The ISME Journal (2011) 5, 519–531

www.nature.com/ismej

ORIGINAL ARTICLE

Unveiling an abundant core microbiota in the human adult colon by a phylogroup-independent searching approach

Monika Sekelja[1,2], Ingunn Berget[3], Tormod Næs[1,4] and Knut Rudi[1,5]
[1]*Nofima, The Norwegian Institute of Food, Fisheries and Aquaculture Research, Aas, Norway;* [2]*Department of Informatics, University of Oslo, Oslo, Norway;* [3]*The Centre for Integrative Genetics (CIGENE), Norwegian University of Life Sciences, Aas, Norway;* [4]*Department of Mathematics; University of Oslo, Oslo, Norway and* [5]*Hedmark University College, Hamar, Norway*

The potential presence of widespread and stable bacterial core phylogroups in the human colon has promoted considerable attention. Despite major efforts, no such phylogroups have yet been identified. Therefore, using a novel phylogroup- and tree-independent approach, we present a reanalysis of 1 114 722 V2 region and 71 550 near full-length 16S rRNA sequences from a total of 210 human beings, with widespread geographic origin, ethnic background and diet, in addition to a wide range of other mammals. We found two highly prevalent core phylogroups (cores 1 and 2), belonging to the clostridial family *Lachnospiraceae*. These core phylogroups showed a log-normal distribution among human individuals, while non-core phylogroups showed more skewed distributions towards individuals with low levels compared with the log-normal distribution. Molecular clock analyses suggest that core 2 co-evolved with the radiation of vertebrates, while core 1 co-evolved with the mammals. Taken together, the stability, prevalence and potential functionality support the fact that the identified core phylogroups are pivotal in maintaining gut homeostasis and health.
The ISME Journal (2011) **5,** 519–531; doi:10.1038/ismej.2010.129; published online 26 August 2010
Subject Category: microbial ecology and functional diversity of natural habitats
Keywords: core microbiota; *16S rRNA* gene; human; evolution

Introduction

The interaction between the host and its gut microbiota is of fundamental importance for host health and disease. Recent research has shown that human intestinal microbiota can be linked to obesity (Ley *et al.*, 2006; Turnbaugh, 2006), allergies, inflammatory bowel disease (Frank *et al.*, 2007) and colon cancer (Scanlan *et al.*, 2008). It is hypothesized that the microbiota is composed of a variable part that differs between individuals and a stable part that is found in the vast majority of human beings (Turnbaugh *et al.*, 2007, 2009). Factors affecting the variable part of the microbiota are now rather well documented and depend on, among others, the environment we are exposed to, our genetic profile and our diet (Alm *et al.*, 1999; Zoetandal *et al.*, 2001; Stewart *et al.*, 2005; Ley *et al.*, 2006; Khachatryan *et al.*, 2008; Turnbaugh *et al.*, 2009).

Correspondence: M Sekelja or K Rudi, The Norwegian Institute of Food, Fisheries and Aquaculture Research, Osloveien 1, Ås, Akershus 1430, Norway.
E-mail: monika.sekelja@nofima.no or knut.rudi@nofima.no
Received 3 May 2010; accepted 5 July 2010; published online 26 August 2010

Defining the stable part of the microbiota (core microbiota) has proven to be much more difficult (Turnbaugh *et al.*, 2009). Limitations with previous studies, however, are that they are based on either searching for phylogroups at predefined evolutionary depths or comparison of phylogenetic trees. Owing to the clonal growth of bacteria, that is, asexual reproduction, there is no real rationale for defining phylogroup depths (Doolittle and Zhaxybayeva, 2009). The current discussion about the existence of operational taxonomic units (OTUs) or predefined taxonomic core phylogroups is therefore more or less semantic, that is, depending on defined phylogroup cutoff values. The main problem with the phylogenetic tree-based comparative approaches, on the other hand, is the uncertainties in tree construction and the nearly infinite combinatorial possibilities of phylogenetic trees (Liu *et al.*, 2009), making large-scale tree comparisons difficult or impossible. Thus, it is likely that the current approaches have failed to recognize important ancient co-evolution events between the core microbiota and the host (Turnbaugh *et al.*, 2009, 2010).

The aim of this work was to search for a human core microbiota independent of both predefined

phylogroup depths and phylogenetic trees. Our strategy was to use alignment-independent bi-linear multivariate modelling (AIBIMM) for global analyses of 16S rRNA phylogeny, in which gene clones cluster in a principal component analysis (PCA) map based on their sequence similarity (details are given in Rudi *et al.*, 2006). The PCA map was divided into a grid, and the core potential of each square of the grid was calculated as a function of the density and the diversity of the cloned sequences within the square. In this case, high clone diversity means that many different individuals were represented within the square. Larger regions in the PCA map with high core potential were subsequently investigated further as possible core phylogroups. A schematic outline of the approach is given in Figure 1.

To address co-evolution of core bacterial groups, we used both fixed- and relaxed-rate molecular clock analyses using data from chicken and termites as calibrators. This enabled the determination of approximate time points for the co-evolution events between the host and the core microbiota.

We analysed both the largest full-length and deep sequencing *16S rRNA* gene data sets on human gut microbiota available today, in addition to a range of simulated data sets for method validation. The most extensive near full-length data set comprised of more than 14 000 *16S rRNA* gene sequences isolated from stool samples collected over the course of 1 year from 12 adult obese individuals (Ley *et al.*, 2006). It contains temporal samples collected from various individuals divided into two diet-treatment groups (fat or carbohydrate restricted). We also analysed three other large full-length data sets on adult humans (Dethlefsen *et al.*, 2008; Li *et al.*, 2008; Turnbaugh *et al.*, 2009), one on infants (Palmer *et al.*, 2007), one on 83 different mammals (Ley *et al.*, 2008) and one on the multiple colonic sites of three adult humans (Eckburg *et al.*, 2005). Finally, a deep sequencing data set of 1 114 722 *16S rRNA* gene V2 region sequences from 154 persons were included (Turnbaugh *et al.*, 2009). All data sets are summarized in Table 1.

We present the discovery of two core phylogroups (cores 1 and 2), which have co-evolved with the radiation of vertebrates and mammalians, respectively.

Materials and methods

To analyse the data, we used a newly developed AIBIMM approach for global *16S rRNA* gene analyses of prokaryotes (Rudi *et al.*, 2006). The Ley human data were used to make the basic phylogenetic PCA model. All other data sets were projected onto this model. Using an in-house developed approach described in section 'Grid-based search', regions with high core potential were identified. The core potential (*cp*) was developed for this purpose and varies between 0 and 1. High *cp* values indicate high potential for belonging to a core, that is, a bacterial phylogroup with a large number of different individuals and time points. In order to make the results more robust against possible influence from individual data points, we accompanied the search with a resampling procedure.

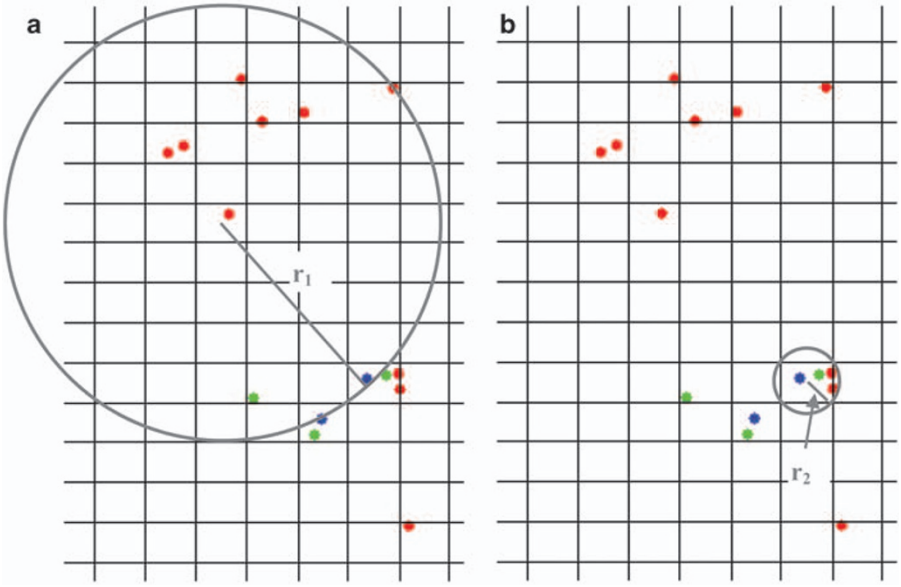

Figure 1 Graphical representation of core searching approach. (**a**) Square with low core potential and (**b**) square with high core potential. Each colour in the plot represents one specific category. The PCA map is divided into small grid squares of equal size. With centre point in the middle of each grid square, the smallest radius needed to include bacteria from all different individuals is found. The bacterial clones located in each grid square are assigned a value, defined as 'core potential', based on the value of the smallest radius. The smaller the radius, the higher the 'core potential'. The areas with highest 'core potential' values are investigated further as the potential core bacterial phylogroups. The core potential (*cp*) of the data points in the grid placed in the centre of the circles is (a) $1—r_1/r_{max}$ and (b) $1—r_2/r_{max}$, where r_{max} is a scaling factor. Thus, the grid with the densest clustering of different categories has the highest core potential.

Table 1 Overview of all data sets used in the analysis

Reference to the published articles	*No. of 16S rRNA sequences*	*Data set description*	*Short title of the data set*
Ley *et al.* (2006)	14 431	12 individuals sampled at four time points (faeces)	Ley human data
Turnbaugh *et al.* (2009)	7161	10 families; adult monozygotic/dizygotic twin pairs and their mothers (faeces from 29 individuals)[a]	Turnbaugh data
Ley *et al.* (2008)	20 470	83 individual mammals (faeces), excluded humans	Ley mammal data
Li *et al.* (2008)	7247	Four-generation Chinese family. Seven members (faeces)	Li data
Dethlefsen *et al.* (2008)	7093	Three individuals (USA/China) treated with antibiotics (faeces)	Dethlefsen data
Palmer *et al.* (2007)	3389	Random samples isolated from seven different infants and two mothers (faeces). In addition, samples from breast milk and vaginal swab of one mother were collected[b]	Palmer data
Eckburg *et al.* (2005)	11 759	Samples isolated from faeces and six colon subdivisions from three adult individuals	Eckburg data
Turnbaugh *et al.* (2009)	1 114 722	Samples isolated from 154 adult individuals (monozygotic/dizygotic twin pairs and their mothers) at one or two time points (time 0 and 57 ± 4 days later)	Turnbaugh V2 data

[a]The faeces from one individual was not analysed due to short sequences.
[b]The 16S rRNA clone library was used as a reference to an extensive microarray analysis of the infant gut microbiota.

Data sets

The extensive human bacterial clone library of 71 550 near full-length and 1 114 722 V2 16S rRNA sequences was assembled from the published articles enlisted in Table 1. The 14 431 16S rRNA sequences collected over the course of 1 year from 12 adult obese individuals were provided by Ley *et al.* (2006). This data set was used to search for possible core microbiota in the human gut as described below. In addition, 7161 full-length and 1 114 722 V2 *16S rRNA* gene sequences from 10 and 54 families, respectively, consisting of twin pairs and their mothers (Turnbaugh *et al.*, 2009), 7247 sequences from a four-generation Chinese family (Li *et al.*, 2008), 7093 sequences from three adult individuals treated with antibiotics (Dethlefsen *et al.*, 2008), 11 759 sequences isolated from mucosal tissues from the six subdivisions of the colon as well as faecal samples from three adult individuals (Eckburg *et al.*, 2005), 3389 sequences from infants and their mothers (Palmer *et al.*, 2007) and 20 470 sequences isolated from 83 different mammals (Ley *et al.*, 2008) were used to validate the cores identified from the Ley human data set. *Lachnospira* clones isolated from termites and chickens, downloaded from the Ribosomal Database Project 10 (RDP10) (Cole *et al.*, 2009), were used to estimate divergence times of core bacterial groups. Ten data sets, each consisting of 998 simulated 16S rRNA sequences, were used to compare OTU- and phylogenetic tree-based method with AIBIMM method and to validate our core search algorithm.

Sequence filtering and preprocessing

All full-length 16S rRNA sequences were pairwise aligned to *Escherichia coli* 16S rRNA sequence and trimmed to start position 49 and end position 1295 of the latter mentioned sequence. The V2 sequences were trimmed to area comprising 159–335 positions in *E. coli 16S rRNA* gene sequence. Sequences that were not long enough to include the predefined start and end positions were removed from the analysis.

Alignment-independent bi-linear multivariate modelling

The DNA sequences were transformed into DNA pentamer frequency table using the computer program PhyloMode (http://www.nofimamat.no/phylomode). A window of five nucleotides is moved along each DNA sequence and the frequencies of different pentamers present in the sequence are stored in a frequency table. The window size of five nucleotides was chosen because it enabled the insight into the phylogeny instead of multimer equalities happening simply by chance. The AIBIMM approach is described in more detail in Rudi *et al.* (2006, 2007). The size of the pentamer frequency table for the data sets corresponded to 'the number of 16S rRNA sequences' × 1024, for example, the size of the pentamer frequency table for the Ley human data was 14 431 × 1024.

The pentamer frequency table was compressed using PCA, a projection method that transforms a data table consisting of possibly correlated variables (in our case, pentamers) into a smaller number of uncorrelated latent variables, called principal components. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. The result is a new set of variables ('loadings') that represent linear functions of the original variables. The new variables are uncorrelated and reflect the most important structure of the data. Values for each sample (in our case, DNA sequence) projected onto these new variables, loadings, are then calculated and called 'scores'. The data points (in our case, unknown bacteria isolated from stool samples) that cluster in the PCA plot are interpreted as closely related. The Ley human data were used to make the main model of gene phylogeny and all other data sets were projected onto this model.

Grid-based search

For each observation in the Ley human data, we have information on its origin, that is, which individual it belongs to, and the specific time period (week number) at which it was sampled. The core microbiota should consist of bacteria that are both widespread (found in each individual) and stable (found at each time period). Therefore, we should search for areas with high density (many similar bacteria) where a large number of different individuals and time points are represented.

The searching process implemented consists of two steps: (1) define 'individual core' microbiota, that is, the stable (over time) microbiota for each individual, and (2) use only the stable microbiota from each individual to search for the universal 'core' microbiota, that is, those that are found in all individuals. The purpose of the first step is to obtain detailed information on each individual separately for the purpose of removing the irrelevant data points (the fluctuating bacteria) before step 2. Step 1 is repeated for each individual. In step 2, a reduced data set consisting of stable bacteria from each individual is used. The criterion and data mining approach is the same in both steps. The PCA score space is divided into a grid, and the *cp* value is computed for each point in the grid. The *cp* is based on a diversity index calculated for each square in the grid. The diversity index $d_i(r)$ for all points within square number i in the grid is the number of different categories (time points in step 1 or individuals in step 2), represented within the region spanned by a circle with radius r and centred within the square, as illustrated in Figure 1. The corresponding *cp* for square i is defined by Equation (1), where r_i is the smallest radius where the diversity index is larger than a given threshold d_{opt}:

$$cp_i = 1 - \frac{r_i}{r_{\max}} \tag{1}$$

Here $r_{\max}$ is the distance between the two points furthest apart in the data set. By scaling the minimum radius r_i in this way, the *cp* is constrained to the interval [0 1], and will hence not be data dependent. A value of 0.5 will, for instance, mean that the radius needs to be half of the maximum to reach d_{opt}.

The minimum radius is calculated as shown in Equation (2), and the diversity index is given by Equation (3):

$$r_{\mathrm{i}} = \arg\min_{\mathrm{r}}(d_{\mathrm{i}}(r) \geqslant d_{\mathrm{opt}}) \tag{2}$$

$$d_{\mathrm{i}}(r) = \#\left\{ (l \in S_{\mathrm{r}}) | (\sum_{j=1}^{N} K_{jl}) > 0 \right\} \tag{3}$$

Here S_{r} is the region spanned by the circle with radius r centred in square i and $\mathbf{K}$ is the $N \times L$ category matrix, where each column is a dummy variable representing the different categories (time points or individuals) and each row j corresponds to a data point in S.

In our case, d_{opt} was chosen to be the maximum number of categories present in the data set. Thus, when searching for the stable bacteria within each individual (step 1), d_{opt} was set to be 4, that is, four different time points. The search for common stable bacteria between individuals was performed with $d_{\mathrm{opt}} = 12$, that is, 12 different individuals. See supplementary material for a discussion on the choice of d_{opt}.

The outcome from step 1 is used to filter out unstable bacteria by removing all data points with *cp* value $< cp_{\mathrm{mean}}$ from the total data set before step 2. Finally, the potential core groups consisting of sequences (data points) with low explained variance are eliminated. In order to speed up the search, *d* and *cp* are only calculated for grids with at least one data point.

Results are visualized using the PCA plot of gene phylogeny and colouring each data point based on its calculated *cp* value. The data points that are greyed out have a *cp* value lower than the mean *cp* value (cp_{mean}) for the total data set. The dark blue data points have *cp* value near cp_{mean}, while the dark red points have the highest possible *cp* value.

Resampling

The above procedure could have been used as is, but in order to make it more robust to possible influence from individual data points locally, we decided to accompany the search with a resampling procedure. More specifically, we resampled the original data after PCA to create 1000 new data sets. The minimum radius (Equation (2)) was calculated for all grids with $cp_i \geqslant cp_{\mathrm{mean}}$, where cp_{mean} is the mean *cp* value for the total resampled data set. Grids with smaller *cp* values are unlikely to represent cores, and leaving them out speed up the procedure. The *cp* values for the original data set were updated to the minimum radius for which all categories were included in its circle in 95% of all resampled data sets (Supplementary Figure S6). This ensures that within this radius there will always be samples belonging to all categories.

Exclusion of potential core groups

Explained variance for each data point was calculated using the four first principal components (PC1–PC4). Data points with low explained variance are not well modelled by the PCA model of gene phylogeny, that is, their location in the plot do not necessarily tell us much about their phylogeny. Thus, these points were disregarded as potential core.

Requirements
The searching process, including search with resampling on all individual sub data sets as well as the final reduced total data set, was performed in the Matlab programming environment (1994–2007; The MathWorks Inc., Natick, MA, USA) on a personal computer with average performances (dual core 2×2.33 GHz, 2 GB RAM) measuring the total elapsed time to be approximately 85 min.

Investigation of stability over time
The next step in our procedure was to investigate whether the core bacterial phylogroups displayed more stability over the individuals and over time than other bacterial phylogroups with similar sequence similarity. For each phylogroup i, the number of bacteria N_i varied. To avoid the sampling error, we selected randomly 750 samples from each phylogroup. This number was chosen on the basis of the minimum amount of bacteria among the four phylogroups. Next, the relative abundance of each group was calculated for each individual and each time point according to Equation (4), where $R_i(j,t)$ is the number of bacteria from group i for individual j and time point t:

$$R_{\mathrm{i}}^{\mathrm{norm}}(j,t) = R_i(j,t) / \sum_t R_i(j,t) \qquad (4)$$

These values were then compared for the different individuals for the different time points using simple plotting techniques. In order to get a quantitative measure of the variability between the individuals at the different time points and also of the variability over time and individual, regular variance estimates were computed for each bacterial group. The variances across individuals at each time point were compared for the different bacteria by using the Bartlett multiple-sample test for equal variances (function in Matlab's Statistics Toolbox called vartestn) to assess whether there are significant differences. The variance was calculated across the individuals for each time point and compared between phylogroups to investigate whether the correspondence between individuals in bacterial abundance was significantly higher in the core compared with the other phylogroups. In addition, the variance of change in bacterial abundance in each phylogroup during the controlled diet period (weeks 12–52) was compared between the phylogroups. Individual nos. 7 and 10 were taken out from the significance testing owing to the lack of sample collection at time points 4 (week 52) and 3 (week 26), respectively.

Validating the distribution of human ethnicity and mammals
The test data sets Turnbaugh data, Ley mammal data, Li data, Dethlefsen data, Palmer data and Eckburg data were preprocessed (see section 'Sequence filtering and preprocessing') and transformed into pentamer frequency table (see section 'Alignment-independent bi-linear multivariate modelling'). Each table was consequently projected onto the PCA model constructed using pentamer frequency table of the Ley human data. Finally, the bacterial abundance for each individual in each data set within the borders of each core phylogroup was determined.

Distribution of genome-sequenced and known butyrate-producing bacteria
To gain insight into the functionality of bacteria within the core groups, we downloaded 18 full-length *16S rRNA* gene sequences of both genome-sequenced and butyrate-producing bacteria isolated from human colon (Louis *et al*., 2004). These sequences were preprocessed and projected into the PCA model as described above.

Validation of human population distribution by deep sequencing V2 data
To evaluate the distribution of sequences belonging to core and two major non-core phylogroups, *Bacteroidetes* and *Faecalibacteria*, we used the Turnbaugh V2 data set and a PCA model constructed using the V2 region of *16S rRNA* gene sequences from the Ley human data. The Turnbaugh V2 data were projected onto the PCA model of the V2 Ley human data, and bacterial sequences belonging to core and non-core phylogroups were counted for each individual. Furthermore, a function called 'Individual Distribution Identification' found in the statistics package Minitab was utilized to investigate population distribution within the four bacterial phylogroups.

Comparison between mucosal and stool samples
Using Eckburg data, we investigated the differences in the distribution of the core bacterial phylogroups inhabiting the colon (mucosa) compared with faeces. The relative amounts of mucosal and stool core bacterial sequences were found for each individual and compared using ordinary Student's *t*-test for the comparison of two means.

Alignment-based phylogenetic analyses
DNA multiple sequence alignments were generated using RDP10 aligner (Cole *et al*., 2009). Phylogenetic analysis was performed using program Mega 4.1 (Tamura *et al*., 2007). The evolutionary history was inferred using the neighbour-joining method (Saitou and Nei, 1987). The bootstrap consensus tree inferred from 500 replicates (Felsenstein, 1985) is taken to represent the evolutionary history of the taxa analysed (Felsenstein, 1985). The evolutionary distances were computed using the maximum

composite likelihood method (Tamura *et al*., 2004) and are in the units of the number of base substitutions per site. All positions containing gaps and missing data were eliminated from the data set (complete deletion option).

We used Unifrac (Lozupone and Knight, 2005; Lozupone *et al*., 2006) to define which lineages in the tree were significantly different. Here, one has to define at which distance from the root to cut the tree so that each lineage that exists at the defined distance from the root will be treated individually in the analysis. Finally, sequences on lineages that did not differ significantly between individuals were defined as shared sequences.

To assign sequences to OTUs, we used Mothur (Schloss *et al*., 2009). This program is widely used for the analysis of community sequence data. The sequence pairwise distances and the cluster groups based on these distances were found using commands dist.seqs and cluster (). To find shared bacterial sequences between individuals, get. sharedseqs () command was used. This command includes an option called label, where you specify at which distance you wish to investigate whether there are shared sequences. Usually, this distance is not known, so one has to try all possible distances where there exist shared sequences.

The hypothesis testing was performed on classification and core identification results from 10 simulated data sets using AIBIMM-, OTU- and tree-based methods. Using two-sample *t*-test, we tested the null hypothesis such that the number of false positives/negatives was equal between AIBIMM- and OTU/tree-based method compared with the alternative hypothesis that AIBIMM gave fewer false positives/negatives.

Divergence time estimation

To infer co-evolution of potential core bacterial groups with their host, we calculated pairwise distances within each core and assumed a 16S rRNA divergence rate of 1% per 50 million years (Ochman and Wilson, 1987),

In addition, we utilized a Bayesian Markov chain Monte Carlo analysis package called BEAST 1.5.3 (Drummond and Rambaut, 2007) (including BEAUti for generating BEAST input files and TreeAnnotator for summarizing BEAST output files). We sampled full-length 16S rRNA sequences from each of the following bacterial groups: core 1, core 2, termite's (invertebrates) *Lachnospira* and chicken's (vertebrates) *Lachnospira*, with a total of 17 sequences. Thereafter, we defined three taxon sets: all sequences (taxon set 1), one sequence from core 1 and one from chicken's *Lachnospira* group that is the closest outgroup to core 1 bacterial group (taxon set 2), and one sequence from core 2 and one from termite's *Lachnospira* group that is the closest outgroup to core 2 bacterial group (taxon set 3). We chose to use 'Hasegawa, Kishino and Yano' as nucleotide substitution model with estimated base frequencies and gamma distributed rates and 'Relaxed Clock: Uncorrelated Log-normal' as molecular clock model, in which the rate at each branch is drawn from a log-normal distribution. The distribution of divergence between the members of taxon sets 2 and 3 was set to 'normal' centred in 330 and 600 million years, with a standard deviation of 10 and 20 million years, respectively. This corresponds to the date estimate of the common ancestor of host species chicken and termites, reported by Blair and Hedges, (2005) and Peterson *et al*. (2008), correspondingly. The program was set to run 10 million generations sampling every 1000 steps. TreeAnnotator found the best tree discarding the first 10% as burn-in and annotating only nodes with frequency of 0.5 in all trees.

Simulation of 16S rRNA sequence evolution

We used Seq-Gen (Rambaut and Grass, 1997) to generate 16S rRNA sequences for method comparison and validation purposes. The input file was a Phylip phylogenetic tree (Supplementary Figure S7) constructed using the Matlab programming environment. The tree consisted of two sequence groups with sequence similarity of 90% within each group (we varied divergence values from common ancestor from 0.1% to 1%). The command 'seq-gen -mHKY -l1500-a0.15 -t2.0 <Mytree.txt >Result.txt' generated 998 sequences with a length of 1500 λbp using Hasegawa, Kishino and Yano model for nucleotide substitution and gamma distribution to allow for more site-specific rate heterogeneity (with γ parameter describing the shape of γ distribution, estimated to be 0.15 using 30 randomly selected sequences from *Lachnospira* group of the Ley human data). The simulated sequences using various divergence values from common ancestor were analysed using out- and phylogenetic tree-based methods and our AIBIMM approach.

Results

The data set published by Ley *et al*. (2006) was used as a master data set in the construction of a coordinate-based phylogeny. The rationale was the extensive size of this data set, and that the data include temporal analyses of the microbiota in 12 individuals. Figure 2a shows the PCA plot of gene phylogeny. Each data point represents one 16S rRNA clone sequence. The main phylogroups identified are: *Bacteroidetes*, *Firmicutes* (*Faecalibacteria* and *Lachnospiraceae*) and *Actinobacteria* (separated along principal component 3, not shown here) and some minor phylogroups, such as *Akkermansia*, *Eubacterium* and *Subdoligranulum* (a complete list of taxonomic assignments is given in Supplementary Table S4).

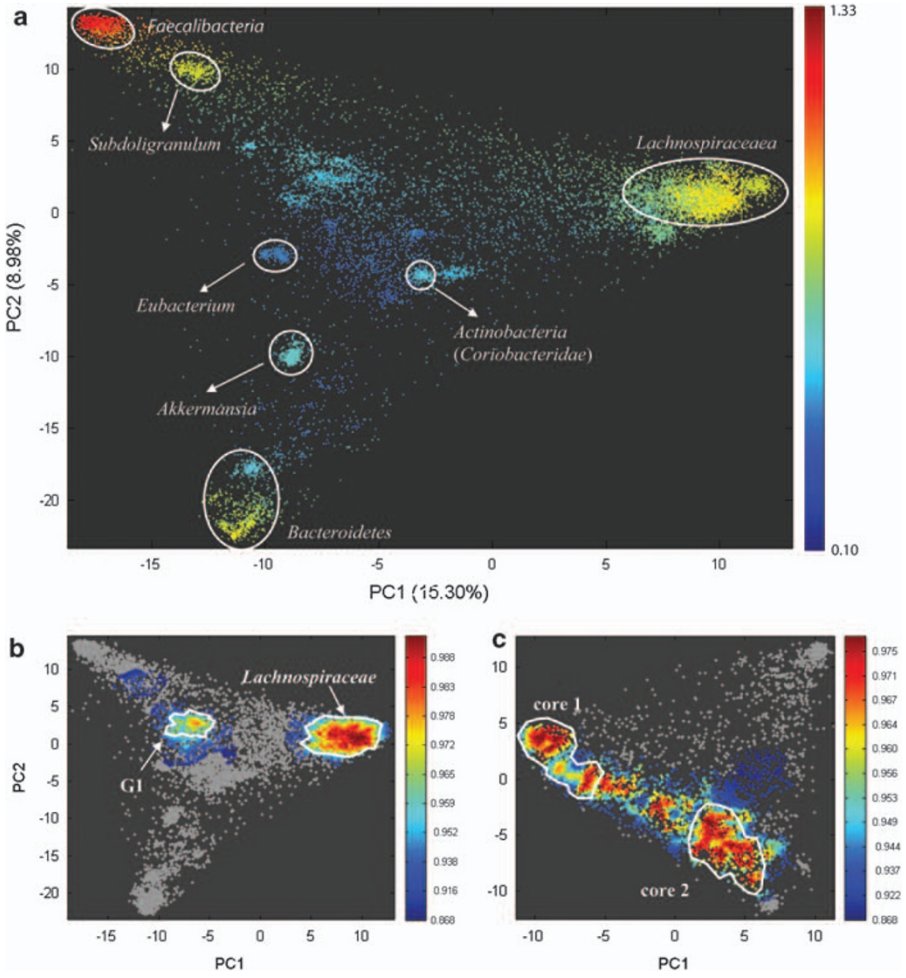

Figure 2 PCA plot of gene phylogeny constructed using AIBIMM method, here presented by the first two principal components of the model. Each data point corresponds to one *16S rRNA* gene clone. The labels describing different phylogroups are for descriptive purposes only (no real/statistical taxonomic group assignment). Supplementary Table S4 gives an overview of sequence names, their best matches in RDP10 database and PCA coordinates. (**a**) The data points are coloured according to their explained variance, that is, how well the PCA model is describing them. The explained variance for a data point is high when its model approximation value complies well with its observed value. (**b**) Grid-based search reveals potential core areas. The data points having *cp* value below cp_{mean} are greyed out. The colour scale ranges from red (cp_{max}) to blue ($cp \approx cp_{mean}$). *Lachnospiraceae* core group (40% of the total data points count) was extracted as a sub-data set and further analysed to find more specific core groups, cores 1 and 2, as shown in (**c**).

This data set was divided into 12 individual-specific data sets and the proposed grid-based data mining approach was applied to each subdata set separately (Supplementary Figure S1). Two bacterial clusters, G1 and *Lachnospiraceae*, were found to be stable over time within each individual and common for all individuals as shown in Figure 2b. Owing to high density of bacterial clones in *Lachnospiraceae* bacterial group, we performed a new PCA only on *Lachnospiraceae* subdata set and applied grid-based search with resampling to investigate whether we could find more specific core bacterial phylogroups within this group. Two separate core phylogroups, *Lachnospiraceae* cores 1 and 2 were identified (Figure 2c).

Sequence similarity within each of the potential bacterial core phylogroups and two of the other major (non-core) phylogroups, *Bacteroidetes* and *Faecalibacteria*, are summarized in Table 2. The sequence similarity comparison of bacteria within G1 bacterial group showed deep branching (more than 20% sequence divergence). Owing to its deep phylogenetic divergence and its low explained variance in the PCA model of gene phylogeny, G1 group was not analysed further in this work. An additional analysis of the structure of phylogroups *Lachnospiraceae* cores 1 and 2 was carried out using alignment-based phylogenetic reconstruction (Figure 3). This analysis showed that core 1 was a relatively shallow monophyletic phylogroups, while core 2 was deeper branching, having core 1 as a subgroup. We also found that one of the *Lachnospiraceae* outgroup (outgroup 2 in Figure 3) was within core 2.

We applied two approaches in estimating cores 1 and 2 divergence times. Using the pairwise distances approach, we observed that the histogram of pairwise distances within cores 1 and 2 had a major

Table 2 Sequence similarity within each phylogroup in the Ley human data

Bacterial phylogroup	*Sequence similarity (%)*	*No. of sequences*	*Rel. abundance (%)*
Lachnospiraceae	⩾85.28	5764	39.94
Core 1	⩾93.99	1854	12.85
Core 2	⩾89.54	1457	10.1
G1	⩾78.01	693	4.8
Faecalibacteria	⩾95.36	774	5.36
Bacteroidetes	⩾85.97	879	6.09
Total data set	⩾72.28	14 431	100

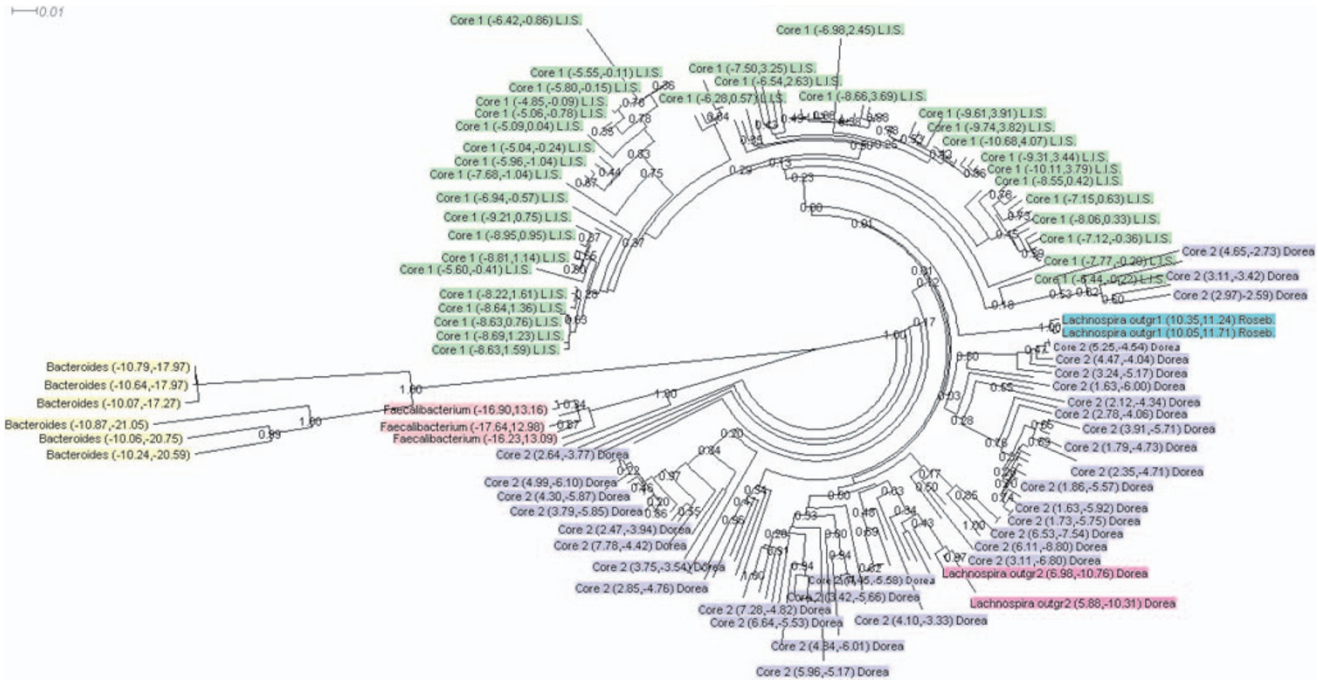

Figure 3 Phylogenetic reconstruction of sequences belonging to cores 1 and 2. One hundred sequences were chosen randomly from each core phylogroup, three from *Faecalibacteria* and *Bacteroides* and six from each of the manually derived bacterial outgroups: *Lachnospiraceae* outgroup 1 (yellow) and *Lachnospiraceae* outgroup 2 (blue). Sequence alignment was performed using RDP10 aligner (Cole *et al*., 2009). The neighbour-joining tree was constructed using program Mega 4.1, and 12 with default parameters. Dendroscope (Huson *et al*., 2007) made it possible to draw the tree as circular phylogram and to colour the labels. Detailed description is available in Supplementary Figure S5. The name of the each sequence consists of the phylogroup name, coordinates in PCA plot Figure 2a and closest match in RDP10 database. The majority of core 2 sequences have highest match to *Dorea* in RDP10 database, while most core 1 sequences are described either as *Lachnospiraceae Incertae Sedis* (LIS) or uncultured *Lachnospiraceae*. Representatives from *Lachnospiraceae* outgroup 1 shared high similarity to *Roseburia* (Roseb).

peak at 5% and 9%, respectively (Supplementary Figure S2). The divergence time for the two cores was then estimated using divergence rate for 16S rRNA of 1% per 50 million years (Ochman and Wilson, 1987). This corresponds to divergence time at about 250 and 450 million years ago for cores 1 and 2, respectively. The second approach was conducted using *Lachnospira* sequences from termites and chickens. The termites were chosen to reflect invertebrates and chickens to indicate vertebrates. The pairwise distances analyses showed that there was a major peak for chickens at 6% and termites at 10% (Supplementary Figure S8). The sequences positioned adjacent to the core phylogroups when projected onto the PCA model of the Ley human data were used as calibration points for relaxed molecular clock analyses (see Materials and methods for details). The divergence times estimated using this approach were 250 and 480 million years for cores 1 and 2, correspondingly.

A characteristic of the core microbiota is that it should be stable over time in a given individual, and that the core phylogroup should be widely distributed among individuals. We therefore investigated the temporal stability in the Ley human data. We compared the stability of *Lachnospiraceae* group and cores 1 and 2 with the common bacterial phylogroups *Bacteroides* and *Faecalibacteria*. These were chosen because of the high abundance in the data (Figure 2a). The core phylogroups had a significantly ($P<0.05$) lower variance in bacterial abundance among individuals than the other

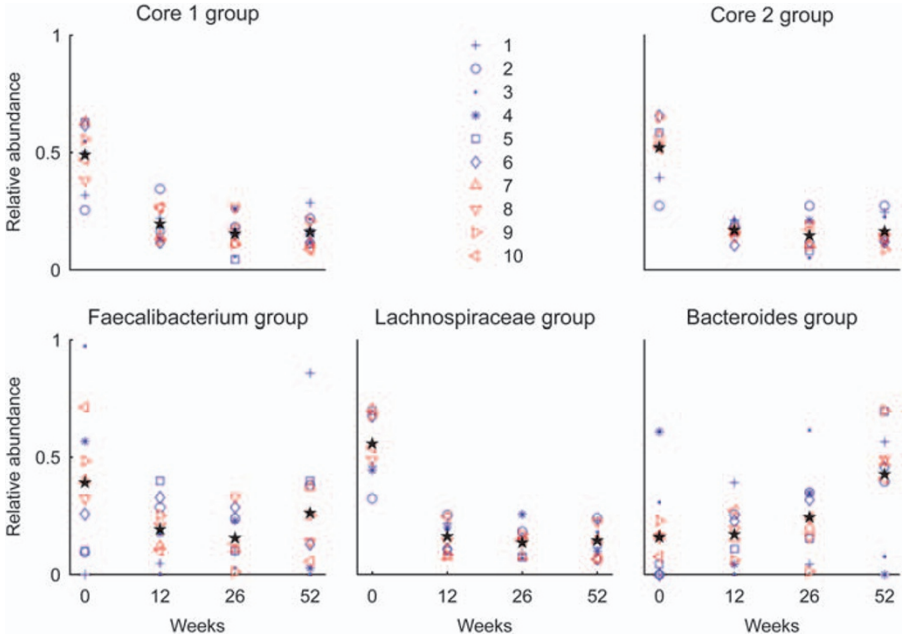

Figure 4 Relative abundance of the five bacterial phylogroups within each individual for each time point. The normalization was performed per individual per phylogroup (Equation (4)). The mean value for each time point for each phylogroup is outlined with the star symbol. The colouring distinguishes between the two diet-treatment groups: fat restricted (blue) and carbohydrate restricted (red).

dominant phylogroups (Supplementary Table S1). This was also reflected when analysing each time point separately (Figure 4 and Supplementary Table S2). This means that the bacterial abundance among individuals is significantly more similar in the core compared with non-core bacterial phylogroups, and that the core phylogroups are significantly more stable over time. Finally, the abundance of bacteria from both cores 1 and 2 showed a consistent drop with the diet—independent of nutrient (fat/carbohydrate) restriction.

We investigated the spatial distribution of cores 1 and 2 within the gut using the Eckburg data (Table 1). Both cores 1 and 2 were found in all six subdivisions of the human colon: cecum, ascending colon, transverse colon, descending colon, sigmoid colon and rectum, as well as in the stool samples. The pervasiveness of core 1 in mucosal samples was significantly higher ($P<0.01$) (Supplementary Figure S3) compared with stool samples. The pervasiveness of core 2 phylogroup did not differ significantly between mucosal and stool samples.

We used the Palmer data (Table 1) to investigate the first appearance of core phylogroups after birth. This data set was used only as a reference by Palmer *et al.* (2007), thus it does not include a complete coverage of the individuals analysed. Nevertheless, this is, to our knowledge, the only full-length 16S rRNA data set available on infant microbiota and was therefore included in our analysis. Core 1 was found in only 1, while core 2 was found in five of seven babies sampled. The babies whose faeces did not contain core 2 bacteria were only sampled when they were $\leqslant 2$ and 12 days old, respectively. Core 1 bacteria were not observed in infants younger than 7 months. Neither of the two core phylogroups was found in milk or vaginal swabs collected from one of the mothers. In this data set, 16S rRNA sequences were provided from two mothers immediately after they gave birth. Interestingly, they were the only adults in all downloaded data sets (Table 1) who lacked core 1. On the other hand, both mothers contained bacteria belonging to core 2.

The ethnic distributions of cores 1 and 2 were analysed using a Chinese data set (Table 1). The subjects selected for this study belonged to a four-generation Chinese family: child (1.5 years old), mother, father, uncle, grandmother, grandfather and great-grandmother. We found both core phylogroups in all members of the Chinese family. There was no significant correlation between the age of the family members and the abundance of the core bacteria.

Twin data were analysed using the Turnbaugh data set (from full-length 16S rRNA sequence-based survey) comprised of stool samples from 10 adult monozygotic/dizygotic twin pairs and their mothers. The twins were of European or African ancestry, were born in the same town in USA, but now lived throughout the country. Both core phylogroups were found in all individuals analysed.

The stability towards external perturbations was investigated for a data set concerning the effect of antibiotic treatment on the human gut microbiota. Using the comprehensive Dethlefsen data (Table 1), we observed that the antibiotics treatment did not influence significantly the frequency of the bacteria in the two core phylogroups. All individuals in this data set contained bacteria from both cores 1 and 2.

The distribution within a human population was determined by analysis of the Turnbaugh V2 data from 154 individuals. This analysis showed that the population distributions of cores 1 and 2 were log normal, while the distributions of non-core phylogenetic groups *Bacteroidetes* and *Faecalibacteria* were negatively skewed compared with the log-normal distribution (Supplementary Figure S10).

To address the co-evolution between mammalian gut microbiota, we investigated the distribution of the core phylogroups 1 and 2 among mammals (including humans) with the additional information from the Ley mammal data (Table 1). The core group 1 was found in Lagomorpha ($n=1$) and Xenarthra ($n=1$), 96% of the humans ($n=56$), 79% of the non-human Primates ($n=19$), 67% of Rodentia ($n=3$), 50% of Chiroptera ($n=2$), 47% of Carnivora ($n=17$), 19% of Artiodactyla ($n=21$), 14% of both Perissodactyla ($n=7$) and Proboscidae ($n=7$) and none in the other mammalian orders included in the study. The core 2 group on the other hand was found in all humans ($n=56$), Proboscidae ($n=7$), Xenarthra ($n=1$), Diprotodontia ($n=2$), Hyracoidea ($n=2$), Lagomorpha ($n=1$), and Monotremata ($n=1$), 95% of Artiodactyla (cloven-hooved mammals), 89% of the non-human Primates, 71% of Perissodactyla, 67% of Rodentia, 50% of Chiroptera, 47% of Carnivora and none in Insectivora ($n=1$).

Genome-sequenced strains were projected onto cores 1 and 2 and none of them were within the core phylogroups. Eubacterium eligens (strain ATCC 27750), however, showed homology to core 2 (Supplementary Figure S4). The potential functionality of the core phylogroups was further evaluated by comparison with known butyrate-producing bacteria. We discovered that one of the butyrate-producing bacteria isolated from human colon (Louis *et al.*, 2004) belongs to core 1 and several to core 2. From the literature, however, we also identified non-butyrate-producing bacteria within core 2 (Taras *et al.*, 2002).

To validate our core search findings, we reanalysed the *Lachnospira* subset using an OTU approach. Here, we used Mothur to search for core microbiota in the *Lachnospiraceae* group (Figures 2b and c). This group contains both core phylogroups found by our search algorithm. Mothur found only a part of core 1 at distances <0.15, a part of core 2 at distances between 0.11 and 0.17 and most *Lachnospiraceae* at larger distances (Supplementary Figure S13).

We used simulated data with a known evolutionary history as described in Materials and methods, testing the effect of divergence between the two groups on the accuracy of group assignments by AIBIMM-, OTU- and tree-based methods. With a sequence similarity of 90% within each group and 0.3% distance from the root and below (Supplementary Table S3), OTU-based and phylogenetic tree building methods resolved the two bacterial groups poorer than the PCA plot of gene phylogeny using AIBIMM approach ($P<0.01$) (Supplementary Figure S11). For larger distances from the root ($>0.8\%$), all three methods accurately separated the two groups.

As a final methods validation, we simulated the distribution of bacterial sequences among 12 individuals for core and non-core group resembling the observed individual clone distribution within core 2 and outgroup 1, respectively. The results from core search using our approach and the Mothur's shared sequence search module is summarized in Supplementary Figure S12. In general, our approach had lower false-positive ($P\approx0.16$) and false-negative ($P\approx0.15$) rates compared with Mothur's approach (two-sample *t*-test). A further challenge with the latter approach was that the best separation was achieved at different OTU depths, although the same parameters were used for all 10 data sets during simulation.

Discussion

The strongest claim for the lack of a core phylogroups was made by Turnbaugh *et al.* (2009), suggesting that there were shared core functions of the microbiota among individuals, rather than a taxonomic core. Recent reanalysis of this data set both by us and Turnbaugh *et al.* (2010) revealed that there were in fact stable phylogroups within the data set. Turnbaugh *et al.* (2010) have now found an unclassified *Lachnospiraceae* taxonomic group that was present in 99% of the individuals, whereas we found that both cores 1 and 2 were present in all individuals.

Other discoveries of a human gut core microbiota were made by Rajilic-Stojanovic *et al.* (2009). Using a microarray approach, they uncovered that certain bacterial groups are shared among 10 adult individuals at genus level (approx. 90% sequence similarity) and that *Clostridium* cluster XIVa was the most frequent core phylogroup. The two bacterial core phylogroups identified by our methodology belong to *Clostridium* cluster XIVa (Cotta and Forster, 2006), thus our results support that of Rajilic-Stojanovic *et al.* (2009), but at a larger scale and higher resolution. Another recent gut microbiota investigation was conducted by Claesson *et al.* (2009). They investigated the gut microbiota of four adult subjects with two high-throughput culture-independent methods, pyrosequencing and HITChip, and revealed core gut microbiota in many different bacterial phylogroups, including *Clostridium* cluster XIVa. *Dorea*, related to core 2, and a few other genera were found to be shared in more that 50% of faecal samples collected from 17 adult individuals (Tap *et al.*, 2009). Thus, the existence of core gut microbiota is now being confirmed by different research groups.

The phylogenetic depths of cores 1 and 2 suggest that they are both ancient phylogroups. Moreover,

core 2 corresponds to the development of jawed vertebrates and adaptive immunity approximately 500 million years ago (Pancer and Cooper, 2006). Adaptive immunity certainly represented a bottleneck for host/bacterial interactions, enabling the host to establish better control of the microbiota. It is not unlikely that the adaptive immune system triggered the development of new host/bacterial co-evolution. With respect to core 1, we estimated that this co-evolution developed 250 million years ago, with the emergence of the first mammals. In addition, the current distribution suggests that core 1 is over-represented in Euarchontogles, indicating co-evolution within the lineages leading to rodents and primates (Gibbs *et al.*, 2004).

The widespread and dominant association of the two core phylogroups with humans and other mammals suggests either a mutualistic or commensal interaction. Pathogenic or parasitic interactions are not likely owing to fitness loss of the host (Moran, 2006). The log-normal population distribution among individuals for the core phylogroups suggests that these are controlled by multiplicative rather than additive effects (Limpert *et al.*, 2001). As the distribution pattern for the core was different from the non-core, we believe that this information can be important for future understanding of the functionality of the core microbiota.

It has recently been found that under-representation of the broader group *Lachnospiraceae* correlates with inflammatory bowel disease, indicating an essential role of this phylogroup in maintaining gut homeostasis (Frank *et al.*, 2007). Further supporting an essential role of *Lachnospiraceae* comes from a mouse colon cancer model, in which this phylogroup apparently prevented polyp formation (Mai *et al.*, 2007). It is known that *Lachnospiraceae* are common butyrate producers in the human gut (Cotta and Forster, 2006), and the association between butyrate and cancer protection is well established (Roediger, 1990; Medina *et al.*, 1998; Orchel *et al.*, 2005; Comalada *et al.*, 2006; Tan *et al.*, 2008). Unfortunately, no genome-sequenced bacteria are within core 1 or 2, and the closest core 2 relative *Eubacterium eligens* is not very well characterized phenotypically (Mahowald *et al.*, 2009). Thus, the functionality of cores 1 and 2 cannot be deduced from genome sequences. The limited data set of butyrate-producing bacteria isolated from human colon (Louis *et al.*, 2004) provides evidence that at least some of the bacteria in core 1 and 2 groups are able to produce butyrate. However, 16S rRNA sequence match to all available bacteria in RDP10 (Cole *et al.*, 2009), and revealed that a few of the core 2 bacteria have similarity to the genus *Dorea*, which contains representatives that are not butyrate producing (Taras *et al.*, 2002). Given a butyrate-producing role, the abrupt relative drop in bacteria belonging to the core 1 and 2 phylogroups caused by energy restriction diet observed in the Ley human data (Figure 4) was surprising. Nonetheless, as only samples from faeces were collected, we cannot really determine whether the drop is caused by a diet-induced change in gut transit time (Brodribb *et al.*, 1980) or an actual change in the composition of the gut microbiota.

Interestingly, we found a very low prevalence of *Lachnospiraceae* among infants, and core 1 was not detected until 7 months of age. In addition, the two adults out of 210 with undetectable levels of core 1 were women who gave samples instantly after giving birth. Assuming that the main function of *Lachnospiraceae* is butyrate production, this corresponds with the fact that this specific acid is found in very low levels among infants (Wolin *et al.*, 1998). Butyrate has a dichotomy effect, with low levels leading to cell proliferation and cancer risk, while high levels are toxic and can lead to gut leakage (Peng *et al.*, 2007). For preterm infants, butyrate has been associated with necrotic enterocolitis. Therefore, butyrate-producing bacteria, such as those belonging to *Lachnospiraceae*, are probably not recruited until later in the infant life (Peng *et al.*, 2007).

In summary, our results show that that the two core phylogroups identified are ancient, abundant and stable in the human gut. Linked with previous knowledge, these results also suggest that the core phylogroups are of fundamental importance for human health and disease. The knowledge of such core phylogroups may offer opportunity to design adequate therapeutic strategies for the prevention and treatment of colon-related conditions and diseases.

Acknowledgements

This work was financially supported by the Foundation for Research Levy on Agricultural Products.

References

Alm JS, Swartz J, Lilja G, Scheynius A, Pershagen G. (1999). Atopy in children of families with an anthroposophic lifestyle. *Lancet* **353**: 1485–1488.

Blair JE, Hedges SB. (2005). Molecular phylogeny and divergence times of deuterostome animals. *Mol Biol Evol* **22**: 2275–2284.

Brodribb J, Condon RE, Cowles V, DeCosse JJ. (1980). Influence of dietary fiber on transit time, fecal composition, and myoelectrical activity of the primate right colon. *Digest Dis Sci* **25**: 260–266.

Claesson MJ, O'Sullivan O, Wang Q, Nikkilä J, Marchesi JR, Smidt H *et al.* (2009). Comparative analysis of pyrosequencing and a phylogenetic microarray for exploring microbial community structures in the human distal intestine. *PloS One* **4**: e6669.

Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ *et al.* (2009). The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res* **37**: D141–D145.

Comalada M, Bailon E, de Haro O, Lara-Villoslada F, Xaus J, Zarzuelo A *et al.* (2006). The effects of short-chain fatty acids on colon epithelial proliferation and survival depend on the cellular phenotype. *J Cancer Res Clin Oncol* **132**: 487–497.

Cotta M, Forster R. (2006). The family *Lachnospiraceae*, including the genera *Butyrivibrio*, *Lachnospira* and *Roseburia*. *Prokaryotes* **4**: 1002–1021.

Dethlefsen L, Huse S, Sogin ML, Relman DA. (2008). The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLoS Biol* **6**: e280.

Doolittle WF, Zhaxybayeva O. (2009). On the origin of prokaryotic species. *Genome Res* **19**: 744–756.

Drummond A, Rambaut A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* **7**: 214.

Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M *et al.* (2005). Diversity of the human intestinal microbial flora. *Science* **308**: 1635–1638.

Felsenstein J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**: 783–791.

Frank DN, St Amand AL, Feldman RA, Boedeker EC, Harpaz N, Pace NR. (2007). Molecular–phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc Natl Acad Sci USA* **104**: 13780–13785.

Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S *et al.* (2004). Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**: 493–521.

Huson DH, Richter DC, Rausch C, Dezulian T, Franz M, Rupp R. (2007). Dendroscope: an interactive viewer for large phylogenetic trees. *BMC Bioinform* **8**: 460.

Khachatryan ZA, Ktsoyan ZA, Manukyan GP, Kelly D, Ghazaryan KA, Aminov RI. (2008). Predominant role of host genetics in controlling the composition of gut microbiota. *PloS One* **3**: e3064.

Ley RE, Hamady M, Lozupone C, Turnbaugh PJ, Ramey RR, Bircher JS *et al.* (2008). Evolution of mammals and their gut microbes. *Science* **320**: 1647–1651.

Ley RE, Turnbaugh PJ, Klein S, Gordon JI. (2006). Microbial ecology: human gut microbes associated with obesity. *Nature* **444**: 1022–1023.

Li M, Wang B, Zhang M, Rantalainen M, Wang S, Zhou H *et al.* (2008). Symbiotic gut microbes modulate human metabolic phenotypes. *Proc Natl Acad Sci USA* **105**: 2117–2122.

Limpert E, Stahel WA, Abbt M. (2001). Log-normal distributions across the sciences: keys and clues. *BioScience* **51**: 341–352.

Liu K, Raghavan S, Nelesen S, Linder CR, Warnow T. (2009). Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science* **324**: 1561–1564.

Louis P, Duncan SH, McCrae SI, Millar J, Jackson MS, Flint HJ. (2004). Restricted distribution of the butyrate kinase pathway among butyrate-producing bacteria from the human colon. *J Bacteriol* **186**: 2099–2106.

Lozupone C, Hamady M, Knight R. (2006). UniFrac—an online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinform* **7**: 371.

Lozupone C, Knight R. (2005). Unifrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* **71**: 8228–8235.

Mahowald MA, Rey FE, Seedorf H, Turnbaugh PJ, Fulton RS, Wollam A *et al.* (2009). Characterizing a model human gut microbiota composed of members of its two dominant bacterial phyla. *Proc Natl Acad Sci* **106**: 5859–5864 (E-pub ahead of print 24 March 2009).

Mai V, Colbert LH, Perkins SN, Schatzkin A, Hursting SD. (2007). Intestinal microbiota: a potential diet-responsive prevention target in ApcMin mice. *Mol Carcinogen* **46**: 42–48.

Medina V, Afonso J, Alvarez-Arquelles H, Hernandez C, Gonzalez F. (1998). Sodium butyrate inhibits carcinoma development in a 1,2-dimethylhydrazine-induced rat colon cancer. *J Parent Enteral Nutr* **22**: 14–17.

Moran NA. (2006). Symbiosis. *Curr Biol* **16**: R866–R871.

Ochman H, Wilson AC. (1987). Evolution in bacteria: evidence for a universal substitution rate in cellular genomes. *J Mol Evol* **26**: 74–86.

Orchel A, Dzierzewicz Z, Parfiniewicz B, Weglarz L, Wilczok T. (2005). Butyrate-induced differentiation of colon cancer cells is PKC and JNK dependent. *Digest Dis Sci* **50**: 490–498.

Palmer C, Bik EM, Digiulio DB, Relman DA, Brown PO. (2007). Development of the human infant intestinal microbiota. *PLoS Biol* **5**: e177.

Pancer Z, Cooper MD. (2006). The evolution of adaptive immunity. *Annu Rev Immunol* **24**: 497–518.

Peng L, He Z, Chen W, Holzman IR, Lin J. (2007). Effects of butyrate on intestinal barrier function in a caco-2 cell monolayer model of intestinal barrier. *Pediatr Res* **61**: 37–41.

Peterson KJ, Cotton JA, Gehling JG, Pisani D. (2008). The Ediacaran emergence of bilaterians: congruence between the genetic and the geological fossil records. *Philos Trans R Soc B* **363**: 1435–1443.

Rajilic-Stojanovic M, Heilig HGHJ, Molenaar D, Kajander K, Surakka A, Smidt H *et al.* (2009). Development and application of the human intestinal tract chip, a phylogenetic microarray: analysis of universally conserved phylotypes in the abundant microbiota of young and elderly adults. *Environ Microbiol* **11**: 1736–1751.

Rambaut A, Grass NC. (1997). Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput Appl Biosci* **13**: 235–238.

Roediger W. (1990). The starved colon—diminished mucosal nutrition, diminished absorption, and colitis. *Dis Colon Rectum* **33**: 858–862.

Rudi K, Zimonja M, Kvenshagen B, Rugtveit J, Midtvedt T, Eggesbo M. (2007). Alignment-independent comparisons of human gastrointestinal tract microbial communities in a multidimensional *16S rRNA* gene evolutionary space. *Appl Environ Microbiol* **73**: 2727–2734.

Rudi K, Zimonja M, Næs T. (2006). Alignment-independent bilinear multivariate modelling (AIBIMM) for global analyses of *16S rRNA* gene phylogeny. *Int J Syst Evol Microbiol* **56**: 1565–1575.

Saitou N, Nei M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**: 406–425.

Scanlan PD, Shanahan F, Clune Y, Collins JK, O'Sullivan GC, O'Riordan M *et al.* (2008). Culture-independent analysis of the gut microbiota in colorectal cancer and polyposis. *Environ Microbiol* **10**: 789–798.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB *et al.* (2009). Introducing mothur: open source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75**: 7537–7541.

Stewart JA, Chadwick VS, Murray A. (2005). Investigations into the influence of host genetics on the predominant eubacteria in the faecal microflora of children. *J Med Microbiol* **54**: 1239–1242.

Tamura K, Dudley J, Nei M, Kumar S. (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* **24**: 1596–1599.

Tamura K, Nei M, Kumar S. (2004). Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proc Natl Acad Sci USA* **101**: 11030–11035.

Tan H, Tan S, Lin Q, Lim T, Hew C, Chung M. (2008). Quantitative and temporal proteome analysis of butyrate-treated colorectal cancer cells. *Mol Cell Proteomics* **7**: 1174–1185.

Tap J, Mondot S, Levenez F, Pelletier E, Caron C, Furet J-P *et al.* (2009). Towards the human intestinal microbiota phylogenetic core. *Environ Microbiol* **11**: 2574–2584.

Taras D, Simmering R, Collins MD, Lawson PA, Blaut M. (2002). Reclassification of *Eubacterium formicigenerans* Holdeman and Moore 1974 as *Dorea formicigenerans* gen. nov., comb. nov., and description of *Dorea longicatena* sp. nov., isolated from human faeces. *Int J Syst Evol Microbiol* **52**: 423–428.

Turnbaugh PJ. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444**: 1027–1031.

Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE *et al.* (2009). A core gut microbiome in obese and lean twins. *Nature* **457**: 480–484.

Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. (2007). The human microbiome project. *Nature* **449**: 804–810.

Turnbaugh PJ, Quince C, Faith JJ, McHardy AC, Yatsunenko T, Niazi F *et al.* (2010). Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins. *Proc Natl Acad Sci USA* **107**: 7503–7508.

Wolin MJ, Yerry S, Miller TL, Zhang Y, Bank S. (1998). Changes in production of ethanol, acids and H2 from glucose by the fecal flora of a 16- to 158-d-old breast-fed infant. *J Nutr* **128**: 85–90.

Zoetandal EG, Akkermans ADL, Akkermans-van Vliet WM, de Visser JA, de Vos WM. (2001). The host genotype affects the bacterial community in the human gastrointestinal tract. *Microb Ecol Health Dis* **13**: 129–134.

Supplementary Information accompanies the paper on The ISME Journal website (http://www.nature.com/ismej)