www.sciencemag.org/cgi/content/full/science.1208344/DC1

Supporting Online Material for

Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes

Gary D. Wu,* Jun Chen, Christian Hoffmann, Kyle Bittinger, Ying-Yu Chen, Sue A. Keilbaugh, Meenakshi Bewtra, Dan Knights, William A. Walters, Rob Knight, Rohini Sinha, Erin Gilroy, Kernika Gupta, Robert Baldassano, Lisa Nessel, Hongzhe Li, Frederic D. Bushman,* James D. Lewis*

*To whom correspondence should be addressed. E-mail: gdwu@mail.med.upenn.edu (G.D.W.); lewisjd@mail.med.upenn.edu (J.D.L.); bushman@mail.med.upenn.edu (F.D.B.)

Published 1 September 2011 on *Science* Express
DOI: 10.1126/science.1208344

This PDF file includes:

Other supporting material for this report includes the following:

Materials and Methods

Human subjects.

Healthy volunteers between the ages of 18 and 40 were recruited to participate in the controlled feeding experiment. To be eligible, participants were required to be free from any chronic gastrointestinal disease, cardiac disease, diabetes mellitus or immunodeficiency diseases, to have a normal bowel frequency (minimum once every 2 days, maximum 3 times per day), to have body mass index (BMI) between 18.5 and 35. Participants could not have taken antibiotics within 6 months prior to enrollment, proton pump inhibitors, H2 receptor antagonists, tricyclic antidepressants, narcotics, anticholinergic medications, laxatives, or anti-diarrhea medications within 4 weeks of enrollment, or NSAIDs, dietary supplements, or antacids within 2 weeks prior to enrollment. Similar eligibility criteria were applied for our cross-sectional study except that the age range for participation was 2 years to 50 years and participants were required to have been weaned from nursing. All participants provided informed consent, or assent in the case of minors. Legal guardians provided informed consent for minors. Participants completed three 24-hour dietary recalls during the week before collection of stool samples to assess recent dietary composition. The third recall was performed for the day preceding collection of the first stool sample or beginning the inpatient stays for the controlled feeding experiments. Dietary

recalls were conducted by trained bionutritionists and nutrient intake was computed with the Nutrition Data System program (University of Minnesota). In addition, each participant or their guardian completed a food frequency questionnaire appropriate for the participant's age that assesses usual dietary composition over the preceding year. The complete list of nutrients studied is in Table S3 (Recall) and Table S4 (FFQ).

In the controlled feeding experiment, healthy volunteers were randomly assigned to either a high fiber/low fat diet or a low fiber/high fat diet. Each participant consumed identical meals for 10 consecutive days. The composition of the two study diets was identical. Only the portion sizes were modified to adjust the distribution of fat, carbohydrate, and protein in the diet. Total calories in the high fat were 38% from fat, 35% from carbohydrates, and 27% from protein. In the low fat diet, total calories were 13% from fat, 69% from carbohydrates, and 18% from protein. Portion sizes were also calculated based on the expected caloric requirements for the participant. Stool samples were collected daily and immediately frozen at -80^{0}C. Sigmoidoscopy without bowel cleansing was performed on the first and last day of the inpatient stay. Mucosal pinch biopsies were obtained from the rectum using large cup forceps and samples were flash frozen in liquid nitrogen. Participants consumed 24 x-ray-opaque markers at the time of entry into the research center. Abdominal x-rays were taken 1, 3, and 5 days later to quantify whole gut transit during this time period. Participants were not allowed to leave the clinical research unit without being accompanied by a

member of the research team. Participants were instructed to consume all food provided to them within each 24-hour period. Water, tea and coffee were provided ad lib but no sweetener or milk products could be added. The characteristics of the CAFE subjects are summarized in table S5, and the compositions of their diets are summarized in table S6.

DNA preparation

Stool or biopsy samples were stored at -80^{o}C prior to use. DNA was purified using the MoBio PowerSoil kit according to the manufacturer's instructions with addition of a high temperature heating step to improve lysis. DNA samples were amplified using V1-V2 region primers targeting bacterial 16S genes and sequenced using 454/Roche Titanium technology. DNA sequence reads from this study are available from the Sequence Read Archive (CaFE: SRX037803, SRX021237, SRX021236, SRX020772, SRX020771, SRX020588, SRX020587, SRX020379, SRX020378 (metagenomic). COMBO: SRX020773, SRX020770). Sample metadata (compliant with the MIMARKS standard) are available as tables S7 and S8 below.

Quantification of absolute levels of bacterial and human DNA.

We analyzed the total abundance of bacterial and human DNA in a subset of our samples using quantitative PCR assays for bacterial 16S rDNA gene segments

and the human beta-tubulin gene. We found on average $5x10^8$ 16S rDNA copies and ~300 beta-tubulin copies per microgram of total DNA in the 95 samples tested. Given some simple assumptions (mean of five 16S gene copies per bacterium and 5 Mb per bacterial genome), we calculate that bacterial DNA accounts for the majority of the mass of DNA in our stool samples, and the contribution of human DNA is several orders of magnitude less. The estimated number of bacteria detected using the 16S Q-PCR assay per gram of stool (wet weight) ranged from $2X10^8$ to $7X10^{10}$, a range that overlaps with earlier studies. Smaller proportions of bacterial DNA were detected in biopsy samples, consistent with a larger contribution of human DNA (data not shown). No differences in absolute bacterial abundance were detected between diets.

Statistical analysis.

Initial analysis was carried out using the QIIME pipeline (*14*)(http://qiime.sourceforge.net/) which implemented taxonomic assignment with RDP(*15*), and distances-based analysis using UniFrac(*16*). We used the default parameter settings in the QIIME pipeline, including Lane masking (*17*), and a single even-depth rarefaction analysis (depth = 2408), chosen to exclude only the lowest-depth sample, to control for sequencing effort (clustering analyses were not, however, sensitive to sequencing depth). The analysis produced weighted and unweighted UniFrac distance matrices and a table of per-sample genus counts.

We then performed clustering by partitioning around medoids (PAM)(*18*) using Jensen-Shannon divergence (JSD) of the normalized genus counts. Weighted UniFrac distance, Euclidean distance and Bray-Curtis distance of the normalized genus counts were also compared. The optimal number of clusters was chosen by the maximum average silhouette width, known as the silhouette coefficient (SC)(*19*). The quality of those clusters was assessed by the same measure, following the accepted interpretation that SC values above 0.5 indicate a reasonable clustering structure (*20*).

Lane masking was used in generating alignments for the analysis described above, which involves masking hypervariable regions to avoid artifacts due to convergent sequence evolution. Results without Lane masking were compared and found to be generally parallel, showing strong support for two clusters dominated with *Bacteroides* and *Prevotella*, and weak support in one case for a third cluster. Results were compared in fig. S2.

Overall associations between dietary / demographic variables and microbiome compositions were assessed using PERMANOVA based on weighted or unweighted UniFrac distances. Associations between dietary nutrients and individual taxa proportions were assessed by Spearman's rank correlation test while associations between enterotypes and individual taxa proportions or dietary nutrients were assessed by Wilcoxon rank sum test. False discovery rate (FDR) control was used to account for multiple comparisons when evaluating these

associations. Nutrient intake was normalized using the residual method to standardize for caloric intake. Quantities of nutrients were standardized over all 98 samples to have mean=0 and standard deviation=1.

For the shotgun metagenomic analysis, reads were aligned using BLAST and annotated using KEGG. We used a two-sample t test to compare the changes for each functional category between diets. All category counts were converted to relative abundance before performing the analysis. Columns were normalized to mean 0 and standard deviation 1. P values for shotgun metagenomic analysis were not corrected for multiple comparisons. Most statistical analysis was conducted using the statistical software package R.

Additional analysis

Association of gut microbial composition with BMI

We investigated the relationship of microbiome data to demographic data using PERMANOVA (table S1). Among the variables tested, BMI was among the most strongly associated with microbiome composition (p=0.001 unweighted analysis; p=0.145 weighted analysis). Additional demographic variables achieving significance included sex, race and consumption of yogurt and alcohol. Presence of pets, appendectomy history and drinking of tap water achieved marginal significance. Since weighted UniFrac analysis puts relatively more weight on common taxa and unweighted analysis puts more weight on rare taxa, some differences were observed between these two types of analysis, as expected. In general, quantitative and qualitative metrics provide different insights into relationships between community structure and external factors that may be important(*16*).

We compared short-term and long-term diet using two types of questionnaires. A validated food frequency questionnaire ("FFQ") queried usual (long term) diet. Three 24 hour dietary recalls within 7 days of sample collection were used to assess recent food intake ("Recall"). The dietary patterns measured between the two were partially correlated, so we first used principal component analysis to

summarize the diets. The 154 nutrients for Recall and 214 nutrients for FFQ were used to generate five principal components, which explained 55% and 58% of the total variation, respectively. We then performed PERMANOVA to test the overall association between these 5 PCs and microbiome variation, where the microbiome data were summarized as weighted or unweighted pairwise UniFrac distances. The composition of both the recent diets ($p<0.001$ unweighted; $p=0.047$ weighted) and usual diets ($p=0.011$ unweighted; $p=0.003$ weighted) were associated with microbiome composition.

To test the independent association of diet and BMI on the composition of the gut microbiome, multivariate models were constructed including BMI, sex, race and the principle components for recent diet. Adjusting for sex and race had minimal impact on the associations of recent diet and BMI with the microbiome. In a model including only BMI and diet, the strength of the associations was less strong, suggesting that at least part of the association of BMI with microbiome composition was due to the influence of diet and vice versa (table S9). Similar results were observed when BMI and total fats or saturated fatty acid intake was included in the model.

Analysis of effects of red wine consumption

At FDR 25%, we identified five anthocyanidins of various forms and total anthocyanidins to be associated with microbiome composition (unweighted

UniFrac). In an American diet, the major sources of anthocyanidins are red wine and some fruits. These anthocyanidins showed weak to moderate correlations with red wine consumption (Spearman correlation 0.23 to 0.62). Red wine consumption is also associated with overall microbiome composition ($p = 0.003$; PERMANOVA, unweighted analysis). To determine whether anthocyanidins and red wine consumption are independently associated with the overall microbiome composition, we tested the association of anthocyanidins and red wine consumption with microbiome composition after adjustment for each other. Possibly due to the relatively weak correlation between anthocyanidin and red wine consumption, the strength of the associations were minimally attenuated in the adjusted analyses (table S10). Thus, red wine consumption and anthocyanidin consumption appear to be independently associated with the microbiome composition.

Application of published enterotype clustering methodology

To reproduce a previously published enterotype clustering methodology (*4*), we performed clustering by PAM using the square root of the Jensen-Shannon divergence, and chose the number of clusters by the Caliński-Harabasz (CH) index of the relative clustering quality as defined in the original publication of the method (*20*). The CH index indicated that three clusters were preferred, but the quality score for three clusters (SC=0.17) indicated no substantial structure. We also applied the CH Index to clustering using several alternative distance

measures (Bray-Curtis, Euclidean, Jensen-Shannon, weighted UniFrac, and weighted normalized UniFrac). Interestingly, in all but one case (weighted UniFrac) the CH index chose three as the optimal number of clusters, even though the silhouette scores were substantially higher for two clusters. No reasonable support (SC ≥ .5) for three clusters was found using any distance measure.

Prevotella-Bacteroides gradient analysis

The enterotype clustering is driven primarily by the ratio of the two dominant genera, *Prevotella* to *Bacteroides*; this ratio defines a clear gradient across the putative COMBO enterotypes (fig. S5; note that 69 samples with no *Prevotella* were excluded), emphasizing that the boundary between enterotypes is not sharply defined. When we removed these genera, the structure was undetectable (17 clusters, SC=0.115). Also, these genera compose between 12% and 83% of the relative abundance in the communities (mean +/- s.d. = 0.46 +/- 0.17), and the only distance measures that produced reasonable support for clustering, JSD and Euclidean distance, are measures that emphasize differences in the largest components of a distribution. These findings highlight the importance of these two genera.

Supplementary Figures

fig. S1. Associations between bacterial taxa and food groups. A) Heat map as in Fig. 1, but showing lower level phylogenetic assignments of bacterial taxa (right of the figure panel). Rows correspond to bacterial taxa, columns correspond to nutrients or nutrient classes. Colors represent the taxa-nutrient Spearman's correlations, where red color indicate positive association, blue color negative association, and * indicates the significant association at the false discovery rate (FDR) of 25%. The "F" or "R" following the label for each nutrient indicates the origin of the data from either FFQ or recall, respectively. Clustering was carried out using neighbor joining. B) A view of two-way clustering of Spearman's correlations between nutrients and microbiome taxa, where bacterial lineages are grouped by Phylum.

fig. S2. Enterotype clustering under different data processing and clustering methods. Clustering in the COMBO data was probed using several methods with and without Lane masking (in addition to the analysis in Figure 2 showing Jensen-Shannon clustering on data processed with Lane masking). For each silhouette scores provide a measure of the relative strength of clustering.

fig. S3. Bacterial taxa associated with each enterotype in the COMBO data. A) Proportional representation. The proportion of each genus is standardized using

z-score transformation to reflect the relative abundance. B) Absolute representation.

fig. S4. Analysis of CAFE samples using shot-gun metagenomics. A) Comparison of Phylum level proportions for the ten CAFE1 subjects determined using 16S tag sequencing (left) or shot-gun metagenomic sequencing (right). A total of 423 sequences from the metagenomic dataset were identified as containing an rRNA gene region and used in the analysis. The identified rDNA containing sequences, as well as the amplified sequences, were classified using RDP classifier (bootstrap cutoff 80%). B) Analysis of gene functional categories (KEGG) that changed in reciprocal directions depending on the diet. Red and blue bars indicate samples obtained from subjects fed a high fat/low fiber and low fat/high fiber diet, respectively. P values are not corrected for multiple comparisons. C) Domain level lineages detected in shot-gun metagenomic analysis of the controlled feeding study. Proportions of reads in KEGG categories detected in the controlled feeding samples are in table S11.

fig. S5. *Prevotella*-*Bacteroides* gradient.

The log ratio of the relative abundance of *Prevotella* to *Bacteroides* in a given sample is plotted against the first principal coordinate of that sample based on Jensen-Shannon divergence (Fig. 2). Samples are colored to match their enterotype assignments (green: *Bacteroides* predominant enterotype; red: *Prevotella* predominant). This plot excludes those subjects with zero *Prevotella*

abundance due to the undefined log ratio (69 of the subjects). The clear gradient across the enterotype boundary emphasizes that the boundary is not sharply defined.

Supplementary Tables

table S1. Demographic variables and their association with microbiome composition. P-values are not adjusted for multiple comparisons.

table S2. Association of nutrients and enterotype partitioning. The order of the nutrients is the same as in Figure 2C. Mean values of the normalized nutrient intake are given for both enterotypes with standard deviation indicated in the bracket. Unadjusted P values from Wilcoxon rank sum test as well as the FDR adjusted P values (Q values) are also listed.

table S3. Complete list of the dietary categories analyzed for Recall questionnaire and subject data.

table S4. Complete list of the dietary categories analyzed for the FFQ questionnaire and subject data.

table S5: Characteristics of participants in CAFE study.

table S6: Comparison of usual, recent and assigned diets of CAFE participants.

table S7: Metadata for COMBO samples, arranged according to the MIMARK standard.

table S8: Metadata for CAFE samples, arranged according to the MIMARK standard.

table S9: Multivariate models to test the interaction between BMI and diet on the strength of association with microbiome composition.

table S10. Association of anthocyanidin consumption with the microbiome composition after adjustment for red wine consumption and association of red wine consumption with the microbiome composition after adjustment for anthocyanidin. The * indicates the p value is for the association of red wine and the microbiome composition.

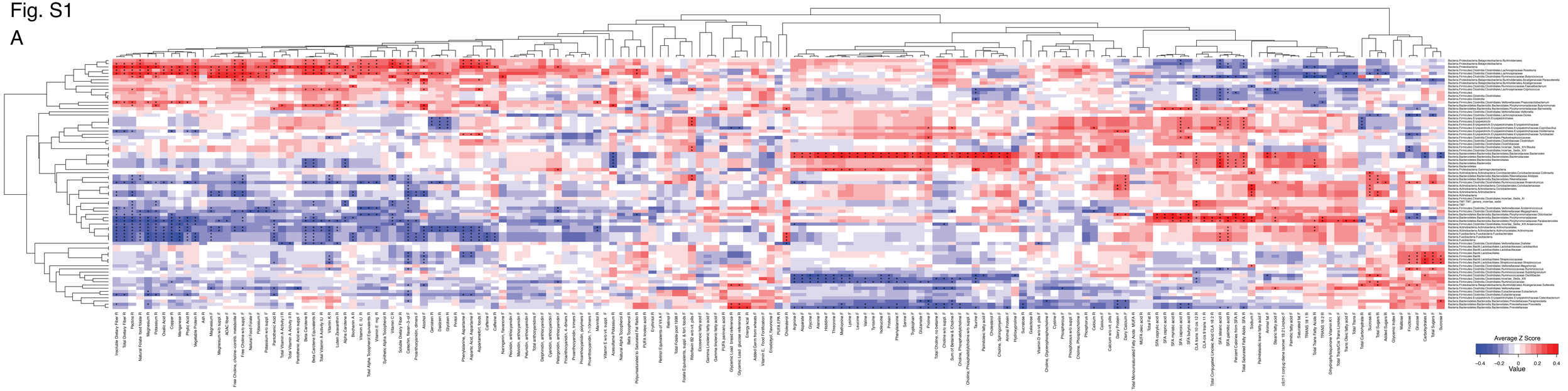
table S11. Proportions of reads in KEGG categories detected in the controlled feeding samples.

Supplementary References

14. J. G. Caporaso *et al.*, *Nat Methods* 7, 335 (2010).
15. Q. Wang, G. M. Garrity, J. M. Tiedje, J. R. Cole, *Applied and Environmental Microbiology* 73, 5261 (2007).
16. C. A. Lozupone, M. Hamady, S. T. Kelley, R. Knight, *Applied and Environmental Microbiology* 73, 1576 (2007).
17. D. J. Lane, in *Nucleic Acid Techniques in Bacterial Systematics,* E. Stackebrandt, M. Goodfellow, Eds. (John Wiley and Sons, Chichester, 1991), pp. 115-175.
18. L. Kaufman, P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis.*, (1990).
19. P. J. Rousseeuw, *J. Comput. Appl. Math.* 20, 53 (1987).
20. T. Calinski, J. Harabasz, *Communications in statistics* 1, 1 (1974).

Fig. S1

A


Average Z Score
−0.4
−0.2
0
0.2
0.4
Value

B

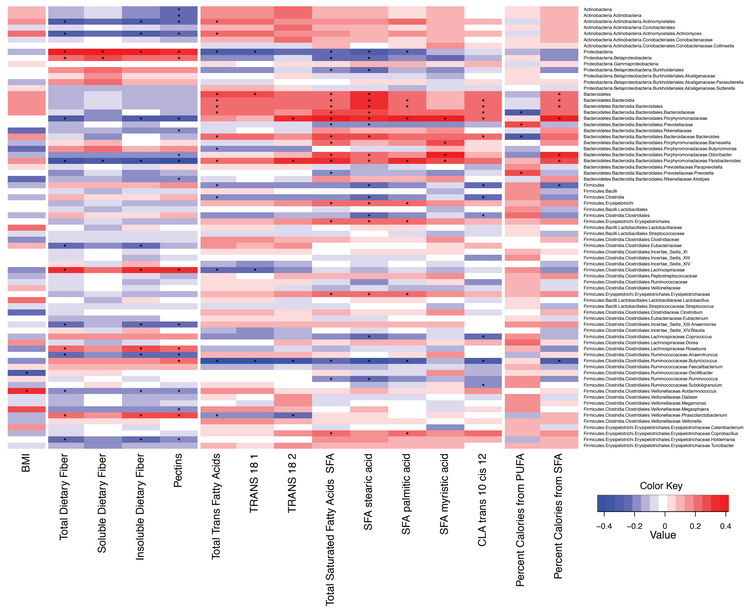

BMI
Total Dietary Fiber
Soluble Dietary Fiber
Insoluble Dietary Fiber
Pectins
Total Trans Fatty Acids
TRANS 18:1
TRANS 18:2
Total Saturated Fatty Acids
SFA stearic acid
SFA palmitic acid
SFA myristic acid
CLA trans 10 cis 12
Percent Calories from PUFA
Percent Calories from SFA
Color Key
−0.4
−0.2
0
0.2
0.4
Value

Fig. S2

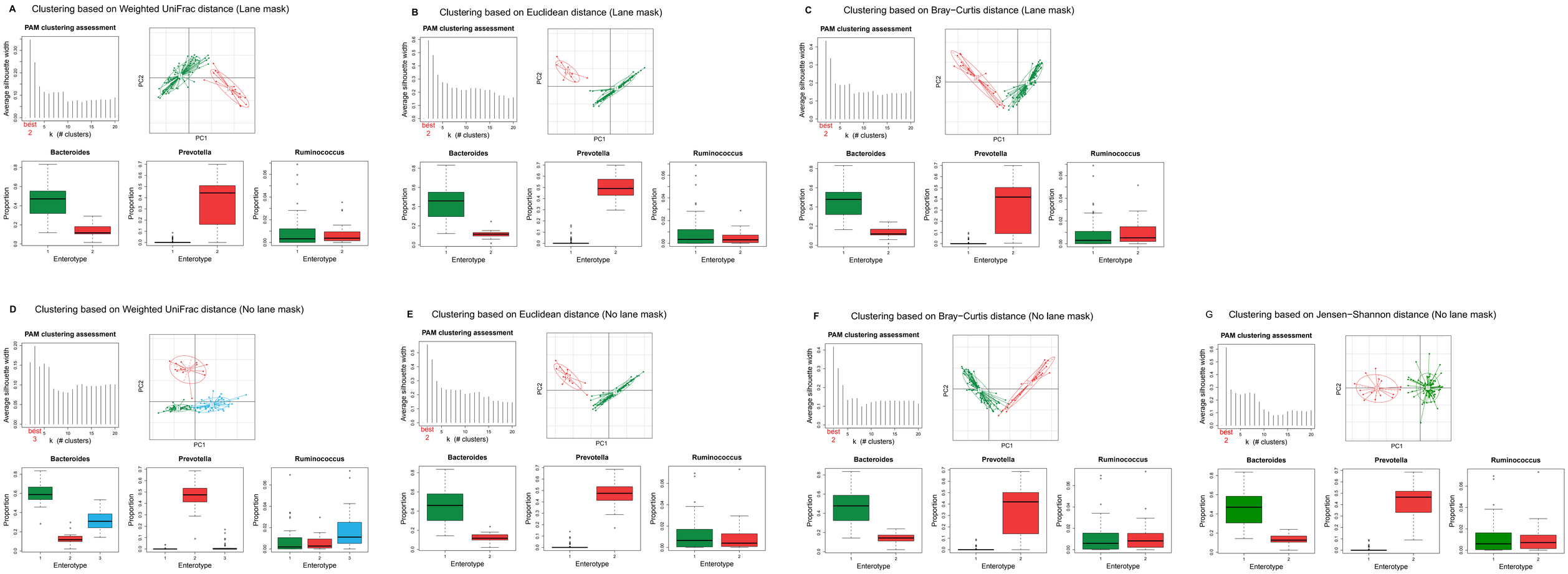

A
Clustering based on Weighted UniFrac distance (Lane mask)
B
Clustering based on Euclidean distance (Lane mask)
C
Clustering based on Bray−Curtis distance (Lane mask)
D
Clustering based on Weighted UniFrac distance (No lane mask)
E
Clustering based on Euclidean distance (No lane mask)
F
Clustering based on Bray−Curtis distance (No lane mask)
G
Clustering based on Jensen−Shannon distance (No lane mask)
PAM clustering assessment
Average silhouette width
k (# clusters)
best
PC1
PC2
Bacteroides
Prevotella
Ruminococcus
Proportion
Enterotype

Fig. S3

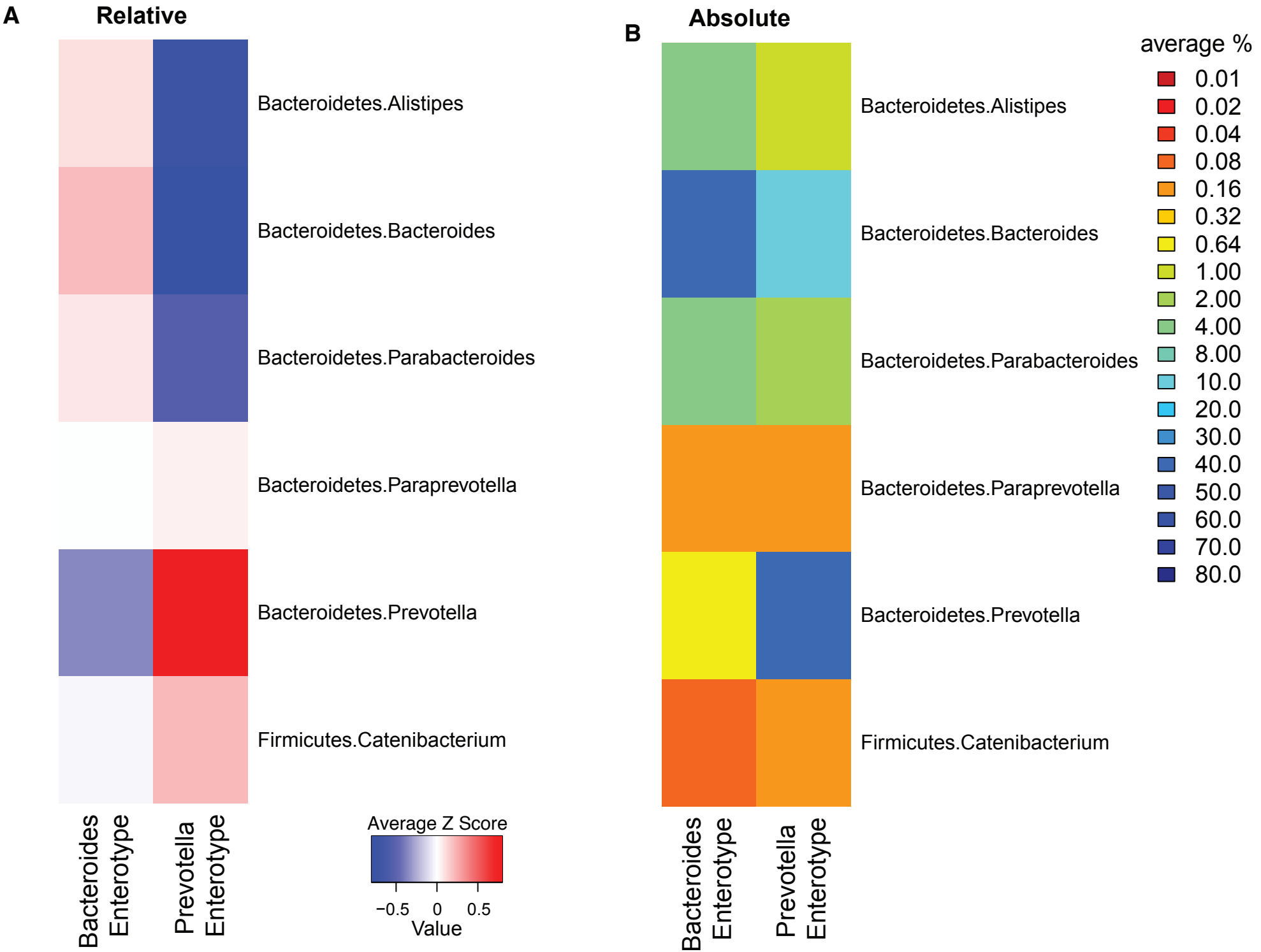

A
Relative
Bacteroidetes.Alistipes
Bacteroidetes.Bacteroides
Bacteroidetes.Parabacteroides
Bacteroidetes.Paraprevotella
Bacteroidetes.Prevotella
Firmicutes.Catenibacterium
Bacteroides Enterotype
Prevotella Enterotype
Average Z Score
−0.5
0
0.5
Value
B
Absolute
Bacteroidetes.Alistipes
Bacteroidetes.Bacteroides
Bacteroidetes.Parabacteroides
Bacteroidetes.Paraprevotella
Bacteroidetes.Prevotella
Firmicutes.Catenibacterium
Bacteroides Enterotype
Prevotella Enterotype
average %
0.01
0.02
0.04
0.08
0.16
0.32
0.64
1.00
2.00
4.00
8.00
10.0
20.0
30.0
40.0
50.0
60.0
70.0
80.0

Fig. S4

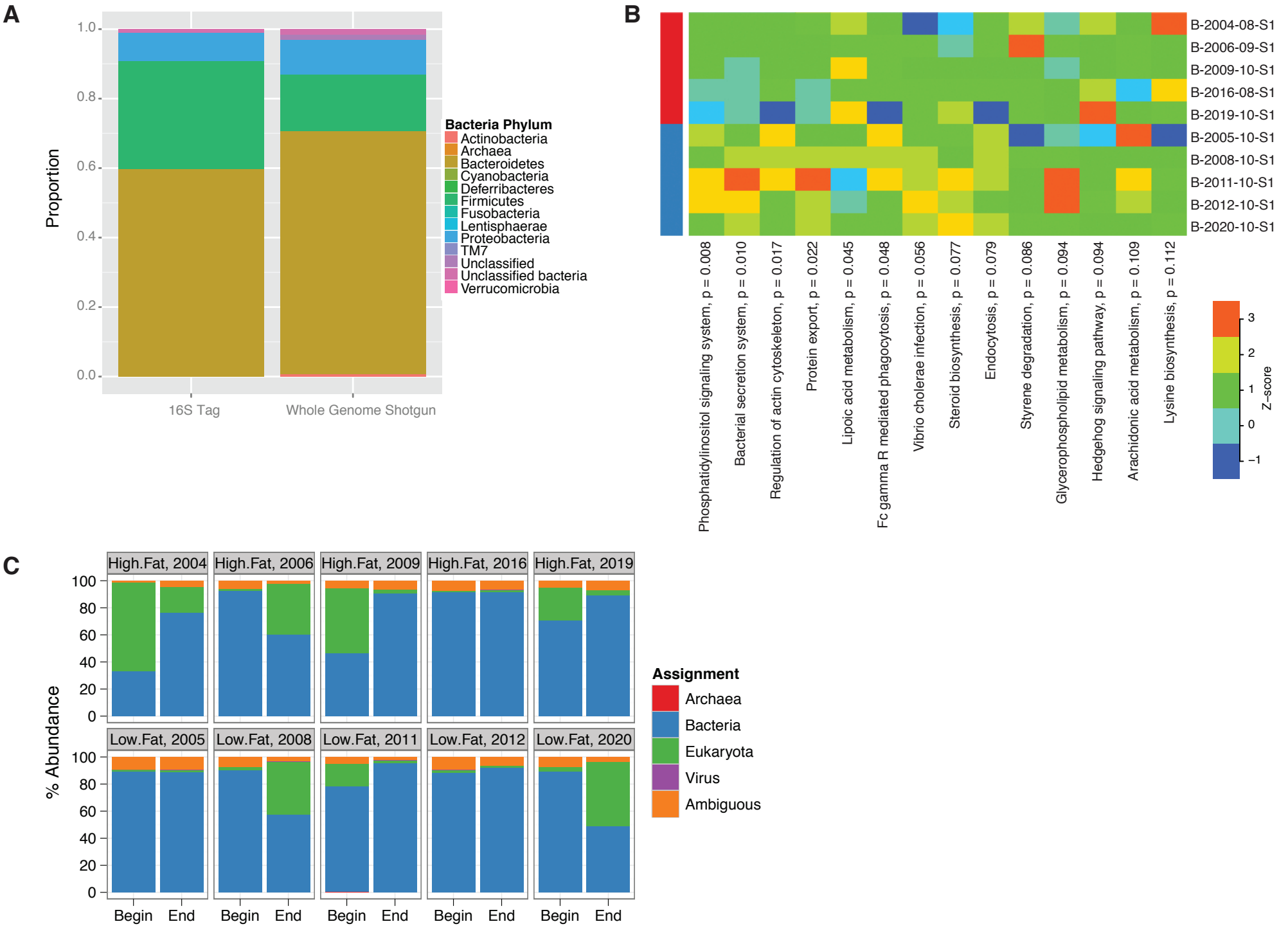

A
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
16S Tag
Whole Genome Shotgun
Bacteria Phylum
Actinobacteria
Archaea
Bacteroidetes
Cyanobacteria
Deferribacteres
Firmicutes
Fusobacteria
Lentisphaerae
Proteobacteria
TM7
Unclassified
Unclassified bacteria
Verrucomicrobia
B
B-2004-08-S1
B-2006-09-S1
B-2009-10-S1
B-2016-08-S1
B-2019-10-S1
B-2005-10-S1
B-2008-10-S1
B-2011-10-S1
B-2012-10-S1
B-2020-10-S1
Phosphatidylinositol signaling system, p = 0.008
Bacterial secretion system, p = 0.010
Regulation of actin cytoskeleton, p = 0.017
Protein export, p = 0.022
Lipoic acid metabolism, p = 0.045
Fc gamma R mediated phagocytosis, p = 0.048
Vibrio cholerae infection, p = 0.056
Steroid biosynthesis, p = 0.077
Endocytosis, p = 0.079
Styrene degradation, p = 0.086
Glycerophospholipid metabolism, p = 0.094
Hedgehog signaling pathway, p = 0.094
Arachidonic acid metabolism, p = 0.109
Lysine biosynthesis, p = 0.112
Z−score
3
2
1
0
−1
C
High.Fat, 2004
High.Fat, 2006
High.Fat, 2009
High.Fat, 2016
High.Fat, 2019
Low.Fat, 2005
Low.Fat, 2008
Low.Fat, 2011
Low.Fat, 2012
Low.Fat, 2020
% Abundance
100
80
60
40
20
0
Begin
End
Assignment
Archaea
Bacteria
Eukaryota
Virus
Ambiguous

Fig. S5