Title	Gut microbiota composition correlates with diet and health in the elderly
Authors	Claesson, Marcus J.;Jeffery, Ian B.;Conde, Susana;Power, Susan E.;O'Connor, Eibhlís M.;Cusack, Siobhán;Harris, Hugh M. B.;Coakley, Mairead;Lakshminarayanan, Bhuvaneswari;O'Sullivan, Orla;Fitzgerald, Gerald F.;Deane, Jennifer;O'Connor, Michael;Harnedy, Norma;O'Connor, Kieran;O'Mahony, Denis;van Sinderen, Douwe;Wallace, Martina;Brennan, Lorraine;Stanton, Catherine;Marchesi, Julian R.;Fitzgerald, Anthony P.;Shanahan, Fergus;Hill, Colin;Ross, R. Paul;O'Toole, Paul W.
Publication date	2012-07-13
Original Citation	Claesson, M. J., Jeffery, I. B., Conde, S., Power, S. E., O'Connor, E. M., Cusack, S., Harris, H. M. B., Coakley, M., Lakshminarayanan, B., O'Sullivan, O., Fitzgerald, G. F., Deane, J., O'Connor, M., Harnedy, N., O'Connor, K., O'Mahony, D., van Sinderen, D., Wallace, M., Brennan, L., Stanton, C., Marchesi, J. R., Fitzgerald, A. P., Shanahan, F., Hill, C., Ross, R. P. and O'Toole, P. W. (2012) 'Gut microbiota composition correlates with diet and health in the elderly', Nature, 488, pp. 178–184. doi: 10.1038/nature11319
Type of publication	Article (peer-reviewed)
Link to publisher's version	10.1038/nature11319
Rights	© 2012, Macmillan Publishers Limited. All rights reserved. This is a post-peer-review, pre-copyedit version of an article published in Nature. The final authenticated version is available online at: https://doi.org/10.1038/nature11319
Download date	2024-04-25 13:13:27
Item downloaded from	https://hdl.handle.net/10468/10653

University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Supplementary Figures

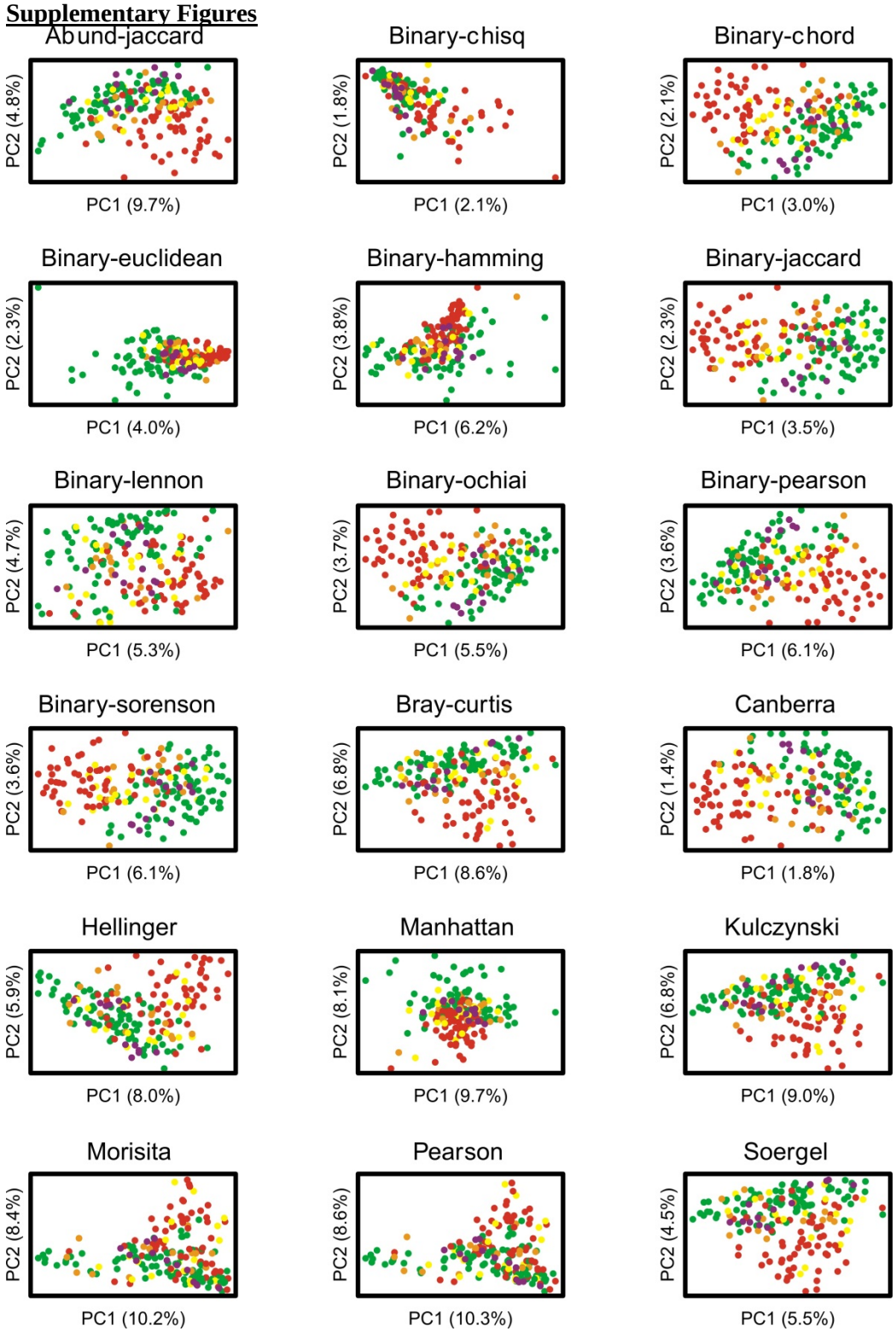


Figure 1 | Principal Co-ordinate Analysis based on 18 different non-UniFrac beta-diversity distances.
Subjects are colour coded according to community setting; Community (green), Day Hospital (yellow), Rehabilitation (orange), Long-stay (red), and Young healthy control subjects (purple).

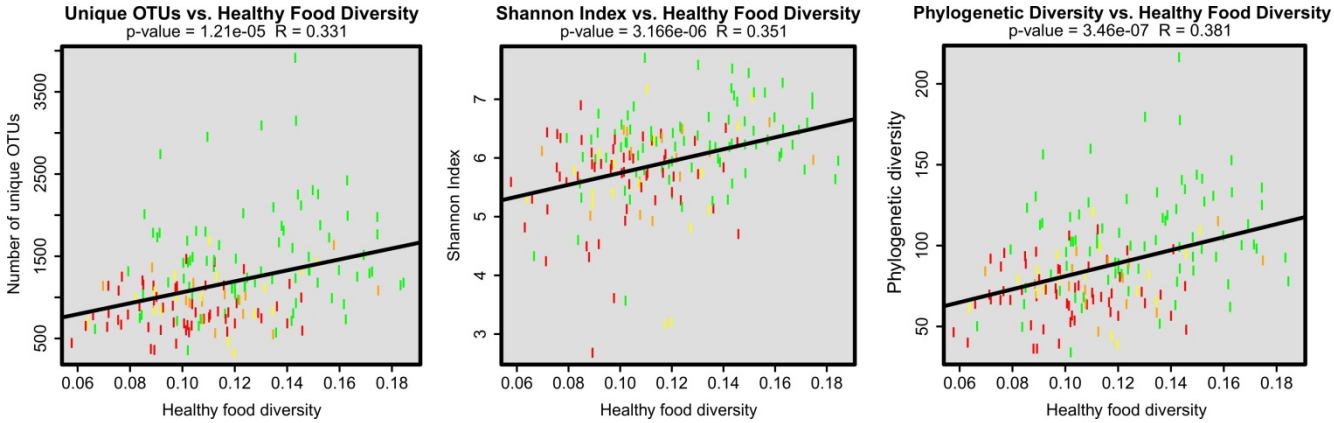


Figure 2 | Pearson correlations and linear regression of Healthy Food Diversity and three microbiota diversity metrics for the 168 subjects where dietary information was available.
Subjects were colour coded as in Figure 1.

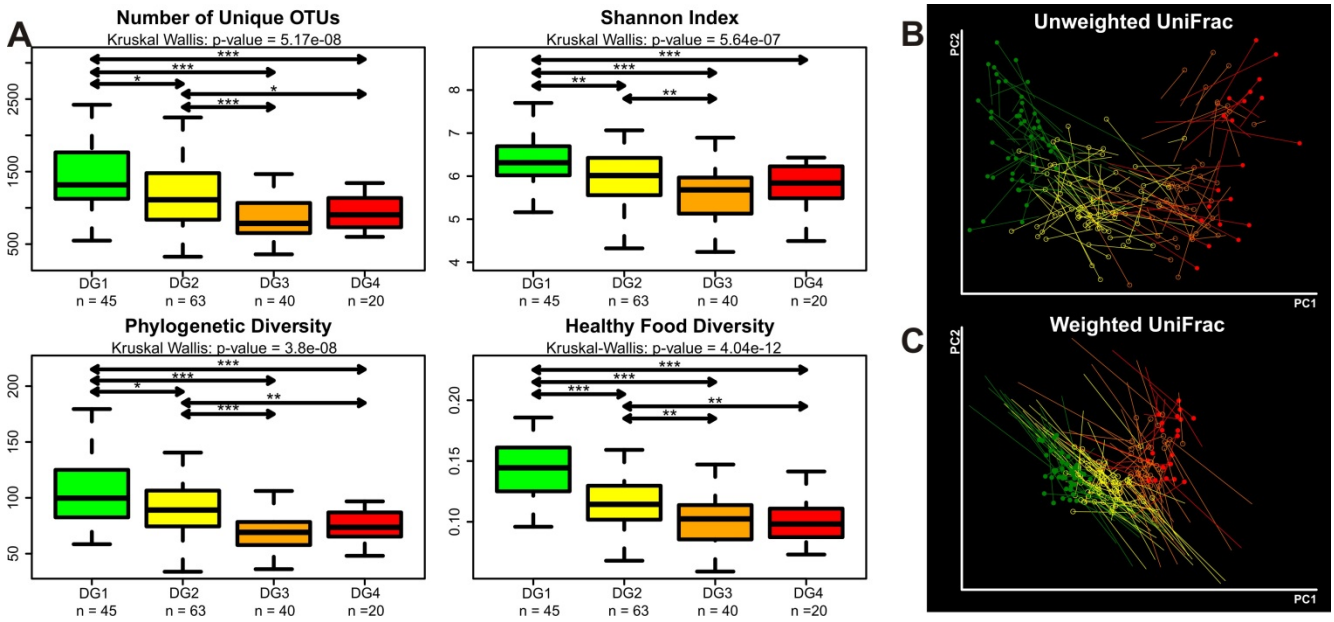


Figure 3 | Microbiota diversity and healthy food diversity correlations
A) Comparison of three microbiota and one healthy food diversity indices across the four dietary groups. B) Procrustes plots of unweighted and C) weighted UniFrac PCoA analysis of microbiota combined with FFQ PCA. The subjects in all panels are colour coded according to diet groups; DG1 (green closed circles), DG2 (yellow open circles), DG3 (orange open circles), and DG4 (red closed circles).

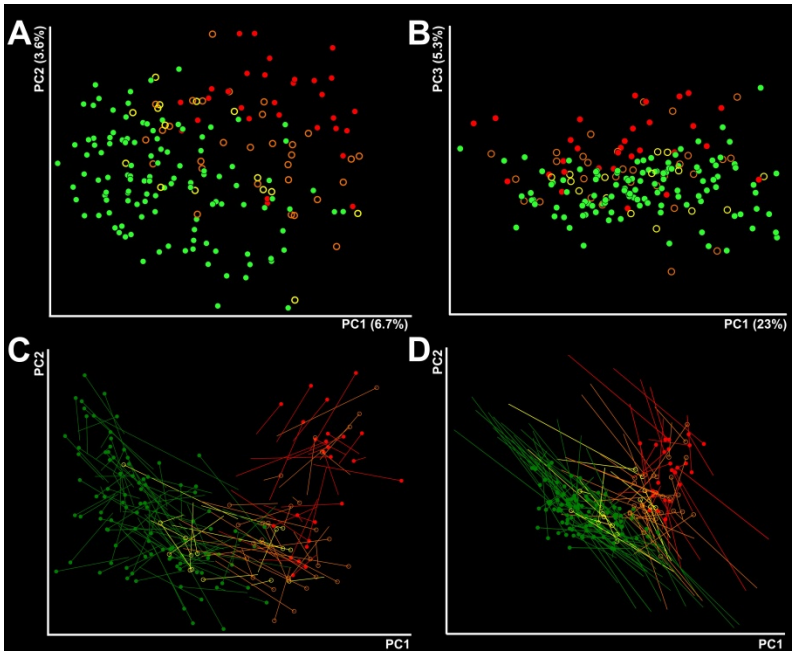
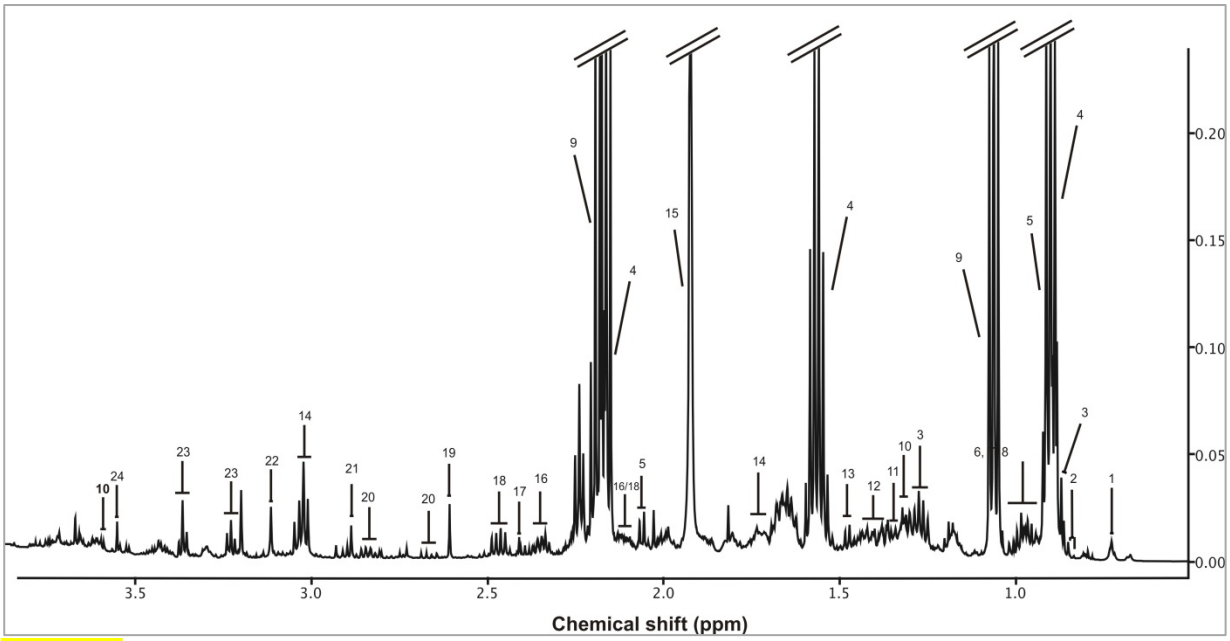
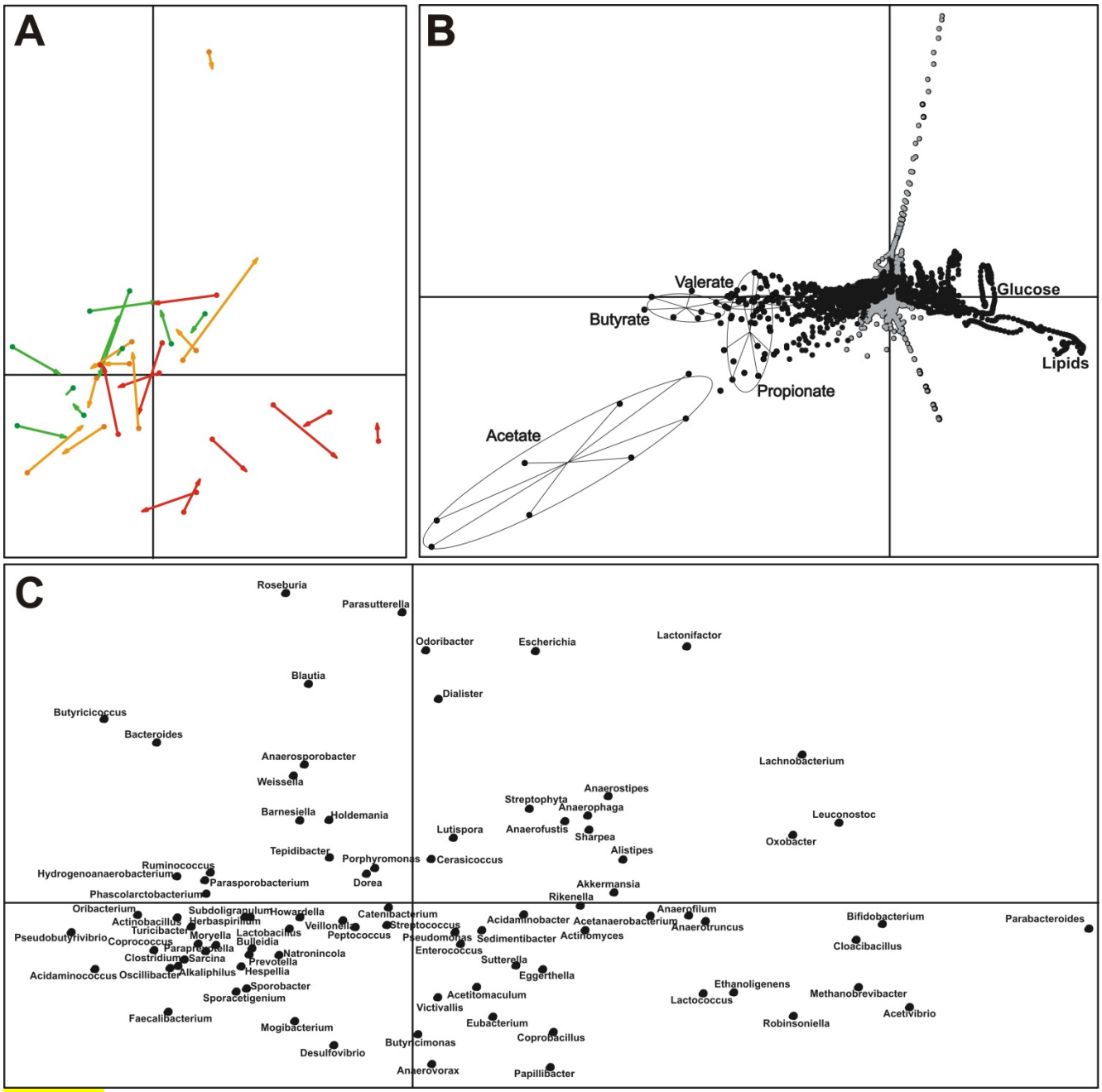


Figure 4 | Duration in long-stay care affects microbiota
A) Unweighted and B) weighted UniFrac PCoA of faecal microbiota from 191 subjects. Panels C and D show Procrustes analysis combining unweighted (C) and weighted (D) UniFrac PCoAs (non-circle end of lines) with Food Type PCA (circle-end of lines). The subjects are colour coded according to duration in long-stay care: N/A (community, day hospital and young healthy controls; green closed circles), less than six weeks (rehabilitation; yellow open circles), from six weeks to one year (long-stay; orange open circles), and longer than one year (long-stay; red closed circles).



NEW FIG

Figure 5 | Representative ^{1}H NMR spectrum of a faecal water extract.
Peaks are labelled to indicate a range of metabolites present in faecal water. The faecal sample was from from a subject in the community residence group. Peak identiities: 1, cholate; 2, caprylate; 3, valerate; 4 butyrate; 5, isovalerate; 6, valine; 7, leucine; 8, isoleucine; 9, propionate; 10, threonine; 11, lactate; 12, isocaproate; 13, alanine; 14, lysine; 15, acetate; 16, glutamate; 17, succinate; 18, glutamine; 19, methylamine; 20, aspartate; 21, trimethylamine; 22, malonate; 23, taurine; 24, glycine.



NEW FIG

Figure 6 | Co-inertia analysis (CIA) of relationships between metabolome, microbiota composition and residence location.

The upper left panel shows the CIA of the metabolomics PCA and microbiota PCA, with arrows indicating where samples position in the metabolite dataset relative to the microbiota dataset. The upper right panel shows NMR loadings data; 95 % confidence intervals were calculated for individual loadings using jack-knife analysis[1]. Loadings that are significantly different from zero are presented in the plot as black dots with those that failed to show significance presented in grey. Relevant metabolites are labelled with dashed lines connecting NMR regions that represent the same metabolite. The lower panel displays the associated microbiota at genus level. Only genera present in at least 20% of the samples were used in the analysis.

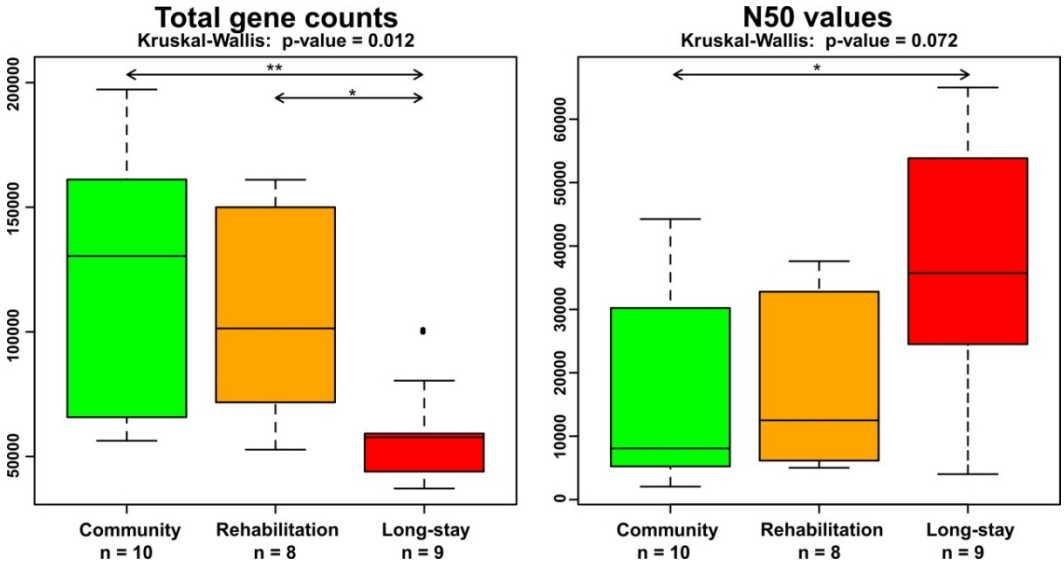


Figure 7 | Predicted gene counts and assembly statistics for the faecal metagenome of 27 selected subjects. The graphs show total gene counts and sequencing assembly N50 values for shotgun sequence data of the faecal metagenome for 27 subjects of indicated residence location.

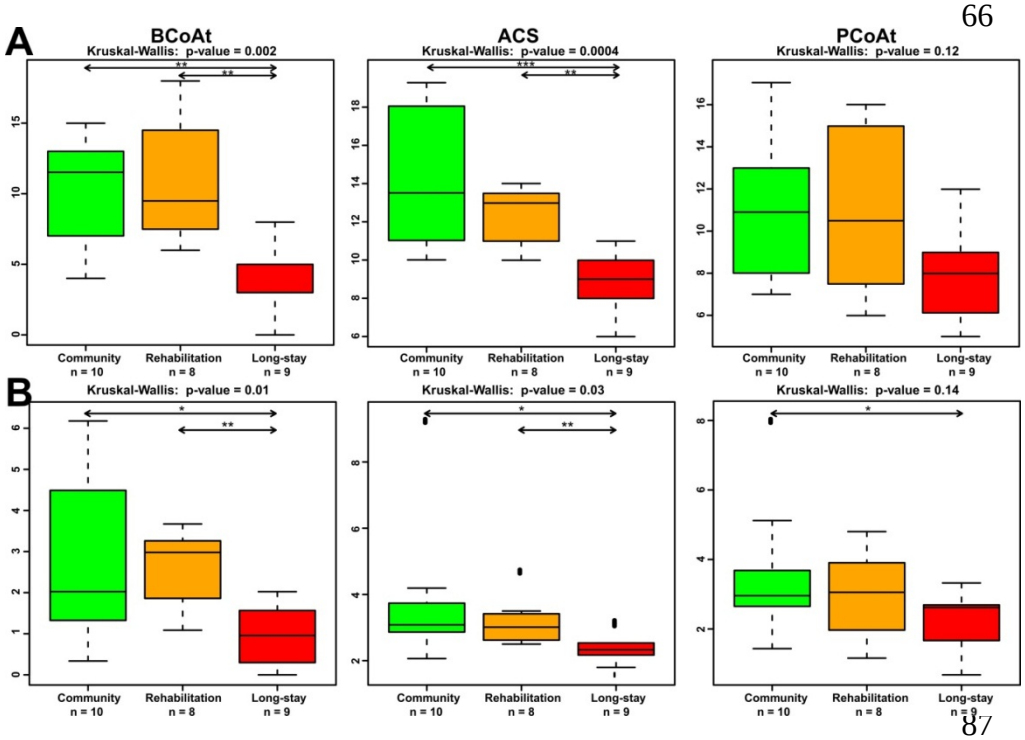


Figure 8 | Frequency of genes relayed to butyrate, acetate and propionate production in the faecal metagenome of 27 subjects.
Comparison of A) gene counts and B) average sequencing coverage, for enzymes involved in butyrate, acetate and propionate production. Gene count values were normalised for 4.79 x 109 of sequenced bases per subject, and coverage values were normalised for the average coverage in each metagenome. BCoAt: Butyryl-CoA transferase/Acetyl-CoA hydrolase; ACS: Acetate-formyltetrahydrofolate synthetase/Formate-tetrahydrofolate ligase; PCoAt: Propionyl-CoA:succinate-CoA transferase/Propionate CoA-transferase.

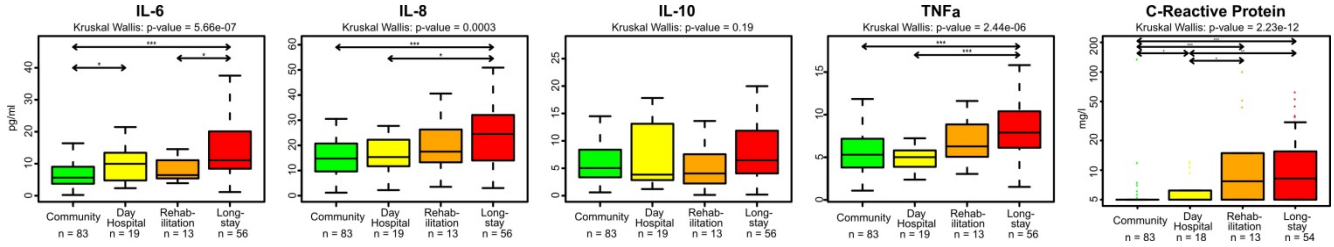
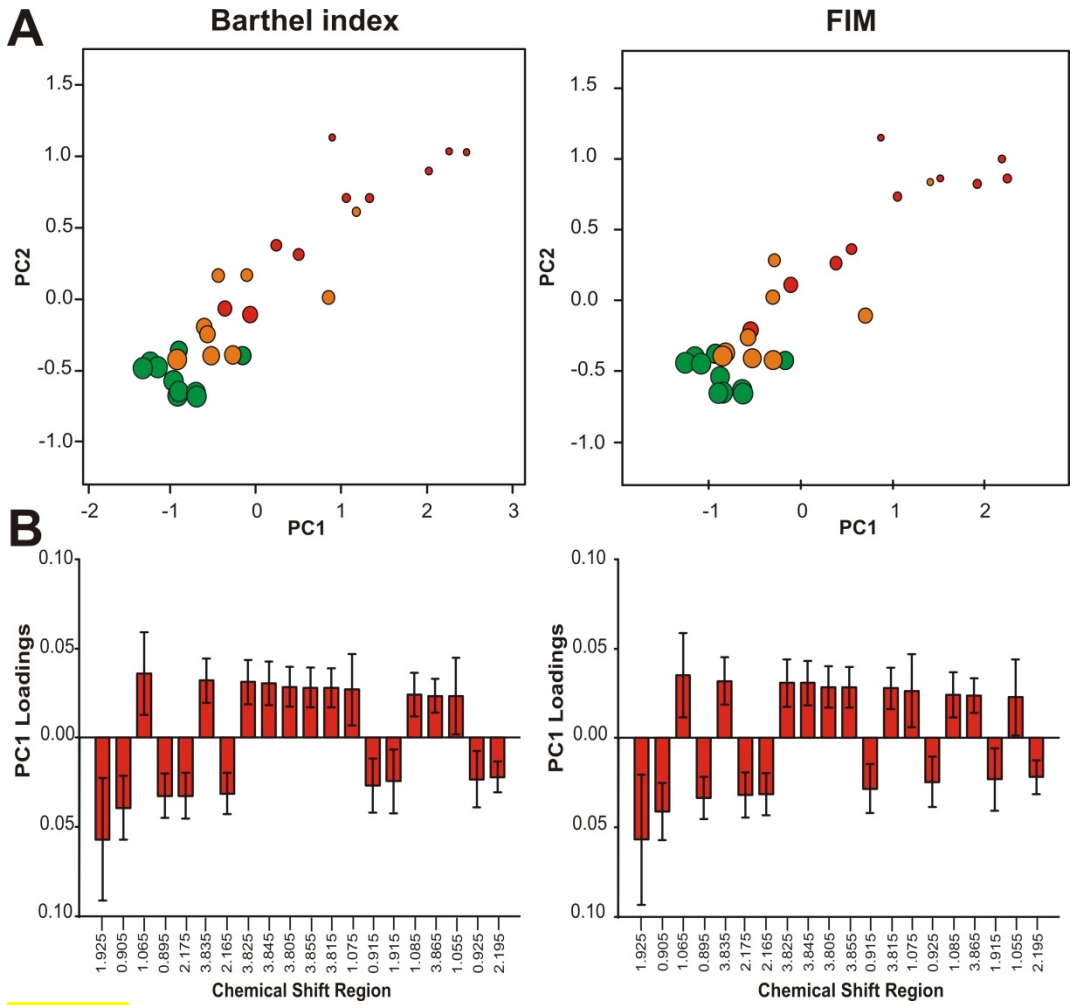


Figure 9 | Cytokine and C Reactive Protein levels vary across subject residence location.
Box plots are color coded by residence location according to the scheme in Figure 1. Kruskal-Wallis P-values refer to tests performed across all four community locations, and Mann-Whitney test was performed for each pair-wise comparison.



NEW FIG

Figure 10 | Correlation of faecal metabolome with indices of frailty.
A. PPCCA score plots derived from ^{1}H NMR spectra of faecal water with FIM or Barthel index as covariate. Dot size reflects covariate value. The influence of the covariate was quantified by examining regression parameter estimates and the associated 95% CI. Barthel index has a significant effect on PC1 with an intercept of 1.62 (2.96, 0.28) (95% CI). FIM has a significant effect on PCI with an intercept of 1.58 (0.19, 2.96). Spectra from community (n=10) are represented by green circles, rehabilitation (n=9) are represented by orange circles and long-stay residents (n=10) are represented by red circles. B. The top loadings for PC1 for each model. Acetate (1.925, 1.915 ppm), butyrate (0.905, 0.895, 0.915, 2.165 ppm) and propionate (2.175, 2.195 ppm) increase with increasing FIM and Barthel values. Error bars represent the 95% confidence intervals. .

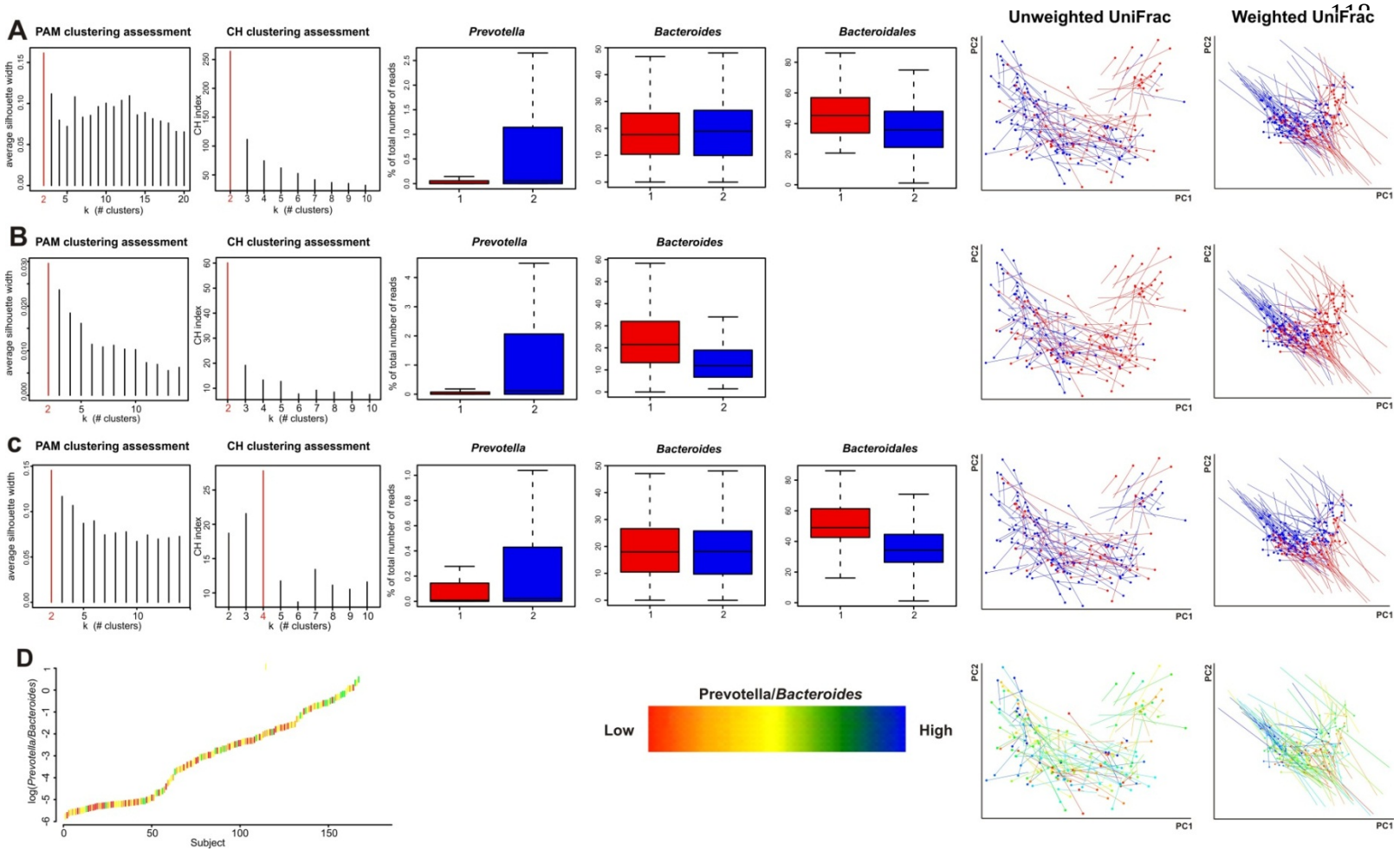


Figure 11 | Enterotype definition is highly dependent on choice of clustering method.
A) Enterotype clustering when based on Jensen-Shannon distances according to approach by Arumagam et al. 2010. B) Enterotype clustering when based on unweighted, and C) weighted UniFrac distances according to the approach by Wu et al. 2011. Abundances of *Prevotella* and *Bacteroides*/*Bacteroidales* in the two clusters were displayed as boxplots and in the unweighted UniFrac Procrustes plots to the right of figure. Clusters generated from the three methods were validated using both the silhouette technique (left bar chart) and the Calinski-Harabasz Index (right bar chart). D) In replacement of clustering approach, the *Prevotella*/*Bacteroides* ratio was ordered by size and subject colour-coded by residence location (left), and displayed as colour gradient in the Procrustes plots.

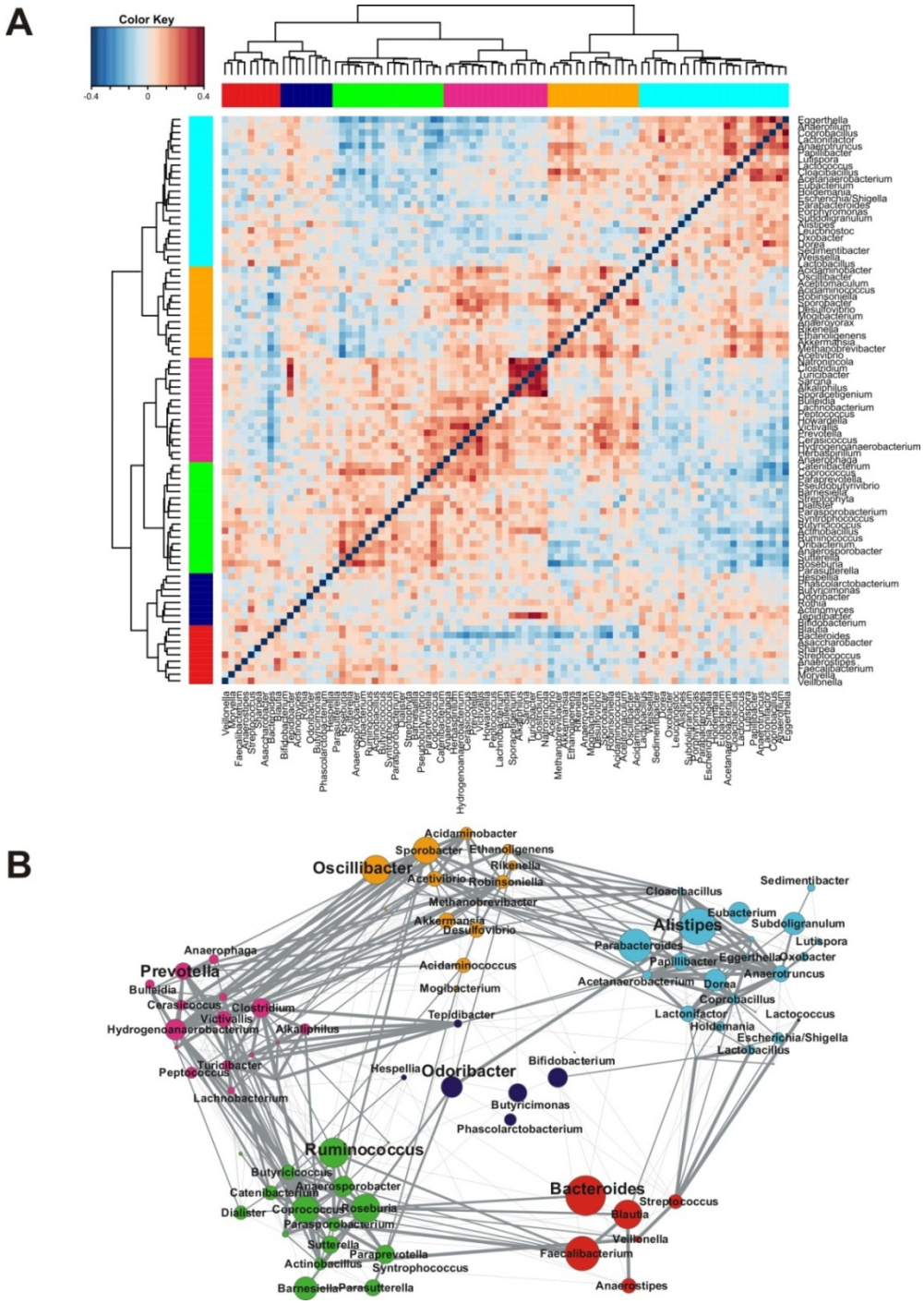


Figure 12 | Definition of bacterial Co-abundance groups (CAGs)

CAGs were defined by A) heat plot showing Kendall correlations between genera clustered by the Spearman correlation coefficient and Ward linkage hierarchical clustering; B) Network plot highlighting correlation relationships between five CAGs and a central group, for the whole cohort studied. Circle sizes indicate genus abundances. The thickness of the lines is proportional to correlation strength.

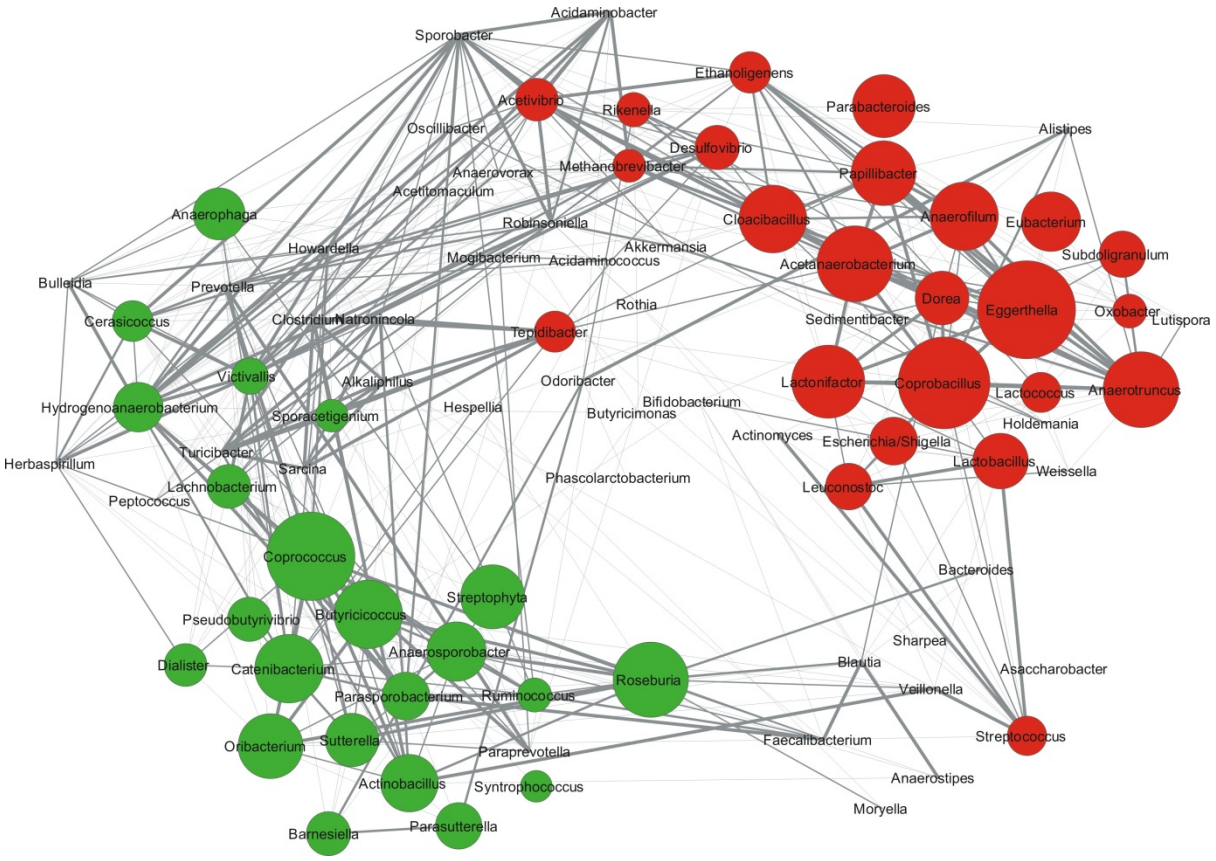


Figure 13 | Genera that are significantly differentially abundant between Community (green) and Long-stay (red) subjects.

Circle sizes indicate significant differential over-abundance in either of the two residence locations.

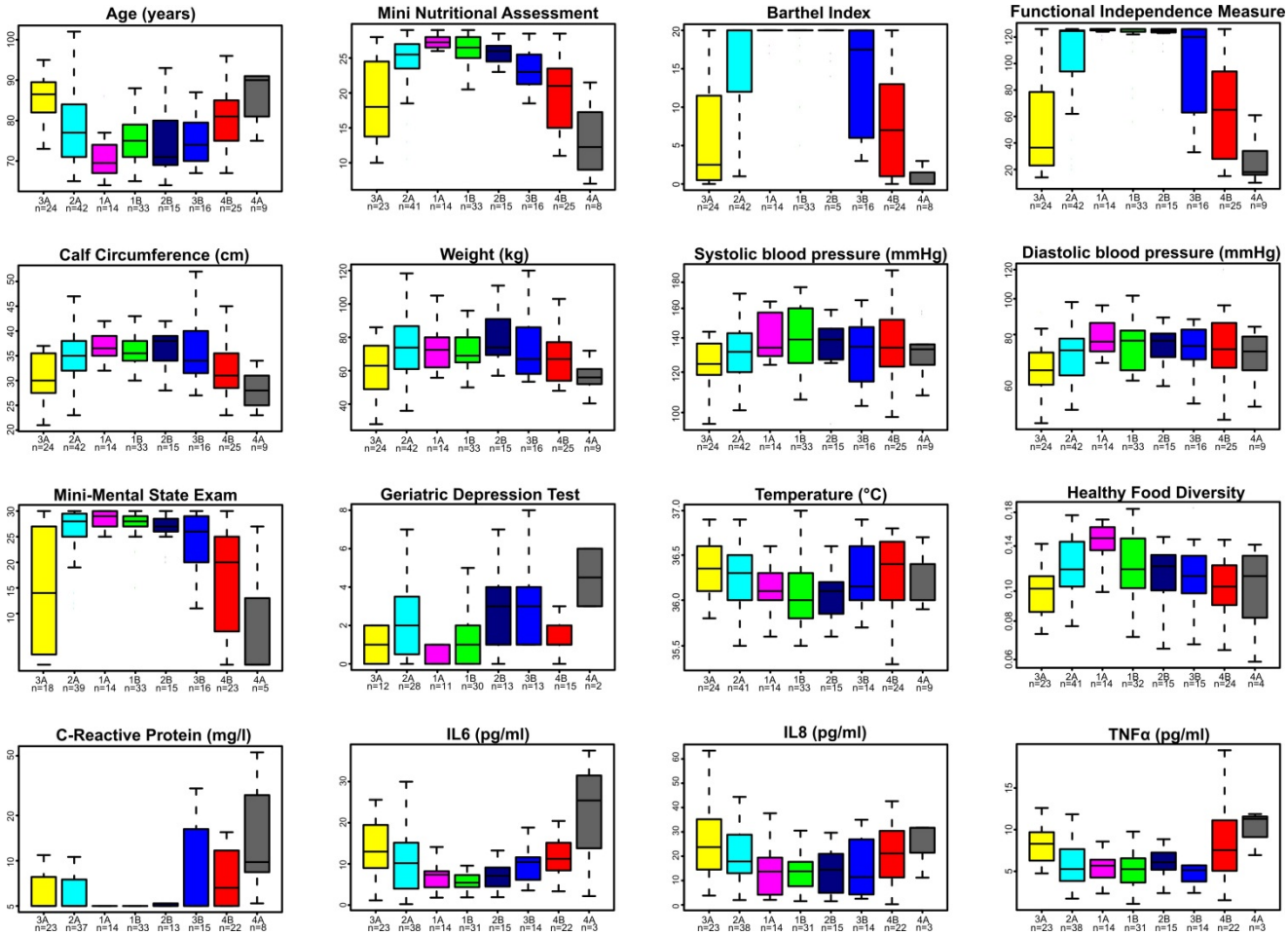


Figure 14 | Relationship between the main eight microbiota groupings and health indicators.

Box-plots showing health measures and indices, and food indices, as a function of main microbiota groupings 1a - 4b from UniFrac analysis in Fig.4.

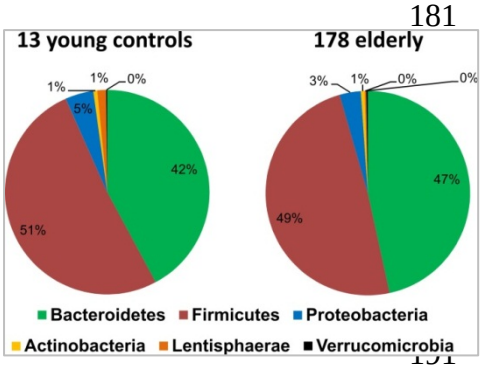


Figure 15 | Aggregate reads classified at phylum level in faecal microbiota of young control and elderly subjects.

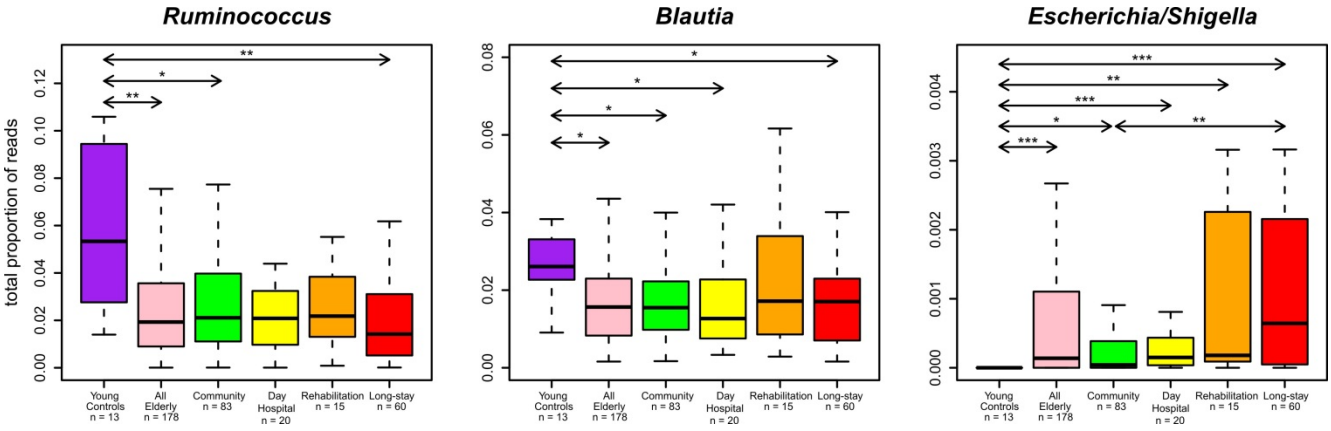


Figure 16 | Major microbiota differences by community location.
The box-plots show significant genus-level differences between microbiota of young control subjects, the complete older person group studied here (all elderly), and cohorts defined by residence location.

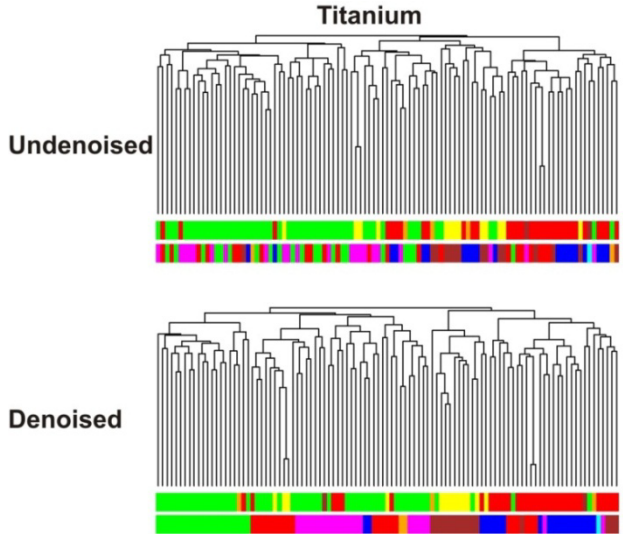


Figure 17 | Hierarchical clustering of denoised and un-denoised OTUs, for both sequencing platforms, based on the Pearson correlation similarity metric and average linkage.
The top horizontal line for each heat map shows community location and the bottom line shows the different pyrosequencing runs with one colour each.

Supplementary methods

Subject recruitment and sample collection

This study was approved by the Cork Clinical Research Ethics Committee. Subjects older than 64 years were recruited and clinically investigated in two local hospitals, which serve a population base of ~481,000 in the Cork city and county region. They were defined as (a) community-dwelling (Community); (b) attending an out-patient day hospital (Out-patient); (c) in short-term rehabilitation hospital care (Rehabilitation; under six-weeks stay) or (d) in long term institutionalized care (long-stay; more than six weeks). The mean age of the subjects was 78 (±8) years, with a range of 64 to 102 years. The subjects were all of Irish ethnicity. None of the faecal samples from elderly subjects from our previous study[2] were analyzed in the current analysis, because we did not have food frequency data for all that cohort. Exclusion criteria were a history of alcohol abuse, participation in an investigational drug evaluation or antibiotic treatment within prior 30 days, or advanced organic disease. Informed consent was obtained from all subjects or, in cases of cognitive impairment, by next-of-kin in accordance with the local research ethics committee guidelines. Data collected included anthropometric measurements, clinical history and status and medication history. Antibiotic use prior to the one-month exclusion period was also recorded for each subject. Thirteen younger adult subjects of age ranging 28-46 years, which had not been treated with antibiotics within 30 days, were also recruited by informed consent.

Clinical and nutritional data collection

Habitual dietary intake was assessed using a validated, semi-quantitative, food frequency questionnaire (FFQ) based upon the SLAN study[3]. Food properties were determined using the UK Food Standards Agency Nutrient databank[4]. The Mini Nutritional Assessment (MNA) was used as a screening and assessment tool to identify subjects at risk of malnutrition.

Non-fasted blood samples were collected and analysed at Cork University Hospital clinical laboratories. Cytokines were measured using validated, commercial multi-spot microplates (Meso Scale Diagnostics). Anthropometric measures included height, weight, calf and mid-arm circumference. Charlson Comorbidity Index, Mini mental State Exam, Geriatric Depression Test,

Barthel Score and Functional Independence Measures were carried out on all participants. For long-term care, day-hospital and rehabilitation subjects, a research nurse reviewed the medical records for information on disease and current medication usage.

Molecular methods and bioinformatics

DNA was extracted from fecal samples, and the V4 region of the 16S rRNA gene was amplified, sequenced and analyzed, as described previously[5]. Briefly, V4 amplicons were sequenced on a 454 Genome Sequencer FLX Titanium platform (Roche Diagnostics Ltd, West Sussex, UK and Beckman Coulter Genomics SA, Grenoble, France). Raw sequencing reads were quality trimmed using the QIIME pipeline[6] according to the following criteria: i) exact matches to primer sequences and barcode tags; ii) no ambiguous bases (Ns); iii) read-lengths not shorter than 150 bp or longer than 350 bp; iv) the average quality score in a sliding window of 50 bp not to fall below 25. For large-scale assignments into the new Bergey bacterial taxonomy[7] we used the RDP-classifier version 2.2 with 50% as confidence value threshold. This was based on what was found suitable for V4 amplicons from the human gut environment[5]. RDP classifications were imported into a MySQL database for efficient storage and advanced querying.

The amplicon reads were truncated at 210 bp prior to OTU picking at 97% similarity level, and filtered for chimeric sequences using ChimeraSlayer. Representative sequences (the most abundant) for each OTU were aligned using PyNAST[8] prior to tree building using FastTree[9]. These phylogenies were combined with absence/presence or abundance information for each OTU to calculate unweighted or weighted UniFrac distances, respectively[10]. Principal Co-ordinate Analysis and Procrustes superimposition were then performed from the UniFrac distances and Food Frequency data. The amplicon sequences were deposited in MG-RAST under the Project ID 154.

Metagenomes were sequenced from libraries with 91 bp paired-end 91 Illumina reads and 350 bp insert size and assembled using MetaVelvet[11]. Samples EM039 and EM173 were sequenced from libraries of 101 bp paired-end Illumina reads with a 500 bp insert size, and subsequently assembled using MIRA[12] in hybrid with 551,726 and 665,164 454 Titanium reads, respectively. Protein sequences from enzymes were screened against the assembled metagenomes using TBLASTN with an

amino acid identity cut-off of 30% and an alignment length cut-off of 200 bp. We screened the metagenome data for enzymes associated with production of butyrate (butyryl-CoA transferase/acetyl-CoA hydrolase), acetate (acetate-formyltetrahydrofolate synthetase/formate-tetrahydrofolate ligase), and propionate (propionyl-CoA:succinate-CoA transferase/propionate CoA-transferase). Genes were predicted using MetaGene[13].

NMR analysis of the faecal water metabolome

Faecal water samples were prepared by the addition of 60 µl D_2O and 10 µl tri-methylsilyl-2,2,3,3-tetradeuteriopropionate to 540 µl faecal water. Spectra of samples were acquired by using a Carr-Purcell-Meiboom-Gill (CPMG) pulse sequence with 32 K data points and 256 scans. Spectra were referenced to TSP at 0.0 ppm, phase and baseline corrected with a line broadening of 0.3 Hz using the processor on Chenomx NMR suite 7 (Chenomx Inc, Edmonton, Alberta, T5K 2J1). The spectra were integrated at full resolution for data analysis (PCA, PLS-DA, CIA) with the water region (4 – 6 ppm) excluded and the data was normalized to the sum of the spectral integral. For PPCCA data analysis, the spectra were integrated into spectral regions (0.01ppm).

Statistical methods and metabolome data analysis

Statistical analysis was carried out using R (version 2.13.2) or Stata (version 11) software packages. Kruskal-Wallis and Mann-Whitney tests were used to find significant differences in microbial taxa, clinical and biochemical measures, alpha diversity, and Healthy Food Diversity (HFD). We used least square linear regression for comparing alpha diversity and HFD. Median regression[14] was used to compare clinical measures and microbiota, while adjusting for age, gender, medications, and when appropriate residence location. For median regression, the median was modelled as a linear function of independent variables. Model parameters are estimated such that they minimised the sum of the absolute differences between observed and predicted values. P-values were generated using the wild bootstrap method[15] to estimate variance.

A linear quantile (median) regression for two variables – a response variable (y) and a predictor variable (x) – is the following: median (y) = $\beta 0 + \beta 1x$ where $\beta 0$ is the intercept (value when

y = 0) and β1 is the slope (change in median of y for a unit change in x). Together, these parameters describe the association between y and x, where x is a predictor of y. In the case of multiple predictor variables, each one is added to the regression equation and so the equation becomes median (y) = β0 + β1x1 + β2x2 and now the slope β1 is interpreted as the median change in x1 after adjusting for x2. This can be likened to a laboratory experiment where the specific effect of one variable on another is isolated by holding all other relevant variables constant.

Following statistical analysis of the taxonomic classifications, we estimated q-values to control for multiple testing using the qvalue and nFDR functions in the R package[16]. At the genus level we estimated the proportion of true null hypotheses with the qvalue function unless the estimated pi0 was less or equal to zero. In all other instances, we used the Benjamini & Hochberg method[17].

Statistical analysis of the NMR data was performed using diverse software packages: PCA and PLS-Da analysis was performed in SIMCA-P+ (Umetrics, Umeå, Sweden); Permutation testing was performed in R and PPCCA was performed in R using the MetabolAnalyze package. The NMR data was Pareto-scaled prior to data analysis. Assignment of the spectral peaks was performed using in-house libraries, statistical correlation analysis and 2D NMR spectra (TOCSY and COSY).

Supplementary Tables

Table 1. Taxa that are found in 50% of the subjects, significantly different across the four community locations, according to a Kruskal-Wallis test, and that are more abundant in Community subjects relative to Long-stay subjects.

C: Community subjects; DH: Day Hospital; R: Rehabilitation; LS: Long-stay. Community location columns show medians of percentages of RDP-classified reads over the total number of reads for each subject. The FDR value represents a Benjamini-Hochberg corrected P-value for the smaller data sets Clostridium cluster and Family, and a Q-value for the larger data sets Genus and Species.

	C	DH	R	LS	FDR
Family					
Lachnospiraceae (Firmicutes/Clostridia/Clostridiales)	10.50	8.94	11.19	5.95	7.1E-05
Pasteurellaceae (Proteobacteria/Gammaproteobacteria/Pasteurellales)	0.03	0.05	0.01	0.00	0.0025
Alcaligenaceae (Proteobacteria/Betaproteobacteria/Burkholderiales)	0.28	0.53	0.33	0.09	0.0034
Genus					
Coprococcus (Firmicutes/Clostridia/Clostridiales/Lachnospiraceae)	1.52	0.67	1.23	0.40	8.40E-08
Roseburia (Firmicutes/Clostridia/Clostridiales/Lachnospiraceae)	1.79	2.51	2.26	0.28	8.40E-07
Butyricicoccus (Firmicutes/Clostridia/Clostridiales/Ruminococcaceae)	0.11	0.08	0.07	0.05	0.00002
Catenibacterium (Firmicutes/Erysipelotrichi/Erysipelotrichales/Erysipelotrichaceae)	0.06	0.02	0.00	0.00	0.00002
Oribacterium (Firmicutes/Clostridia/Clostridiales/Lachnospiraceae)	0.01	0.01	0.01	0.00	0.00006
Anaerosporobacter (Firmicutes/Clostridia/Clostridiales/Lachnospiraceae)	0.44	0.19	0.35	0.12	0.00018
Actinobacillus (Proteobacteria/Gammaproteobacteria/Pasteurellales/Pasteurellaceae)	0.03	0.05	0.01	0.00	0.00018
Lachnobacterium (Firmicutes/Clostridia/Clostridiales/Lachnospiraceae)	0.01	0.00	0.01	0.00	0.00027
Hydrogenoanaerobacterium (Firmicutes/Clostridia/Clostridiales/Ruminococcaceae)	0.47	0.07	0.43	0.12	0.00057
Sutterella (Proteobacteria/Betaproteobacteria/Burkholderiales/Alcaligenaceae)	0.09	0.11	0.06	0.00	0.00085
Parasutterella (Proteobacteria/Betaproteobacteria/Burkholderiales/Alcaligenaceae)	0.08	0.06	0.05	0.01	0.0011
Parasporobacterium (Firmicutes/Clostridia/Clostridiales/Lachnospiraceae)	0.08	0.04	0.06	0.02	0.0019
Barnesiella (Bacteroidetes/Bacteroidia/Bacteroidales/Porphyromonadaceae)	0.48	0.44	0.67	0.11	0.0048

Table 2. Taxa identified as per criteria in SupplementaryTable 1 but showing taxa that are more abundant in Long-stay subjects relative to Community-dwelling subjects.

	C	DH	R	LS	FDR
Family					
Desulfovibrionaceae (Proteobacteria/Deltaproteobacteria/Desulfovibrionales)	0.01	0.01	0.02	0.09	5.8E-05
Lactobacillaceae (Firmicutes/Bacilli/Lactobacillales)	0.00	0.01	0.15	0.02	7.1E-05
Eubacteriaceae (Firmicutes/Clostridia/Clostridiales)	0.18	0.87	0.19	0.89	7.1E-05
Porphyromonadaceae (Bacteroidetes/Bacteroidia/Bacteroidales)	4.41	7.20	5.57	8.53	0.00047
Enterobacteriaceae (Proteobacteria/Gammaproteobacteria/Enterobacteriales)	0.01	0.02	0.03	0.12	0.00059
Coriobacteriaceae (Actinobacteria/Actinobacteria/Coriobacteriales)	0.01	0.01	0.03	0.02	0.00087
Incertae Sedis XI (Firmicutes/Clostridia/Clostridiales)	0.00	0.00	0.00	0.01	0.0054
Streptococcaceae (Firmicutes/Bacilli/Lactobacillales)	0.05	0.09	0.11	0.10	0.042
Genus					
Coprobacillus (Firmicutes/Erysipelotrichi/Erysipelotrichales/Erysipelotrichaceae)	0.01	0.02	0.05	0.15	5.7E-09
Lactonifactor (Firmicutes/Clostridia/Clostridiales/Ruminococcaceae)	0.06	0.10	0.18	0.21	3.7E-07
Anaerotruncus (Firmicutes/Clostridia/Clostridiales/Ruminococcaceae)	0.09	0.08	0.12	0.26	4.3E-07
Acetanaerobacterium (Firmicutes/Clostridia/Clostridiales/Ruminococcaceae)	0.01	0.02	0.02	0.04	5.6E-07
Lactobacillus (Firmicutes/Bacilli/Lactobacillales/Lactobacillaceae)	0.00	0.01	0.15	0.02	3.9E-06
Anaerofilum (Firmicutes/Clostridia/Clostridiales/Ruminococcaceae)	0.00	0.00	0.00	0.01	7.8E-06
Eubacterium (Firmicutes/Clostridia/Clostridiales/Eubacteriaceae)	0.16	0.87	0.19	0.89	1.0E-05
Papillibacter (Firmicutes/Clostridia/Clostridiales/Ruminococcaceae)	0.11	0.11	0.10	0.38	0.00002
Parabacteroides (Bacteroidetes/Bacteroidia/Bacteroidales/Porphyromonadaceae)	2.56	4.18	4.20	5.86	0.00009
Acetivibrio (Firmicutes/Clostridia/Clostridiales/Ruminococcaceae)	0.04	0.01	0.01	0.11	0.00018
Ethanoligenens (Firmicutes/Clostridia/Clostridiales/Ruminococcaceae)	0.03	0.02	0.02	0.06	0.00038
Dorea (Firmicutes/Clostridia/Clostridiales/Lachnospiraceae)	0.30	0.44	0.47	0.66	0.00055
Escherichia/Shigella (Proteobacteria/Gammaproteobacteria/Enterobacteriales/Enterobacteriaceae)	0.00	0.02	0.02	0.06	0.0012
Subdoligranulum (Firmicutes/Clostridia/Clostridiales/Ruminococcaceae)	0.34	0.34	0.42	0.70	0.0026
Desulfovibrio (Proteobacteria/Deltaproteobacteria/Desulfovibrionales/Desulfovibrionaceae)	0.00	0.00	0.00	0.04	0.0032
Oxobacter (Firmicutes/Clostridia/Clostridiales/Clostridiaceae)	0.02	0.02	0.01	0.04	0.0052
Streptococcus (Firmicutes/Bacilli/Lactobacillales/Streptococcaceae)	0.04	0.08	0.11	0.09	0.0054

322 **Table 3. Median values of food consumption for each of the four diet groups (DGs).**
323

Food Names	Food Groups	DG1	DG2	DG3	DG4
White bread and rolls (including ciabatta & pannini)	cereals and potatoes	0.01	0.143	0.4285	0
Brown bread and rolls	cereals and potatoes	0.143	0.286	0.0215	0
Wholemeal bread and rolls	cereals and potatoes	0.714	0.033	0	0
Wheat-free; Rye bread; spelt bread (specify)	cereals and potatoes	0	0	0	0
Cream crackers, cheese biscuits	cereals and potatoes	0	0.01	0	0
Crisp bread	cereals and potatoes	0	0	0	0
Pancakes, muffins, oatcakes	cereals and potatoes	0	0	0	0
Scone (white)	cereals and potatoes	0.01	0.01	0.0215	0.005
Scone (brown)	cereals and potatoes	0	0	0	0
Non-ready to eat - Porridge	cereals and potatoes	1	0.286	1	1
High fibre (Bran/flakes, wheat biscuits, shredded wheat)	cereals and potatoes	0	0	0	0
Corn flakes, popped rice	cereals and potatoes	0	0	0	0
Muesli	cereals and potatoes	0	0	0	0
Sugar-coated cereals	cereals and potatoes	0	0	0	0
Boiled, instant or jacket potatoes	cereals and potatoes	1	0.714	0	0
Mashed potatoes	cereals and potatoes	0.033	0.286	1	1
Chips / Roast potatoes	cereals and potatoes	0.033	0.033	0.033	0.005
Potato Salad	cereals and potatoes	0.01	0.01	0	0
White Rice	cereals and potatoes	0.01	0	0	0
Brown Rice	cereals and potatoes	0.01	0	0	0
White/yellow or green pastas (e.g. spaghetti, noodles)	cereals and potatoes	0.01	0	0	0
Wholemeal pasta	cereals and potatoes	0	0	0	0
Cream (tablespoon)	dairy	0	0.01	0	0.2145
Creme fraiche/soured cream	dairy	0	0	0	0
Low-fat yoghurt	dairy	0	0	0	0
Full-fat yoghurt or Greek style yoghurt (125g carton)	dairy	0	0.01	0	0.286
Dairy desserts (125g carton)	dairy	0	0	0	0.005
Cheddar cheese (medium serving)	dairy	0.143	0.143	0	0
Low-fat cheddar cheese (medium serving)	dairy	0	0	0	0
Cottage cheese	dairy	0	0	0	0
Eggs as boiled, fried, scrambled, poached (one)	dairy	0.286	0.286	0.143	0
Butter	dairy	0	0	0	0.0715
Lite Butter	dairy	0	0	0	0
Low-fat margarine	dairy	0	0	0	0
Cholesterol Lowering margarine	dairy	0	0	0	0
Cream & Veg. Oil spread	dairy	0	0	0	0
Olive oil spread	dairy	0	0	0	0
Meat or cream soups: homemade / fresh (1 bowl)	dairy	0	0.01	0	0
Meat or cream soups: tinned / packet (1 bowl)	dairy	0	0	0	0
Whole milk (cup) - cow, goat, soya, rice milk -	dairy	0	0	0.143	1

specify					
Semi-skimmed milk (cup)	dairy	0	0	0	0
Probiotic Yoghurts / cheese / milk	dairy	0.286	0	0	0
Fish fried (batter/breadcrumbs), baked, grilled, fingers/cakes	fish	0	0.01	0.088	0.143
White fish, fresh or frozen (e.g. cod, haddock, plaice, sole)	fish	0.143	0.143	0	0
Oily fish, fresh or canned (e.g. mackerel, kippers, tuna, salmon, sardines, herring)	fish	0.143	0.033	0	0
Shellfish (e.g. crab, prawns, mussels)	fish	0	0	0	0
Apples	fruit	0.286	0.143	0	0.2145
Pears	fruit	0.143	0.033	0	0
Oranges, satsumas, mandarins	fruit	0.286	0.033	0	0
Grapefruit	fruit	0	0	0	0
Bananas	fruit	0.286	0.286	0.143	0.2145
Grapes	fruit	0.143	0.01	0	0
Melon	fruit	0.01	0	0	0
Peaches, plums	fruit	0.01	0	0	0
Apricots	fruit	0	0	0	0
Strawberries, raspberries, kiwi fruit	fruit	0.143	0.01	0	0
Blueberries	fruit	0	0	0	0
Tinned fruit (specify)	fruit	0	0.01	0	0
Dried fruit e.g. raisins	fruit	0.01	0	0	0
Frozen fruit (specify)	fruit	0	0	0	0
Other fruit	fruit	0	0	0	0
Pure fruit drinks e.g. orange juice (small glass)	fruit	0.033	0.033	0.286	1
Fruit squash (small glass)	fruit	0	0	0	0
Beef (roast / steak)	meat	0.033	0.143	0.143	0.286
Beef: stew	meat	0.033	0.143	0.143	0.143
Beef burger (1 burger)	meat	0	0	0	0
Pork (roast / chops / escalopes)	meat	0.01	0.033	0.143	0.143
Lamb (roast / chops / stew)	meat	0.033	0.143	0.143	0.143
Chicken or other poultry e.g. turkey: roast	meat	0.143	0.143	0.286	0.286
Breaded chicken, chicken nuggets, chicken burger	meat	0	0	0	0
Bacon / Ham	meat	0.033	0.143	0.143	0.143
Processed meat (Corned beef, Luncheon meats, sausages)	meat	0	0.033	0.143	0.143
Savoury pies (e.g. meat, pork, steak & kidney, sausage roll)	meat	0	0	0	0
Liver, heart, kidney, paté	meat	0	0	0	0
Chocolate coated sweet biscuits e.g. digestive (one)	sweets, cakes and alcohol	0.143	0.033	0	0
Plain biscuit e.g. digestives, rich tea (one)	sweets, cakes and alcohol	0.033	0.286	0.286	0.2145
Cakes e.g. fruit, sponge	sweets, cakes and alcohol	0.033	0.033	0.033	0
Buns, pastries e.g. croissants, doughnuts	sweets, cakes and alcohol	0	0	0	0
Fruit pies, tarts, crumbles	sweets, cakes and alcohol	0.01	0.033	0.005	0
Sponge puddings	sweets, cakes and alcohol	0	0	0	0.005
Milk puddings e.g. rice, custard, trifle	sweets, cakes and alcohol	0.01	0.033	0.286	0.286

Ice cream, choc ices, frozen desserts	sweets, cakes and alcohol	0.01	0.143	0.143	0.286
Chocolates, singles or squares	sweets, cakes and alcohol	0.033	0.033	0.005	0
Sweets, toffees, mints	sweets, cakes and alcohol	0.01	0.033	0	0
Sugar added to tea coffee, cereals (teaspoons)	sweets, cakes and alcohol	0	0	0	3
Sugar substitute e.g. canderel (teaspoon)	sweets, cakes and alcohol	0	0	0	0
Crisps or other packet snacks	sweets, cakes and alcohol	0	0	0	0
Jam, marmalade, honey, syrup (teaspoon)	sweets, cakes and alcohol	1	1	0.5	0
Wine (glass)	sweets, cakes and alcohol	0.143	0	0	0
Beer, Larger or Cider (half pint)	sweets, cakes and alcohol	0	0	0	0
Low alcohol / alcohol free beer / larger (half pint)	sweets, cakes and alcohol	0	0	0	0
Port, Sherry, Vermouth, liqueurs (glass)	sweets, cakes and alcohol	0	0	0	0
Spirits e.g. Gin, Whiskey (Single measure)	sweets, cakes and alcohol	0	0	0	0
Low calorie or diet soft fizzy drink (glass)	sweets, cakes and alcohol	0	0	0	0.005
Fizzy Soft drinks e.g.soda pop	sweets, cakes and alcohol	0	0	0	0.2145
Carrots	vegetables	0.286	0.286	0.286	0.714
Spinach	vegetables	0.01	0	0	0
Broccoli, spring greens, kale	vegetables	0.286	0.143	0	0
Brussels sprouts	vegetables	0.01	0.01	0.033	0.143
Cabbage	vegetables	0.033	0.143	0.143	0.143
Peas	vegetables	0.033	0.143	0.143	0.286
Green beans, broad beans, runner beans	vegetables	0.01	0	0	0.143
Courgettes	vegetables	0.01	0	0	0
Cauliflower	vegetables	0.033	0.143	0.143	0.143
Parsnips, turnips	vegetables	0.143	0.143	0.143	0.143
Leeks	vegetables	0.033	0	0	0
Onions	vegetables	0.286	0.143	0	0
Garlic	vegetables	0.143	0	0	0
Mushrooms	vegetables	0.143	0.033	0	0
Sweet peppers	vegetables	0.143	0	0	0
Green salad, lettuce	vegetables	0.286	0.143	0	0
Cucumber, celery	vegetables	0.143	0.033	0	0
Tomatoes	vegetables	0.286	0.286	0.088	0
Sweet-corn	vegetables	0	0	0	0
Beetroot	vegetables	0.033	0.01	0	0
Coleslaw	vegetables	0.01	0.01	0	0
Baked beans	vegetables	0.033	0.143	0	0
Dried lentils, beans, peas	vegetables	0	0	0	0
Tofu, soya meat, TVP, vege-burger	vegetables	0	0	0	0
Lasagne (meat based)	vitamins / minerals / tea etc.	0.01	0	0	0
Lasagne (vegetarian)	vitamins / minerals / tea etc.	0	0	0	0
Pizza	vitamins / minerals / tea etc.	0.01	0	0	0
Quiche (medium serving)	vitamins / minerals / tea etc.	0	0	0	0
Light salad cream or light mayonnaise (tablespoon)	vitamins / minerals / tea etc.	0.01	0	0	0
Salad cream, mayonnaise (tablespoon)	vitamins / minerals / tea etc.	0	0	0	0

French dressing (tablespoon)	vitamins / minerals / tea etc.	0.01	0	0	0
Other salad dressing (tablespoon)	vitamins / minerals / tea etc.	0	0	0	0
Sunflower margarine e.g. Flora	vitamins / minerals / tea etc.	0	0	0	0
Peanuts or other nuts	vitamins / minerals / tea etc.	0.01	0	0	0
Vegetable soups: homemade/fresh (1 bowl)	vitamins / minerals / tea etc.	0.033	0.143	0.286	0.286
Vegetable soups: tinned/packet (1 bowl)	vitamins / minerals / tea etc.	0	0	0	0
Sauces e.g. white, cheese, gravy (tablespoon)	vitamins / minerals / tea etc.	0.033	0.286	1	1
Tomato based sauces e.g. pasta sauces	vitamins / minerals / tea etc.	0.01	0	0	0
Curry-type sauces	vitamins / minerals / tea etc.	0.01	0	0	0
Pickles, chutney (tablespoon)	vitamins / minerals / tea etc.	0.01	0	0	0
Marmite, Bovril (tablespoon)	vitamins / minerals / tea etc.	0	0	0	0
Peanut butter (teaspoon)	vitamins / minerals / tea etc.	0	0	0	0
Tea (cup)	vitamins / minerals / tea etc.	2	4	4	4
Herbal tea (cup)	vitamins / minerals / tea etc.	0	0	0	0
Coffee instant (cup)	vitamins / minerals / tea etc.	0.143	0.033	0	0
Coffee ground (cup)	vitamins / minerals / tea etc.	0	0	0	0
Coffee, decaffeinated (cup)	vitamins / minerals / tea etc.	0	0	0	0
Coffee whitener e.g. Coffee-mate (teaspoon)	vitamins / minerals / tea etc.	0	0	0	0
Cocoa, Hot Chocolate (cup)	vitamins / minerals / tea etc.	0	0	0	0
Horlicks, Ovaltine (cup)	vitamins / minerals / tea etc.	0	0	0	0
Vitamin supplements (details)	vitamins / minerals / tea etc.	0	0	0	0
Mineral supplements (details)	vitamins / minerals / tea etc.	0	0	0	0
Other supplements	vitamins / minerals / tea etc.	1	0	0	0
Ready meal (specify)	vitamins / minerals / tea etc.	0	0	0	0
Takeaway (specify)	vitamins / minerals / tea etc.	0	0	0	0

Table 4 Assembly statistics for shotgun sequencing metagenomics.

Of the 27 assemblies, 25 were based on libraries with 91 bp paired-end 91 Illumina reads and 350 bp insert size. Samples EM039 and EM173 were assembled using 101 bp paired-end Illumina reads, from libraries with a 500 bp insert size, and in hybrid assembly with 551,726 and 665,164 454 Titanium reads, respectively. Only contigs larger than 500 bp were used for further analysis.

Sample	Number of contigs	Total contig length (bp)	Largest contig (bp)	N50 (bp)	Total number of bases
EM039	57,873	102,436,267	495,772	2,873	3,317,557,338
EM148	8,121	56,447,615	520,052	30,601	4,795,110,722
EM172	55,298	83,738,852	125,936	2,050	4,472,602,022
EM173	48,821	108,101,567	146,798	4,013	3,130,777,799
EM175	7,706	56,078,081	595,761	30,216	4,772,905,202
EM176	8,004	64,792,202	656,046	36,029	4,780,810,982
EM177	46,980	121,070,988	217,703	5,683	4,776,788,162
EM191	3,990	49,084,620	418,300	59,730	4,789,620,542
EM204	12,745	79,505,409	570,565	44,252	4,779,024,662
EM205	53,697	135,486,275	162,823	5,242	4,786,831,622
EM208	3,313	41,992,026	546,341	64,999	4,790,740,682
EM209	45,871	116,311,093	205,124	5,017	4,788,571,862
EM219	19,319	85,434,733	555,766	14,559	4,796,211,602
EM227	12,837	54,647,530	336,634	26,765	4,794,620,762
EM232	31,942	121,533,960	236,309	10,432	4,794,202,982
EM238	7,351	39,150,291	769,101	48,224	4,793,427,362
EM242	9,797	60,568,098	641,916	35,719	4,796,499,242
EM251	11,087	65,526,982	357,199	21,319	4,786,586,462
EM268	32,610	97,409,224	575,002	8,616	4,789,756,442
EM275	9,814	59,736,375	813,547	24,504	4,796,447,222
EM283	35,823	110,290,277	284,712	7,512	4,788,078,482
EM293	9,512	57,120,941	733,123	53,850	4,795,731,362
EM305	8,755	59,052,708	446,907	34,979	4,793,440,322
EM308	14,486	80,827,555	411,604	21,742	4,794,984,362
EM326	16,809	82,168,325	511,025	37,604	4,795,947,722
EM337	32,940	84,242,311	257,796	5,093	4,798,092,962
EM338	44,075	125,050,275	132,015	7,202	4,793,679,182
Sum	**649,576**	**2,197,804,580**	**11,723,877**	**NA**	**125,889,048,067**
Average	**24,058**	**81,400,170**	**434,218**	**24,031**	**4,662,557,336**

Table 5 Number of sequence matches for genes encoding three SCFA-producing enzymes against 27 assembled metagenomes.

BCoAt: Butyryl-CoA transferase / Acetyl-CoA hydrolase; ACS: Acetate-formyltetrahydrofolate synthetase/Formate-tetrahydrofolate ligase; PCoAt: Propionyl-CoA:succinate-CoA transferase / Propionate CoA-transferase. Columns titled Matching genes show number of TBLASTN hits of enzymes (BCoAt: GI|71081820|; ACS: GI|218353245| and GI|7242549|; PCoAt: GI|29346147| and GI|260588698|) with an amino acid identity greater than 30% and an alignment length longer than 200 bp.

Subject	Location	Predicted genes	Matching BCoAt genes	Matching ACS genes	Matching PCoAt genes
EM039	Community	136,611	9	13	9
EM148	Rehab	52,810	10	10	7
EM172	Community	150,449	14	18	11
EM173	Longstay	65,189	2	6	4
EM175	Community	56,117	4	11	7
EM176	Community	60,966	7	12	12
EM177	Community	156,739	12	18	17
EM191	Longstay	43,959	1	10	11
EM204	Community	78,779	8	11	10
EM205	Community	183,002	11	11	13
EM208	Longstay	37,155	0	9	6
EM209	Rehab	158,301	16	14	16
EM219	Rehab	94,356	9	13	15
EM227	Longstay	57,754	3	8	8
EM232	Rehab	141,771	18	14	13
EM238	Longstay	38,639	3	6	5
EM242	Longstay	58,756	3	9	9
EM251	Community	65,721	4	10	8
EM268	Community	123,398	13	15	7
EM275	Longstay	59,297	8	10	8
EM283	Community	137,351	13	16	9
EM293	Longstay	55,734	5	8	8
EM305	Rehab	57,727	6	10	8
EM308	Longstay	80,512	8	11	12
EM326	Rehab	85,907	8	12	8
EM337	Rehab	108,711	7	13	6
EM338	Rehab	161,152	13	13	15
Community Total		**1,149,133**	**95**	**135**	**103**
Long-stay Total		**496,995**	**33**	**77**	**71**
Community Average		**114,913**	**9.5**	**13.5**	**10.3**
Long-stay Average		**55,222**	**3.7**	**8.6**	**7.9**

Table 6 Descriptive statistics of clinical and health measurements in 178 elderly subjects across the four residence locations.

	Community (*n*=83)	Day Hospital (*n*=20)	Rehabilitation (*n*=15)	Long-stay (*n*=60)
	Mean (SD)			
Age (yrs)	73.6 (6.5)	80.9 (6.6)	77.3 (6.9)	83.6 (7.5)
Male (%)	43	65	53	23
Weight (kg)	75.4 (16.5)	79.2 (17.3)	82.6 (16.9)	59.8 (13.2)
BMI[1] (kg/m^2)	27.4 (5.0)	28.5 (6.1)	30.5 (4.9)	23.4 (4.6)
CC[2](cm)	36.2 (4.1)	37.3 (5.7)	36.3 (5.2)	29.8 (4.4)
MAC[3] (cm)	28.2 (3.2)	29.6 (3.8)	30.9 (4.1)	26.7 (3.3)
SBP[4] (mmHg)	140.6 (17.1)	136.8 (17.0)	122.3 (15.0)	128.0 (18.8)
DBP[5] (mmHg)	76.0 (10.2)	71.5 (11.4)	65.6 (9.6)	70.8 (13.0)
	Median (interquartile range)			
CCI[6] (0, 9)	0 (0, 1)	3 (2, 4)	3 (1, 4)	3 (2, 4)
Barthel[7] (0, 20)	20 (20, 20)	20 (18, 20)	12 (10, 16)	2 (1, 7)
FIM[8] (0, 126)	126 (125, 126)	124 (119, 126)	94 (82, 114)	36 (20, 61)
MMSE[9] (0, 30)	29 (27, 30)	27 (25, 29)	25 (21, 27)	14 (2, 20)
MNA[10] (0, 30)	27 (26, 28)	25 (24, 27)	25 (23, 25)	17 (14, 21)

[1]Body Mass Index, [2]Calf Circumference, [3]Mid-Arm Circumference, [4]Systolic Blood Pressure, [5]Diastolic Blood Pressure

[6]Charlson Index of Comorbidity: Score out of a total of 22. Higher scores indicate higher degree of co-morbidity

[7]Barthel Index of Activities of Daily Living: Score out of a total of 20. Higher scores indicate higher degree of independence

[8]Functional Independence Measure: Score out of a total of 126. Higher scores indicate higher degree of independence.

[9]Mini-Mental State Exam: Score range 0 (worst performance) – 30 (best performance)

[10]Mini-Nutritional Assessment: Score out of a total of 30. >/= 24 = well-nourished; 17-23.5 = at risk of malnourishment; <17 = undernourished

Table 7: Effect of medications on health measures tested for microbiota correlations

Medication types are coded as per foot-note. The Associated Difference column shows the median change in the health parameter measurement in subjects receiving the medication compared to those who did not, and is derived from the slope of the medication regression in the median regression model adjusted for age, gender, location (for all 178 subjects), unweighted UniFrac PCoA axes 1, 2 and 3, and any other medications that were added to the model for a specific health value. Asterisks indicate significant p-values: p <= 0.05 (*), p <= 0.01 (**), p <= 0.001 (***).

Unweighted UniFrac PCoA for all four residence locations (n = 178)				
Health measurement	**Medication type**	**Number of subjects**	**Associated difference**	***p*-value**
GDT	Med L	28	-0.68	0.2
	Med M	37	1.2	0.064
	Med P	10	1.15	0.13
Diastolic BP	Med C	103	3.39	0.053
	Med D	48	-1.19	0.61
IL6	Med Q	13	3.06	0.15
	Med R	56	1.58	0.18
	Med S	10	1.54	0.32
	Med T	9	4.53	0.18
IL8	Med Q	13	6.61	0.19
	Med R	56	-4.28	0.24
	Med S	10	-3.28	0.6
	Med T	9	0.87	0.87
TNFα	Med Q	13	-0.24	0.75
	Med R	56	0.51	0.40
	Med S	10	2.88	0.1
	Med T	9	0.28	0.71
Unweighted UniFrac PCoA for Community-only subjects (n = 83)				
Health measurement	**Medication type**	**Number of subjects**	**Associated difference**	**p-value**
Diastolic BP	Med C	44	2.3	0.44
	Med D	15	-3.51	0.48
GDT	Med L	5	-1.25	0.12
	Med M	7	2.89	0.012*
Unweighted UniFrac PCoA for Longstay-only subjects (n = 60)				
Health measurement	**Medication type**	**Number of subjects**	**Associated difference**	**p-value**
Barthel	Med C	29	-1.03	0.31
	Med D	13	2.15	0.21
	Med I	9	-0.95	0.55
	Med L	23	-1.48	0.15
	Med N	15	-1.2	0.38
	Med O	7	1.54	0.46
	Med Q	10	-4.74	0.009**
	Med R	44	1.05	0.41
	Med S	4	6.41	0.001**
	Med T	6	-1.96	0.33
FIM	Med C	29	-7.45	0.17
	Med D	13	15.01	0.15
	Med I	9	-10.89	0.29
	Med L	23	-2.28	0.71
	Med N	15	-5.86	0.49
	Med O	7	15.79	0.11
	Med Q	10	-17.6	0.062
	Med R	44	-1.08	0.91
	Med S	4	47.9	0.0004***
	Med T	6	-14.98	0.21
Diastolic BP	Med C	29	0.52	0.84
	Med D	13	-5.18	0.17
Systolic BP	Med C	29	-4.6	0.33
	Med D	13	-7.42	0.34
IL8	Med Q	9	-2.56	0.71
	Med R	40	-3.46	0.48
	Med S	4	-1.66	0.84
	Med T	4	-1.78	0.79
CRP	Med T	4	1.05	0.78

Med C = Cardiovascular Medications (Chronic Cardiac Failure / Angina / Hypertension)
Med D = Diuretics Med I = Bronchodilators / Inhalers medications
Med K = Arthritis (osteo and rheumatoid) medications Med L = Insomnia medications
Med M = Mood Disorders (antidepressants / psychoses / anxiety) medications
Med N = Epilepsy / Seizures / Neuropathic Analgesia medications
Med O = Parkinsons medications Med P = Alzheimers / dementia medications
Med Q = Analgesia - Opiod medications Med R = Analgesia (Non - opiod) - medications
Med S = NSAIDs (non-steroidal anti-inflammatory drug) Med T = Steroids

Table 6. Chemical shifts of metabolites identified by NMR metabolomics.

	Metabolite	Chemical shift (multiplicity) and assignment
1	Cholate	0.73 ppm (s) CH_3
2	Caprylate	0.86 ppm (t) CH_3, 1.28 ppm (m) CH_2
3	Valerate	0.88 ppm (t) CH_3, 1.27 ppm (m) CH_2, 1.51 ppm CH_2
4	Butyrate[a]	0.90 ppm (t) CH_3, 1.56 ppm (m) CH_2, 2.16 ppm (t) CH_2
5	Isovalerate	0.91 ppm (d) CH_3, 2.06 ppm (d) CH_2
6	Valine	0.99 ppm (d) CH_3, 1.04 ppm (d) CH_3
7	Leucine	0.95 ppm (t) CH_3
8	Isoleucine	0.94 ppm (t) CH_3, 1.01 ppm (d) CH_3
9	Propionate	1.06 ppm (t) CH_3, 2.18 ppm (q) CH_2
10	Threonine	1.32 ppm (d) CH_3, 3.59 ppm (d) CH
11	Lactate	1.34 ppm (d) CH_3
12	Isocaproate	0.88 ppm (d) CH_3, 1.44 ppm (m) CH
13	Alanine	1.48 ppm (d) CH_3
14	Lysine	1.72 ppm (m) CH_2, 3.02 ppm (t) CH_2
15	Acetate	1.92 ppm (s) CH_3
16	Glutamate	2.34 ppm (m) CH_2, 2.05 ppm (m) CH_2
17	Succinate	2.39 ppm (s) CH_2
18	Glutamine	2.46 ppm (m) CH_2, 2.11 (m) CH_2
19	Methylamine	2.60 ppm (s) CH_3
20	Aspartate	2.66 ppm (m) CH_2, 2.82 ppm (m) CH_2
21	Trimethylamine	2.88 ppm (s) CH_3
22	Malonate	3.11 ppm (s) CH_2
23	Taurine	3.23 ppm (t) CH_2, 3.36 ppm (t) CH_2
24	Glycine[b]	3.55 ppm (s) CH_2
25	Glucose[b]	3.25 ppm (dd) CH, 3.90 ppm (dd) CH_2, 4.63 ppm (d) CH
26	Glutarate[a]	1.79 ppm (m) CH_2
27	Lipid[b]	1.17 ppm
28	Acetone	2.23 ppm (s) CH_3
29	Phenylacetate	3.54 ppm (s) CH_3, 7.28 ppm (m) CH, 7.37 ppm (m) CH
30	Tyrosine	6.89 ppm (m) CH, 7.19 ppm (m) CH
31	Phenylalanine	CH, 7.33 ppm (m) CH, 7.37 ppm (m) CH, 7.43 ppm (m) CH
32	Formate	8.46 ppm (s) CH_2

[a]higher in faecal water of community subjects relative to long stay subjects
[b]higher in faecal water of long stay subjects relative to community subjects
Letters in brackets indicate the multiplicity of the peak. s, singlet; d, doublet; dd, doublet of doublets; t, triplet; q; quartet; m, multiplet.

Supplementary Notes

Microbiota-metabolome co-inertia analysis

To aid interpretation of the relationship between the genus-level microbiota and metabolomics datasets, we integrated them using a multivariate method known as co-inertia analysis (CIA[18]). Principal components analysis (PCA) can be applied to each of the individual datasets, and these PCAs are used as inputs for the CIA which then identifies ordinations of the two datasets that are maximally co-variant and so identifies the shared biological trends within the two datasets. The method is an unsupervised approach and is insensitive to a high variable-to-sample ratio[19].

CIA was applied to the genus-level microbiota composition and metabolome data (Supplementary Figure 6). As noted in Main text, this analysis revealed a significant relationship between the two datasets with an RV coefficient of 0.474, indicating a strong co-structure between the two datasets. The first two components of the CIA account for 58.9% of the variance in the datasets, with component 1 (horizontal) accounting for 48.4% of the variance, and the second, component 2 (vertical) accounting for another 10.5%.

Community and long-stay samples are separated from each other along the primary axis in the analysis. In addition to differentially abundant metabolites mentioned in the main text, there was an increase in glycine, glucose and lipid levels in the long-stay subject faecal metabolomes. This could be because the associated microbiota may have a reduced ability to metabolize these compounds, or that there is a reduced uptake of these nutrients by the gut in long-stay subjects, that may or may not relate their microbiota.

Using a cut-off of 1% abundance, the main genera associated with the community metabolome along the first axis were *Bacteroides*, *Faecalibacterium*, *Ruminococcus*, *Oscillibacter*, *Coprococcus* and *Sporobacter*. *Bacteroides* spp are known to produce acetate[20] but are found to be associated with diets rich in animal produces and low in fruit and veg in this and other studies[21]. This and their lack of cellulolytic ability[22] may explain the strong association with acetate along the first axis but the minor inverse association with the second axis. Instead *Sporobacter* is the genus most closely related to the production of acetate according to this analysis. When the reads mapping to this

genus were investigated, we found that over 70% could be confidently assigned to the species *Sporobacter termitidis*. *Sporobacter termitidis* is known to produce only acetate as an end fermentation product[23].

A number of genera are associated with the higher faecal levels of butyrate - *Faecalibacterium, Ruminococcus, Oscillibacter* and *Coprococcus*. The identified species that have the highest correlation and abundance with these genera are *Faecalibacterium prausnitzii*, *Ruminococcus bromii*, *Coprococcus eutactus* and *Oscillibacter valericigenes*.

Properties of bacterial co-abundance groups (CAGs).

Associations between individual genera were determined using the Kendall correlation coefficient. The full set of associations was visualized and clustered in R using the Made4 package[24] and the function *Heatplot*, whereby Hierarchical clustering with the Pearson correlation distance metric and Ward clustering was used to define co-abundant groups of genera (Supplementary Fig 12A). All significant positive associations are shown in Supplementary Fig 12B. These associations were controlled for multiple testing using the qvalue method[16] and only those with false discovery rate < 0.05 were retained (Supplementary Table 8).

The Kendall correlations were converted to a correlation distance metric and used as input for a Permutational MANOVA to determine if the CAGs were significantly different from each other. Essentially this compared strength of the correlations between the groups to correlation strengths within the groups in a pairwise manner. Permutational MANOVA[25] was performed using the vegan package in R. All five external CAGs displayed significantly different inter-relationships from each other ($p < 0.0001$), except for the relationship between the *Bacteroides* CAG and the middle *Odoribacter* CAG which was the least distinct, with a significance of 0.052. The *Bacteroides* CAG has the smallest number of significant positive correlations between its members, giving it a weak coherence, and is defined more by significant negative correlations with the other four external CAGs but not against the middle *Odoribacter* group. Therefore the *Bacteroides* CAG was strongly negatively associated with the other CAGs but not with the middle group.

The major genera that dominate the CAGs are characterized by their genus-associated abilities to produce a wide range of short-chain fatty acids, including butyrate, acetate, lactate, propionate, formate, and succinate as well as ethanol, hydrogen, and carbon dioxide. However, the *Oscillibacter* CAG shows a large proportion of genera able to produce acetate, and only the acidaminococci are reported to produce butyrate[26]. The other CAGs also contain genera known to produce acetate, but many of these fermentative genera are also able to synthesize a wider range of SCFAs when compared to the genera in the *Oscillibacter* CAG. Genera able to use the fermentation products include *Desulfovibrio* and *Methanobrevibacter* which are able to use lactate or ethanol as electron donors and acetate as a carbon source respectively. The *Desulfovibrio* may also be using the sulphate liberated by the *Akkermansia* species from mucin, as an electron acceptor.

Description of Dietary Groups (DGs)

DG1: The predominant features of this diet type include consumption of complex carbohydrates (including wholegrain breakfast cereals, breads, boiled potato), daily consumption of a wide range of fruit and vegetables and moderate (5 times/week) consumption of protein-rich white meat, fish and eggs. Red-meat was not consumed by this group while oily fish were consumed once/week with low intakes of dairy produce (approx. 3 times/week) and high-sugar/low-nutrient dense foods.

DG2: Both complex (wholegrain breakfast cereals and breads, boiled potatoes) and simple carbohydrates (white bread) were consumed frequently in this diet type. A lower variety of fruit and vegetables were consumed, less frequently (two-three times/daily) compared to DG1. Red-meat, fish or eggs were consumed daily with no consumption of oily fish. Intakes of dairy produce were lower among this group (approx. once weekly) with higher intakes of high-sugar/low-nutrient dense foods compared to diet type 1.

DG3: Overall, this diet group contained the least variety of all the DG's. Porridge and mashed potato were the main daily staples while simple carbohydrate (white bread) was consumed frequently with no inclusion of wholemeal varieties. Consumption and variety of fruit and vegetables were lowest overall among this group (approx. 1-2 portions daily). Processed and unprocessed meat and eggs were

consumed once-twice/week with no fish consumed. Intakes of dairy produce are low (approx. 3 times week) with the lowest consumption of high-sugar/low-nutrient dense foods.

DG4: Similar to diet type 3, mashed potato and porridge were the only staples in this diet type and were consumed daily. Fruit and vegetable consumption was comparable with that of diet group 1 but with much less variety. Processed and unprocessed meat and fried fish were consumed once-twice/week. Fish were not consumed by this group. Dairy products and high-sugar/low-nutrient dense foods were consumed most frequently by this cohort (once-twice and 3-4 times daily, respectively), predominantly full-fat dairy produce and puddings, and sweetened hot beverages which were consumed 1-3 times daily.

Analysis of possible confounders of microbiota-health associations

Antibiotics. None of the subjects had been treated with antibiotics in the month prior to sampling; extending this exclusion window would impact negatively on recruitment rates of older subjects. To examine the effect of treatment prior to this exclusion window, we used quantile (median) regression to examine the association between microbiota alpha diversity and time in days since last antibiotic usage, extending before the one-month recruitment exclusion window. We also tested correlations between the (microbiota) PCoA axes and time since last antibiotic usage. These models were adjusted for age, gender and location, (as for the main microbiota-health analysis for Table 1). No relationship between microbiota diversity and time since last antibiotic treatment was established (data not shown).

Other medications. We adjusted median regression models for the effect of medications on cognate health read-outs. The actual effects of medication on relevant clinical scores in that model are presented in Supplementary Table 8. The majority of medications had small non-significant effects on the relevant health measure, which did not affect the significance of the relationship between the health variables and the microbiota presented in Table 1.

Diet. Although diet clearly impacts on microbiota; the composition of the diet was expected have a direct health effect independent of a diet-microbiota-health effect. To explore this issue, we repeated the median regression analysis, controlling for diet, using the Health Food Diversity (HFD) index. Although diet explained a number of the associations, the majority of the frailty indicators-microbiota

associations reported in Table 1 were still significant. Two variables, weight and BMI no longer had significant associations with microbiota after adjustment for HFD.

Aggregate microbiota composition in the elderly

We previously showed that there was a higher *Firmicutes*/*Bacteroidetes* ratio among nine younger subjects compared to 118 elderly subjects[2]. When we analysed 178 additional faecal samples from elderly subjects, and four additional samples from younger healthy individuals, we did not observe an aggregate microbiota dominated by phylum *Bacteroidetes* (Suppl. Fig 15). At genus level there were significantly higher levels of *Ruminococcus* and *Blautia* and lower levels of *Escherichia*/*Shigella* in young controls compared to the microbiota of the elderly subjects (Suppl. Fig 16). Clearly the balance in a study population of individuals from the *Bacteroidetes-Prevotella-Ruminococcus*-dominated groups (see main text) will affect aggregate estimations. To adequately infer microbiota differences between young and old populations a larger number of samples from the former cohort are needed, including information about food frequency consumption, which affects phylum proportions (ref.[21]; this study).

Denoised versus Un-denoised pyrosequencing analysis

Initially we denoised the pyrosequencing sequences using Denoiser[27] as part of the OTU picking approach according to recommendations from the QIIME documentation. We detected a run-specific bias in the denoised dataset when comparing OTU clustering of OTUs generated with and without a denoising step. Since OTUs were picked separately for each run at flow-gram level, prior to an overall OTU picking at sequence level, an over-clustering within each run seemed to have taken place. This happened in spite of using the recommended “exact” options for the UClust OTU picker. Suppl. Fig. 17 shows the run-specific bias and also how the clustering according to community location is retained for both denoised and un-denoised data sets.

Supplementary References

1 Nyamundanda, G., Brennan, L. & Gormley, I. C. Probabilistic principal component analysis for metabolomic data. *BMC Bioinformatics* **11**, 571, (2010).

2 Claesson, M. J. *et al.* Composition, variability, and temporal stability of the intestinal microbiota of the elderly. *Proc. Natl. Acad. Sci. U S A* **108 Suppl 1**, 4586-4591, (2011).

3 Harrington, J. *et al.* Sociodemographic, health and lifestyle predictors of poor diets. *Public Health Nutr.*, 1-10, (2011).

4 McCance, R. A. & Widdowson, E. M. *The composition of foods*. 6th ed. edn, (Royal Sociey of Chemistry, 2002).

5 Claesson, M. J. *et al.* Comparative analysis of pyrosequencing and a phylogenetic microarray for exploring microbial community structures in the human distal intestine. *PLoS One* **4**, e6669, (2009).

6 Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7**, 335-336, (2010).

7 Lilburn, T. G. & Garrity, G. M. Exploring prokaryotic taxonomy. *Int J Syst Evol Microbiol* **54**, 7-13, (2004).

8 Caporaso, J. G. *et al.* PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics* **26**, 266-267, (2010).

9 Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490, (2010).

10 Hamady, M., Lozupone, C. & Knight, R. Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *ISME J* **4**, 17-27, (2010).

11 Namiki, T., Hachiya, T., Tanaka, H. & Sakakibara, Y. in *ACM Conference on Bioinformatics Computational Biology and Biomedicine* (Assoc. Comput. Machinery, Chicago, Illinois, 2011).

12 Chevreux, B. *et al.* Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res.* **14**, 1147-1159, (2004).

13 Noguchi, H., Park, J. & Takagi, T. MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res.* **34**, 5623-5630, (2006).

14 Koenker, R. & Basset, G. Regression quantiles. *Econometrica* **46**, 33-50, (1978).

15 Feng, X. D., He, X. M. & Hu, J. H. Wild bootstrap for quantile regression. *Biometrika* **98**, 995-999, (2011).

16 qvalue: Q-value estimation for false discovery rate control (2010).

17 Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. Royal Stat. Society* **57**, 289-300, (1995).

18 Fagan, A., Culhane, A. C. & Higgins, D. G. A multivariate analysis approach to the integration of proteomic and gene expression data. *Proteomics* **7**, 2162-2171, (2007).

19 Jeffery, I. B. *et al.* Integrating transcription factor binding site information with gene expression datasets. *Bioinformatics* **23**, 298-305, (2007).

20 Macfarlane, S. & Macfarlane, G. T. Composition and metabolic activities of bacterial biofilms colonizing food residues in the human gut. *Appl. Environ. Microbiol.* **72**, 6204-6211, (2006).

21 Wu, G. D. *et al.* Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334**, 105-108, (2011).

22 Flint, H. J., Scott, K. P., Duncan, S. H., Louis, P. & Forano, E. Microbial degradation of complex carbohydrates in the gut. *Gut Microbes* **3**, 1-18, (2012).

23 Fardeau, M.-L. *et al.* Isolation and characterization of *Sporobacter termitidis* gen. nov., sp. nov. from the digestive tract of the wood-feeding termite *Nasutitermes lujae*. . *Int. J. Sys. Bacteriol.* **46** 512-517, (1996).
24 Culhane, A. C., Thioulouse, J., Perriere, G. & Higgins, D. G. MADE4: an R package for multivariate analysis of gene expression data. *Bioinformatics* **21**, 2789-2790, (2005).
25 Anderson, M. J. A new method for non-parametric mutivariate analysis of variance. *Austral. Ecology* **26**, 32-46, (2001).
26 Jumas-Bilak, E. *et al. Acidaminococcus intestini* sp. nov., isolated from human clinical samples. *Int. J. Sys.t Evol. Microbiol.* **57**, 2314-2319, (2007).
27 Reeder, J. & Knight, R. The 'rare biosphere': a reality check. *Nat. Methods* **6**, 636-637, (2009).