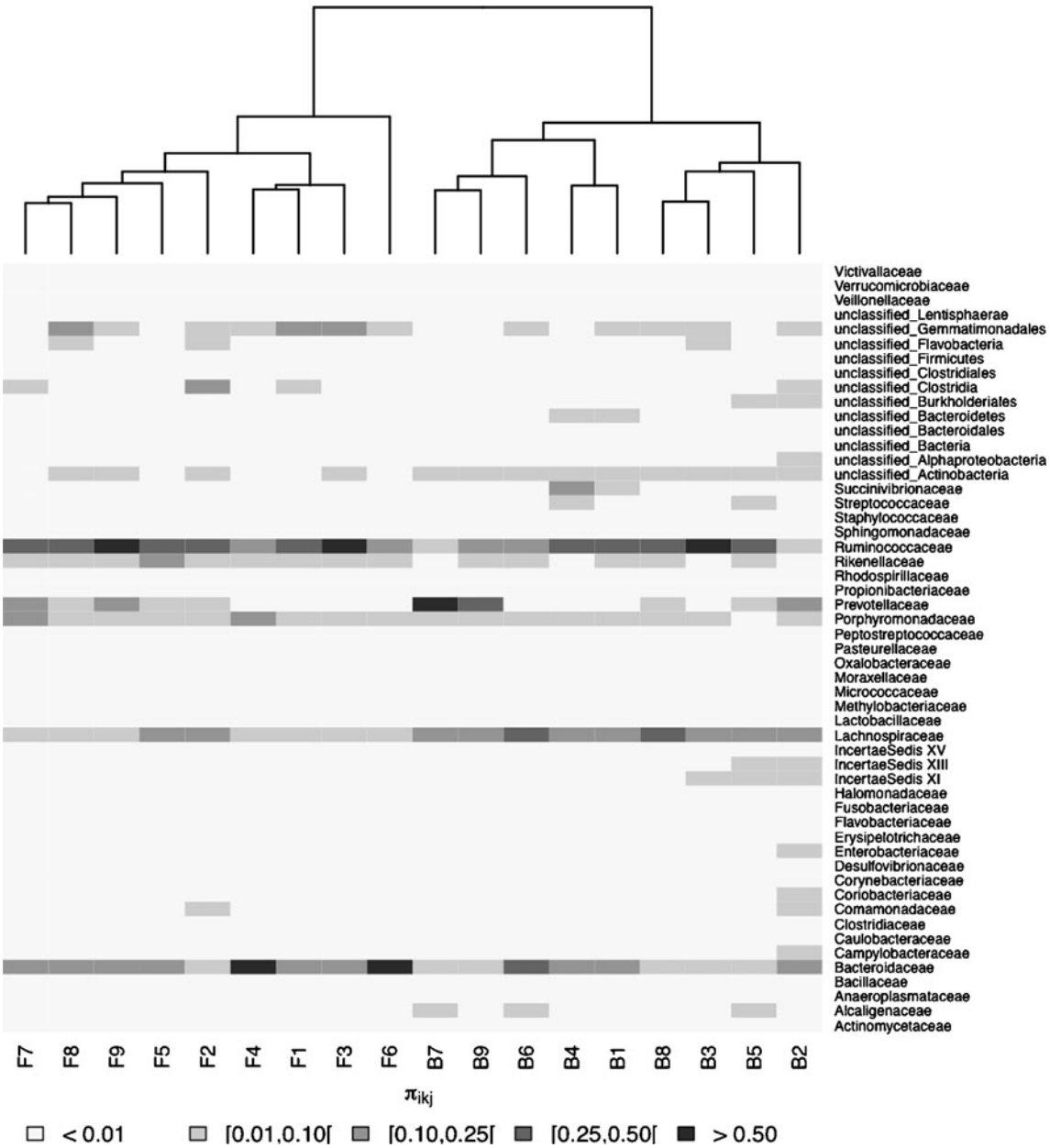

Figure 4 Heatmap at family level of the posterior medians of π_{ikj} grouped into five intervals. On top, hierarchical cluster based on Euclidean distances of the estimated distributions $\boldsymbol{\pi}_{ik}=(\pi_{ik1},\ldots,\pi_{ikJ})^t$ on the log-odds scale. *B* biopsy samples, *F* fecal samples

the log-odds scale) estimated with the Bayesian model (see Appendix) to assess similarities between samples. At family level, a hierarchical cluster based on this distance matrix revealed that the estimated community distributions grouped samples by sample type. Each column in Table 3 lists all samples ranked by increasing distance to the fecal sample that appears in the column header. In most instances, we can see that for a given sample the closest ones are those of the same type, as indicated by the dendrogram in Fig. 4, i.e., the closest samples to a given fecal sample are usually other fecal samples. For individuals six, seven, and nine, the closest biopsy sample to their fecal sample turned out to be their own paired biopsy sample. These are three of the five individuals that provided both samples on the same day. It should be noted that biopsy six is the closest biopsy to seven out of the nine fecal samples and is the second closest to the remaining two. This is partly due to the relative low diversity present in biopsy six that makes its bacterial distribution similar to those found in feces. Similar results were found at genus level.

Finally, LDA and CART did not prove useful to discriminate sample type by community composition, probably due to the large variability observed and the relatively small sample size. For the LDA, we tested multivariate normality and homokedasticity of the covariance matrix of the logit π_{ikj} (the output of the Bayesian model) and did not find significant departures from the null, though this again may be a consequence of the relatively reduced sample size. CART in contrast is a non-parametric method and more robust in general (less sensitive to outliers, invariant to monotonic transformations of the variables, etc.), so it should be less affected by the sample size.

Table 3 Sample ordering based on the Euclidean distance among the logit of the taxonomic distributions π_{ik}

Order	F1	F2	F3	F4	F5	F6	F7	F8	F9
1	F1	F2	F3	F4	F5	F6	F7	F8	F9
2	F4	F8	F8	F1	F8	F4	F8	F7	F8
3	F8	F1	F7	F3	F9	F3	F9	F9	F7
4	F3	F9	F4	F7	F7	F7	F3	F3	F5
5	F9	F3	F1	F8	F1	F8	F5	F5	F1
6	F5	F5	F9	F6	F2	F1	F4	F1	F3
7	F2	F7	F2	F9	F3	F9	F2	F2	F2
8	F7	B6	F5	F5	B6	F5	F1	F4	F4
9	B6	F4	F6	B6	F4	*B6*	*B7*	B6	*B9*
10	F6	B9	B6	F2	B9	F2	F6	B7	B6
11	*B1*	B8	B7	B7	B7	B7	B6	B9	B7
12	B9	B7	B9	B9	B8	B9	B9	*B8*	F6
13	B8	B3	*B3*	*B4*	B3	B8	B8	F6	B8
14	B7	B1	B8	B8	*B5*	B3	B3	B3	B1
15	B3	F6	B1	B1	F6	B1	B1	B1	B3
16	B4	*B2*	B4	B3	B1	B4	B4	B4	B4
17	B5	B4	B5	B5	B4	B5	B5	B5	B5
18	B2	B5	B2	B2	B2	B2	B2	B2	B2

In each column, samples are ordered by increasing distance to the fecal sample in the column header. Highlighted is the paired biopsy of each fecal sample

Discussion

Most of the studies carried out until now on gut metagenomics do not take into consideration that a fecal sample contains a microbial composition that not necessarily can be taken as a good predictor of a corresponding intestinal one. Here, we address the analysis of equivalences and/or correspondences between fecal and rectal mucosal samples from nine healthy individuals. Extrapolation of our results to other sites in the intestine should be made with caution as the microbiota composition may vary along the gut.

Our work confirmed the findings from previous studies that suggested that fecal and mucosal microbial diversity from the same individual are not similar [4, 8, 14, 19, 20, 32]. However, our approach has several differences over previous studies. Firstly, we attempted to provide this diversity comparison from fecal and mucosal samples collected as close as possible in time. Secondly, we did not carry out a bowel cleansing prior to colonoscopy to avoid the potential disturbance of mucosal microbiota associated with this procedure. This however opens the possibility for rectal biopsies to contain bacteria from feces loosely sticking to the mucus but not being actually part of the mucosal microbiota. Biopsies were taken in the absence of macroscopic feces. It is nevertheless unlikely that biopsies were contaminated with fecal material because stools usually have hard consistency and are not attached to the mucosa in the rectum, where they are formed to be expelled outside the body and do not adhere to the mucosa because nothing is absorbed nor secreted there, unlike other sections of the intestine. Given the differences we found in the microbiota between the two types of sample, we think this can hardly be a generalized situation, though the possibility cannot be ruled out completely. Finally, the results presented here have been generated through sequence analysis of clone libraries, which enabled the identification of the microorganisms.

Overall, we found that two phyla, *Firmicutes* and *Bacteroidetes*, dominated those communities, accounting for nearly 85% of all sequences. However, we also observed large between-sample variability in community composition at nearly all taxonomic levels. At phylotype level in particular, the majority of phylotypes detected were sample-specific, showing that each individual carried a

particular combination of bacterial lineages, as previously reported [8, 23, 29, 31]. Strong within-subject variability was also found in the feces-biopsy paired samples.

Our results suggest that community composition in fecal samples is not highly representative of the microbiota in the rectal gastrointestinal tract. In fact, at family and genus levels, taxa distributions group samples by type rather than individual, even for those sample pairs collected the same day. Evaluating the closeness between samples based on distances between their estimated compositions (on the log-odds scale), we found that any fecal sample is more similar to any other fecal sample than to a rectal biopsy sample. We also found that the closest biopsy sample to the feces of an individual was his own paired biopsy sample in three of the five individuals that provided both samples on the same day, a finding that cannot be considered as conclusive given the (statistically) small sample size, especially considering that one of the biopsies is very similar to all fecal samples. These results confirm that the intestinal microbiota is an extremely complex community, the richness and diversity of which seems to be under-represented in fecal samples. However, it has yet to be assessed whether this impoverishment is because not all species in the intestine are susceptible to ending up in feces or whether it is a consequence of the impact of the sudden change in growing conditions (temperature, oxygen, nutrient availability, etc.) on leaving the body. Also, the biopsy samples were frozen immediately after collection, whereas fecal samples were not. It should be considered the possibility that this might have had an impact on reducing the biodiversity found in feces compared to biopsies.

We are well aware that feces will continue to be used in the study of gut microbiota because they are easier to collect than intestinal samples and current work with biopsies is limited to certain methodologies due to the quantity of material that could be obtained from them. Actually, each sample type may provide a distinct and complementary picture of the diversity and ecology found in the human gut microbiota. However, since bacterial diversity in the colon mucosa is under-represented in feces, we think it is important to caution researchers, especially those dealing with bacteria-associated pathologies, on the potential risk of making inference about the intestinal mucosal microbiota, or even about the entire gut one, from that found in feces.

Further research is needed to assess whether, despite the disparity of assemblages of microorganisms found here, there exists a functional microbial core in the microbiome. This could be explored with other "-omic" strategies, as some recent studies have shown [23, 31]. Another alternative is the existence of multiple community configurations that are functionally equivalent despite not sharing a common functional core. Analogously, the question of whether feces are representative of gut microbiota would be better addressed at a functional level rather than at the organism level.

Acknowledgments This work has been funded by grants BFU2008-04501-E and SAF2009-13032-C02-01 from Ministerio de Ciencia y Educación, Spain as well as Prometeo/2009/092 and GVPRE/2008/010 from Generalitat Valenciana, Spain to A.M. A.D. and N.J. are recipients of a fellowship from Instituto de Salud Carlos III, Spain. J.J. A. and G.D. are recipients of a post-doctoral contract from CIBER en Epidemiología y Salud Pública (CIBEResp), Spain. The authors thank all the volunteers for participating in this study.

References

1. Baker GC, Smith JJ, Cowan DA (2003) Review and re-analysis of domain-specific 16S primers. J Microbiol Methods 55:541–555
2. Bibiloni R, Tandon P, Vargas-Voracka F, Barreto-Zuniga R, Lupian-Sanchez A, Rico-Hinojosa MA, Guban J, Fedorak R, Tannock GW (2008) Differential clustering of bowel biopsy-associated bacterial profiles of specimens collected in Mexico and Canada: what do these profiles represent? J Med Microbiol 57:111–117
3. Chao A (1987) Estimating the population size for capture-recapture data with unequal catchability. Biometrics 43:783–791
4. Codling C, O'Mahony L, Shanahan F, Quigley EM, Marchesi JR (2010) A molecular analysis of fecal and mucosal bacterial communities in irritable bowel syndrome. Dig Dis Sci 55:392–397
5. Cole JR, Chai B, Farris RJ, Wang Q, Kulam-Syed-Mohideen AS, McGarrell DM, Bandela AM, Cardenas E, Garrity GM, Tiedje JM (2007) The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. Nucleic Acids Res 35:D169–D172
6. Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, Kulam-Syed-Mohideen AS, McGarrell DM, Marsh T, Garrity GM, Tiedje JM (2009) The ribosomal database project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res 37:D141–D145
7. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 72:5069–5072
8. Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, Gill SR, Nelson KE, Relman DA (2005) Diversity of the human intestinal microbial flora. Science 308:1635–1638
9. Felsenstein J (1989) PHYLIP—Phylogeny Inference Package (Version 3.2). Cladistics 5:164–166
10. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE (2006) Metagenomic analysis of the human distal gut microbiome. Science 312:1355–1359
11. Guarner F, Malagelada JR (2003) Gut flora in health and disease. Lancet 361:512–519
12. Hammer Ø, Harper DAT, Ryan PD (2001) PAST: Paleontological statistics software package for education and data analysis. Palaeontol Electronica 4:9
13. Kelly D, King T, Aminov R (2007) Importance of microbial colonization of the gut in early life to the development of immunity. Mutat Res 622:58–69
14. Lepage P, Seksik P, Sutren M, de la Cochetière MF, Jian R, Marteau P, Doré J (2005) Biodiversity of the mucosa-associated microbiota is stable along the distal digestive tract in healthy individuals and patients with IBD. Inflamm Bowel Dis 11:473–480

15. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659
16. Lozupone C, Hamady M, Knight R (2006) UniFrac—an online tool for comparing microbial community diversity in a phylogenetic context. BMC Bioinform 7:371
17. Neish AS (2009) Microbes in gastrointestinal health and disease. Gastroenterology 136:65–80
18. Oksanen J, Kindt R, Legendre P, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Wagner H (2008) Vegan: community ecology package. R package version 1.15-1. http://cran.r-project.org/, http://vegan.r-forge.r-project.org/
19. Ott SJ, Musfeldt M, Timmis KN, Hampe J, Wenderoth DF, Schreiber S (2004) In vitro alterations of intestinal bacterial microbiota in fecal samples during storage. Diagn Microbiol Infect Dis 50:237–245
20. Ouwehand AC, Salminen S, Arvola T, Ruuska T, Isolauri E (2004) Microbiota composition of the intestinal mucosa: association with fecal microbiota? Microbiol Immunol 48:497–500
21. Pignatelli M, Moya A, Tamames J (2009) EnvDB, a database for describing the environmental distribution of prokaryotic taxa. Environ Microbiol Rep 1:191–197
22. Preidis GA, Versalovic J (2009) Targeting the human microbiome with antibiotics, probiotics, and prebiotics: gastroenterology enters the metagenomics era. Gastroenterology 136:2015–2031
23. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Tap J, Lepage P, Bertalan M, Batto JM, Hansen T, Le Paslier D, Linneberg A, Nielsen HB, Pelletier E, Renault P, Sicheritz-Ponten T, Turner K, Zhu H, Yu C, Li S, Jian M, Zhou Y, Li Y, Zhang X, Li S, Qin N, Yang H, Wang J, Brunak S, Doré J, Guarner F, Kristiansen K, Pedersen O, Parkhill J, Weissenbach J, MetaHIT Consortium, Bork P, Ehrlich SD, Wang J (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464:59–65
24. R Development Core Team (2009) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/
25. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75:7537–7541
26. Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27:379–423
27. Sokol H, Lay C, Seksik P, Tannock GW (2008) Analysis of bacterial bowel communities of IBD patients: what has it revealed? Inflamm Bowel Dis 14:858–867
28. Staden R, Beal KF, Bonfield JK (2000) The Staden package, 1998. Methods Mol Bio 132:115–130
29. Tap J, Mondot S, Levenez F, Pelletier E, Caron C, Furet JP, Ugarte E, Muñoz-Tamayo R, Paslier DL, Nalin R, Dore J, Leclerc M (2009) Towards the human intestinal microbiota phylogenetic core. Environ Microbiol 11:2574–2584
30. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. Nature 444:1027–1031
31. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, Egholm M, Henrissat B, Heath AC, Knight R, Gordon JI (2009) A core gut microbiome in obese and lean twins. Nature 457:480–484
32. Zoetendal EG, von Wright A, Vilpponen-Salmela T, Ben-Amor K, Akkermans AD, de Vos WM (2002) Mucosa-associated bacteria in the human gastrointestinal tract are uniformly distributed along the colon and differ from the community recovered from feces. Appl Environ Microbiol 68:3401–3407
33. Zoetendal EG, Rajilić-Stojanović M, de Vos WM (2008) High-throughput diversity and functionality analysis of the gastrointestinal tract microbiota. Gut 57:1605–1615