Systematic and Applied Microbiology 41 (2018) 1–12

Contents lists available at ScienceDirect

Systematic and Applied Microbiology

journal homepage: www.elsevier.de/syapm

A metagenomic approach to dissect the genetic composition of enterotypes in Han Chinese and two Muslim groups

Jing Li[a,b,*,1], Ruiqing Fu[a,c,1], Yajun Yang[d], Hans-Peter Horz[e], Yaqun Guan[f], Yan Lu[a], Haiyi Lou[a], Lei Tian[a,c], Shijie Zheng[a], Hongjiao Liu[a,g], Meng Shi[a,c], Kun Tang[a], Sijia Wang[a], Shuhua Xu[a,c,h,i,*]

[a] Chinese Academy of Sciences (CAS) Key Laboratory of Computational Biology, Max Planck Independent Research Group on Population Genomics, CAS-MPG Partner Institute for Computational Biology (PICB), Shanghai Institutes for Biological Sciences, CAS, Shanghai 200031, China
[b] School of Life Science and Technology, China Pharmaceutical University, Nanjing 210009, China
[c] University of Chinese Academy of Sciences, Beijing 100049, China
[d] State Key Laboratory of Genetic Engineering and Ministry of Education (MOE) Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, Shanghai 200433, China
[e] Institute of Medical Microbiology, RWTH Aachen University Hospital, 52074 Aachen, Germany
[f] Department of Biochemistry, Preclinical Medicine College, XinJiang Medical University, Urumqi 830011, China
[g] Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD 21218, United States
[h] School of Life Science and Technology, ShanghaiTech University, Shanghai 201210, China
[i] Collaborative Innovation Center of Genetics and Development, Shanghai 200438, China

ARTICLE INFO

Article history:
Received 27 June 2017
Received in revised form
12 September 2017
Accepted 13 September 2017

Keywords:
Metagenome
Next-generation sequencing
16S rRNA
Enterotype
Genome-wide association study

ABSTRACT

Distinct enterotypes have been observed in the human gut but little is known about the genetic basis of the microbiome. Moreover, it is not clear how many genetic differences exist between enterotypes within or between populations. In this study, both the 16S rRNA gene and the metagenomes of the gut microbiota were sequenced from 48 Han Chinese, 48 Kazaks, and 96 Uyghurs, and taxonomies were assigned after *de novo* assembly. Single nucleotide polymorphisms were also identified by referring to data from the Human Microbiome Project. Systematic analysis of the gut communities in terms of their abundance and genetic composition was also performed, together with a genome-wide association study of the host genomes. The gut microbiota of 192 subjects was clearly classified into two enterotypes (*Bacteroides* and *Prevotella*). Interestingly, both enterotypes showed a clear genetic differentiation in terms of their functional catalogue of genes, especially for genes involved in amino acid and carbohydrate metabolism. In addition, several differentiated genera and genes were found among the three populations. Notably, one human variant (rs878394) was identified that showed significant association with the abundance of *Prevotella*, which is linked to *LYPLAL1*, a gene associated with body fat distribution, the waist-hip ratio and insulin sensitivity. Taken together, considerable differentiation was observed in gut microbes between enterotypes and among populations that was reflected in both the taxonomic composition and the genetic makeup of their functional genes, which could have been influenced by a variety of factors, such as diet and host genetic variation.

Introduction

The gut microbiota, mainly bacteria, plays important roles in balancing the immunity and nutritional system of the host, and affects the human health status through multiple host-bacteria interactions. However, the gut microbiota is a very complex ecosystem, encompassing approximately 100 trillion bacterial cells representing more than 1000 species that possess millions of bacterial genes [32]. Although many genes of the microbiome belong to low abundance organisms it remains to be elucidated whether they

* Corresponding authors at: Chinese Academy of Sciences (CAS) Key Laboratory of Computational Biology, Max Planck Independent Research Group on Population Genomics, CAS-MPG Partner Institute for Computational Biology (PICB), Shanghai Institutes for Biological Sciences, CAS, Shanghai 200031, China.
E-mail addresses: lj.cpu@126.com (J. Li), xushua@picb.ac.cn (S. Xu).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.syapm.2017.09.006
0723-2020/

are insignificant for gut ecosystem functioning or whether they represent a "rare biosphere" containing important key stone species [34]. Therefore, the factors representing the forces that drive, shape and maintain the balance of the gut bacterial community represent one of the key questions for current gut microbiome studies [36].

Another emerging question in microbiome studies is to what extent the genetic background of the human host affects the development and stability of the gut microbiome [10]. Although there have been many gut metagenomic studies, such as the Human Microbiome Project (HMP) [27], Metagenomics of the Human Intestinal Tract (MetaHIT) [6], and the BGI's gut meta project [30], it is still unclear to what extent differences in the gut microbiome observed among different human populations [9,24,25] are due to host genetic differences or other factors (e.g. food). Diet has been considered as the major factor that shapes the human gut microbiome [8], and it was reported that different diets are directly associated with distinct gut bacterial compositions (i.e. different enterotypes). For instance, the *Bacteroides* enterotype is associated with a diet rich in protein and animal fat, while the *Prevotella* enterotype is associated with a carbohydrate-enriched diet [39]. However, the definition of enterotype is based on classifying the abundances of the distinct gut bacteria that, in addition, may also be affected by factors other than diet. Moreover, it is still a matter of debate whether the gut microbiota can be truly distinguished into discrete enterotypes or rather "enterotype gradients" [14].

This study used 192 college students from the same university that largely lived in the same environment, were in the same age range, and were in good health (Table S1). The metagenomes of the gut microbiota from 192 samples were sequenced and analyzed to investigate the genetic composition of enterotypes, and further explore the impact of host genetic variation on the composition of the human microbiome.

Materials and methods

Ethical approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee, as well as the 1964 Helsinki declaration, its later amendments or comparable ethical standards.

Informed consent

Informed consent was obtained from all individual participants included in the study.

Sample collection and processing

A total of 386 individuals were recruited, including 65 Han Chinese (HAN), 53 Kazaks (KZK), 235 Uyghurs (UIG) and 33 individuals from other ethnic groups, to voluntarily provide blood (~2 mL), saliva (~2 mL), and stool samples (~2 g). None of the participants had any clinical symptoms and they had not used any antibiotics for one month, according to their self-report declaration. Specimen collection was undertaken in the morning after the participants had stopped eating, drinking and performing oral hygiene 8 h before sampling. Each sample was frozen immediately at −80 °C, and all samples were refrigerated and transported to the laboratory in Shanghai within one week, stored at −80 °C, and used for extracting DNA within four weeks.

DNA extraction

DNA from human blood samples was extracted using the QIAamp® DNA Blood Mini Kit (Qiagen). The DNA from stool samples was extracted by a 3-step procedure, according to the method of Yuan et al. [40]. Briefly, cell lysis was carried out with a cocktail of enzymes (Sigma–Aldrich), followed by bead beating (BioSpec) and extraction with the QIAamp® DNA Mini Kit (Qiagen). Since some participants provided only partial samples, high quality DNA samples were finally selected from 48 HAN, 48 KZK, and 96 UIG for further processing, in which two human blood DNA samples were missing.

Human DNA genotyping and processing

Human DNA genotyping was performed on an Illumina Human OmniZhongHua-8 SNP Array, and the raw intensity data were analyzed with GenomeStudio. After excluding the individuals with a genotype call rate below 90%, SNPs with missing data >10% and SNPs in each population that failed the Hardy–Weinberg equilibrium test ($p < 0.0001$), 859,598 autosomal SNPs were obtained for further analysis. Principal component analysis (PCA) was performed at the individual level using EIGENSOFT V.3.0 [26,28].

16S rRNA gene sequencing and processing

The V1–V3 variable region of the microbial 16S rRNA gene from the DNA extracted from stool samples was amplified with the forward primer for V1 and the reverse primer for V3, and the PCR primers and PCR conditions used were the same as in a previous study [35]. The ~570 bp amplicons were prepared for a sequencing library and paired-end sequencing was performed on an Illumina MiSeq platform for 2×300 cycles with v3 reagents, according to the manufacturer's instructions.

The initial sequences with the correct barcode were assessed and filtered according to the base quality of ($q = 20$, $p = 80$) using FASTX-Toolkit (v0.0.14). Then, the paired-end reads passing the quality filter were merged, and the Chimera sequences were checked and removed by the ChimeraSlayer approach implemented in the QiiME package [4]. To obtain read depths at a comparable level, 20,000 sequences were subsampled from each individual and, after pooling them, the sequences were collapsed into OTUs at an identity level of 0.97. OTUs hit by less than four sequences were removed for the sake of consensus, and then a representative sequence set was built from the pooled sequences for each OTU. Thereafter, the representative sequence set was aligned with the Greengenes core set using the *PyNAST* method implemented in QiiME for taxonomic assignments and relative abundance calculations, as described previously [24].

The distribution of variations based on the frequency distribution of taxa within and between individuals (i.e. analysis of molecular variance; AMOVA) was calculated with Arlequin 3.5 [7]. Alpha-diversity of the gut microbiota was indicated by the results of the rarefaction workflow using QiiME. In detail, rarefied OTU tables from 100 to 10,000 sequences per individual were constructed in steps of 100 sequences, and then the average number of OTUs from ten iterations was used to indicate the alpha diversity of each rarefied OTU table. The beta diversity of the core OTU set (i.e. the OTUs identified in at least 80% (153) of individuals) was indicated by the Sørensen index using the "vegan" package in R. According to the enterotyping tutorials provided by Arumugam et al. (http://enterotype.embl.de/index.html; [1]), the enterotyping of the data was also performed based on the distance matrix calculated from the relative abundance of each OTU or taxon in each sample. The Jensen–Shannon distance (JSD) was used and the partitioning around medoids (PAM) algorithm was applied for par-

titioning all individuals into K groups, while the optimal K was indicated by the Calinski–Harabasz (CH) index. The enterotyping results were visualized by principal coordinates analysis (PCoA).

Metagenome sequencing and processing

The quality checked metagenomic DNA from each stool sample was used for library construction and the paired-end sequencing was performed on an Illumina Hiseq 2500 platform for 2×101 cycles, according to the manufacturer's instructions. The raw paired-end sequencing read duplications were removed with custom script, mapped to the human reference genome (1000 genomes project, v37) using BWA (v0.7.5a) [19] with default settings, and then the human source sequencing reads were removed by SAMtools (v0.1.19) [18,20]. Thereafter, the pure microbiome sequencing reads were assessed and filtered according to the base quality (q = 20, p = 80) using FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit).

The high quality microbiome sequencing reads were assembled by the SOAPdenovo2 package (v2.04) [22] using the parameters avg_ins = 250, K = 63, k = 45, R = Y, M = 3, and others as default settings per individual. After assembling, the contigs with at least 500 bp were further used to predict the genes by MetaGeneMark (v2.8) [41], and then a non-redundant gene set was constructed by pair-wise comparison of all gene sequences identified from 192 individuals using BLAT (v. 35×1) [16] with 95% identity and 90% overlapping thresholds.

The entire translated protein sequences of the non-redundant gene set were locally aligned to the NCBI-NR database using BLASTP (v2.2.29+) [3] and a parameter e-value = 1e-5. Based on the blasting results, the taxonomic assignments and functional annotations of each sequence (i.e. KEGG catalogue) were implemented by the lowest common ancestor (LCA) algorithms in MEGAN5 (v5.2.3) [13].

The high quality microbiome sequencing reads from each individual were aligned to the non-redundant gene set by SOAPaligner in SOAP2 (v2.21) [21] with parameters of r = 2, m = 150, x = 350, and v = 5. The relative abundance of each gene in each individual was calculated by the number of read pairs mapped to the gene over the length of the gene divided by the sum of gene abundance per individual, which was described in detail previously [31]. Furthermore, the relative abundance of each taxonomic or functional group was calculated by the sum of the relative abundance of genes within the group for each individual.

Genetic variation landscape of the metagenome

The reference genomes of 1751 bacterial strains representing 1253 species were obtained from the Human Microbiome Project (HMP) in September 2014. Then the high quality microbiome sequencing reads from 192 individuals were mapped to these reference genomes using Mosaik with the parameters a = all, m = all, hs = 15, mmp = 0.95, mmal = Y, minp = 0.9, mhp = 100, and act = 20, all of which were identical to a previous study [33]. By multiple-pileup of all the alignment results together, a reference genome was considered for further processing by two criteria: first, the cumulative depth of the genome should be $\geq$600X for all individuals, meaning that the average sequencing depth was >3X for each individual, and second, at least one individual covered at least 40% of the whole genome length. Subsequently, the Bayesian model-based approach (i.e. bcftools [20]) was used to call the SNPs from the pooled alignment results for the 111 most enriched bacterial strains. For this, the parameters were set as follows: c = Y, N = Y, e = Y, g = Y, v = Y, and ploidy = 1. SNPs with minor allele frequency <0.02, as well as missing data >20%, were filtered out for further processing.

Principal component analysis (PCA) and neighbor-joining (NJ) tree reconstruction were used to measure the overall bacterial genetic differentiation between individuals. The PCA was performed on all qualified SNPs from 111 bacterial genomes at the individual level using EIGENSOFT, while the NJ tree was built from individual pairwise distances calculated from all qualified SNPs, as described previously [12].

The unbiased F_{ST} following Weir and Cockerham [38] was used to measure the detailed bacterial genetic differentiations between groups (e.g. different ethnic groups, enterotypes, and genders). Particularly, the F_{ST} values of SNPs were calculated for each locus, while the F_{ST} values of 111 bacterial genomes (or genes predicted from those genomes) were the average *F*-statistic over all loci within genes or genomes. To measure the significance of F_{ST} for bacterial genomes, permutation tests (1000 iterations) were performed by randomly shuffling the individuals among groups, and the top 5% highest randomly permutated F_{ST} values for each bacterial genome were set as thresholds.

The ratio of non-synonymous (NS) and synonymous (S) substitutions (i.e. *pN/pS* ratio) was calculated for the 111 bacteria genomes and all genes within these genomes following the method from a previous study [33]. Briefly, the genes and related proteins from the 111 bacteria genomes were predicted by MetaGeneMark. The expected NS and S substitutions were then counted from all possible mutation results of codon changes within genes or genomes, while the observed NS and S substitutions were identified from all qualified SNPs by comparing the genetic variations to the respective codons within the reference genomes. Thereafter, the ratio of *pN* (observed NS over expected NS substitutions) to the *pS* (observed S over expected S substitutions) was calculated for each gene and genome.

For all the genes predicted from the 111 bacterial genomes, the non-redundant gene set was built using BLAT with 95% identity and 90% overlapping thresholds. By ranking the F_{ST} values of these non-redundant genes, the top 40 genes (0.05% of the total of 81,579 non-redundant genes from reference genomes) with the highest F_{ST} values between enterotypes were identified. Furthermore, the median-joining (MJ) haplotype network composed of all NS mutations from the identified SNPs for each gene was constructed using Network (v4.6) (http://www.fluxus-engineering.com), and the 3D structures of homologous proteins according to these 40 genes were extracted from the Protein Data Bank (PDB) database (http://www.rcsb.org) using BLASTP and exhibited using PyMOL software (https://www.pymol.org).

Statistical analysis

The significance of relative abundance distributions between different taxa or functional catalogues was measured by the Mann–Whitney U test (for two groups) and the Kruskal–Wallis test (for more than two groups), and all *P* values were adjusted by Benjamini–Hochberg (BH) correction, which were all performed using R packages. The associations between bacterial relative abundance and human genotypes were calculated by the linear regression model using PLINK (v1.07)

Results

Composition analysis and complexity of the gut microbiota

The principal component analysis (PCA) [26,28] of human genotypes clearly revealed the distinct genetic difference between the three populations Han Chinese (HAN), Kazak (KZK), and Uyghur (UIG) (Fig. S1). By deep sequencing the V1-V3 region of 16S rRNA gene amplicons from fecal samples [33], a total of 9,171,286 paired-

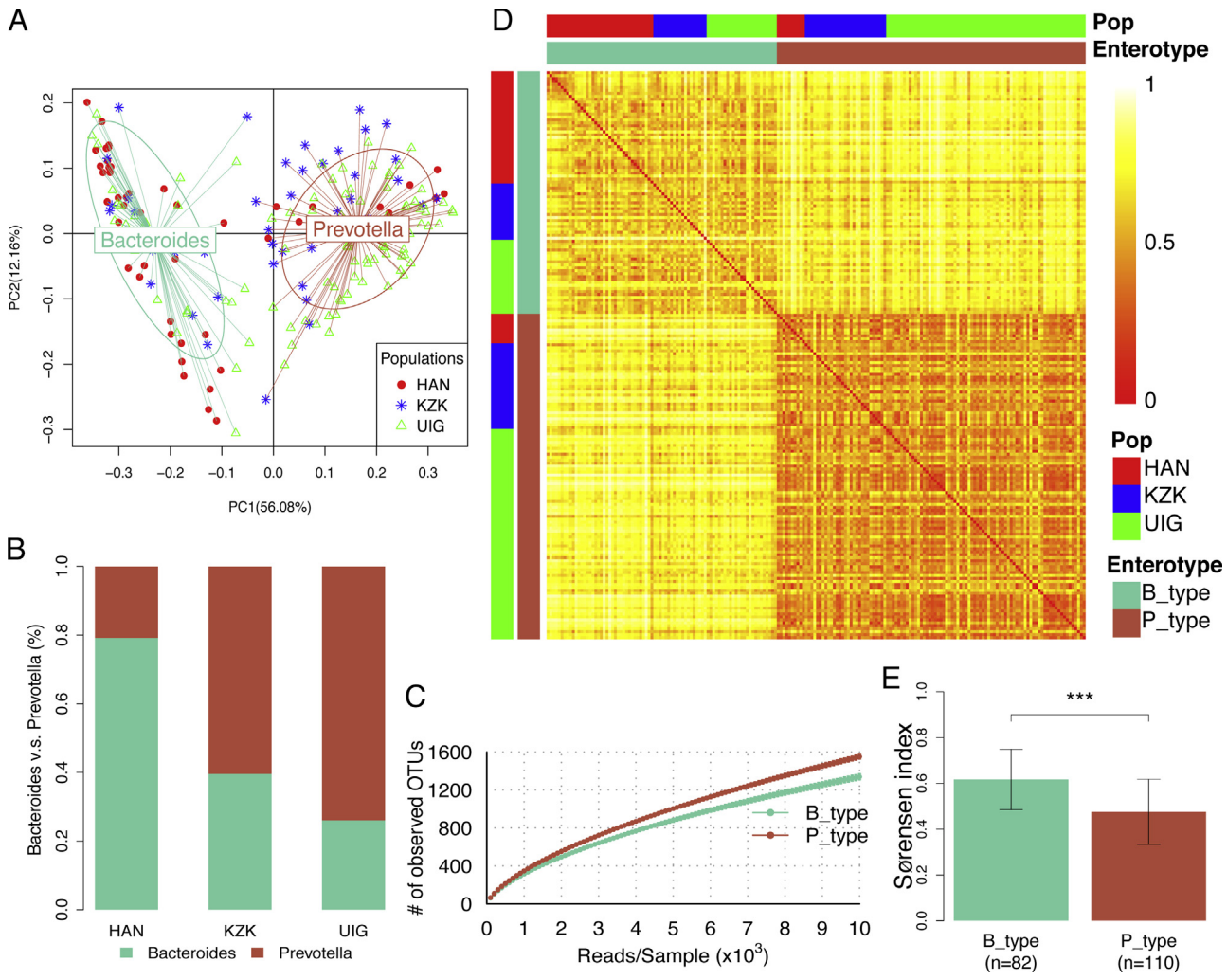


Fig. 1. Enterotype clustering and diversity analysis.
(a) Principal coordinates analysis (PCoA) of 192 individuals based on the composition of bacterial genera in the gut. (b) Proportions of enterotypes in three different human populations. (c) Rarefaction curves based on OTUs observed in individuals from the B_type and P_type, respectively, in which the error bars indicate standard deviation of observed OTUs. (d) Heatmap plot for the inter-individual dissimilarity of the gut microbiota among 192 individuals calculated by the Sørensen index. The dissimilarity is indicated by color according to the scale bar beside the heatmap. (e) Dissimilarity among groups based on the Sørensen index between the B_type and P_type. In the figures, '*' represents $P < 0.05$, '**' represents $P < 0.01$, and '***' represents $P < 0.001$ after BH correction.

end and post-trimmed 16S rRNA sequences were obtained from 192 individuals, with an average length of 518 bp (~90% of the sequences were 500–540 bp) passing the quality control. This resulted in a range of 23,716–401,662 sequences per individual (Table S2). Considering the high variation of sequence numbers among the samples, ~20,000 sequences were randomly subsampled from each sample for downstream analysis [4]. As shown in Table 1, these sequences could be assigned to 14 phyla, 23 classes, 35 orders, 55 families, 94 genera, and 100 core OTUs (present in at least 80% of individuals). Furthermore, there were 84 genera in the HAN population, 84 genera in KZK, and 90 genera in UIG. An analysis of molecular variance (AMOVA) [7] was then carried out at each taxon level to investigate how much of the total variation in the gut microbiome was due to the differences within vs. between individuals from each group. The results (Table 1) showed that most of the variance came from differences within individuals (72.39–90.17% of the total variance), while the largest variance among the three populations was at the genus level (5.93% of the total variance). The relative abundance of the top 25 most abundant genera (i.e. larger than 0.5% in at least one individual) belonged to four phyla: *Actinobacteria*, *Bacteroidetes*, *Firmicutes*, and *Proteobacteria* (Fig. S2A). While 17 genera belonged to *Firmicutes*, only three genera belonged to *Bacteroidetes*, two of which showed the overall highest abundance, ranging from 0.18% to 80.91% in the case of *Bacteroides* and 0–83.56% in the case of *Prevotella*. The abundance of nine genera was significantly different between the three populations (Fig. S2B), with HANs exhibiting the highest abundance in four of the nine genera (i.e. *Bacteroides*, *Blautia*, *Sutterella*, and *Streptococcus*) but also the lowest abundance in five other genera (i.e. *Prevotella*, *Megasphaera*, *Succinivibrio*, *Catenibacterium*, and *Lactobacillus*). Based on the genus level, principal coordinates analysis (PCoA) classified the gut microbiota of the 192 individuals into two distinct enterotypes [1]: one group was dominated by the genus *Bacteroides* (termed B_type hereafter) and the other was dominated by the genus *Prevotella* (termed P_type) (Figs. 1 A, S3A, and S3B). The clustering of the human gut microbiota exhibited a clear population structure: 79.2% of HANs clustered in the B_type and the remaining 20.8% in the P_type, while nearly the opposite was true for UIGs (i.e. B_type: 26.0% and P_type: 74%) (Fig. 1B). The proportion of these two enterotypes in KZKs was intermediate (i.e. B_type: 39.6% and P_type: 60.4%). The differences between B_type and P_type were not restricted to only a few dominant genera, for instance, 16 of the 25 most abundant genera exhibited significant differences between the two enterotypes (Fig. S4A). Quantitatively, the vari-

Table 1
Prevalence distributions at different taxa levels and AMOVA for three human populations and two enterotypes.

Taxa	Total (192)	HAN (48)	KZK (48)	UIG (96)	B_type (82)	P_type (110)	AMOVA (Grouping as populations; avg.)			AMOVA (Grouping as enterotypes; avg.)		
							Between groups (%)	Within group (%)	Within indiv. (%)	Between groups (%)	Within group (%)	Within indiv. (%)
Phylum	14	13	13	14	14	14	0.46	14.58	84.96	5.92	11.53	82.55
Class	23	21	22	23	22	23	0.44	13.34	86.22	5.56	10.52	83.93
Order	35	33	34	35	35	35	0.45	13.36	86.19	5.57	10.53	83.89
Family	55	51	52	54	55	54	3.21	16.42	80.37	18.59	7.76	73.65
Genus	94	84	84	90	93	91	5.93	21.68	72.39	31.12	6.88	62.0
Core_OTUs	100	100	100	100	100	100	1.16	8.67	90.17	7.74	5.21	87.05

ance between the two enterotypes was 31.12% of the total variance at the genus level based on AMOVA, as shown in Table 1, which demonstrated that differences in the bacterial community composition between the two enterotypes were significantly higher than the differences observed among the three human populations. Therefore, the differentiations of the gut microbiota among the three human populations were mainly due to the uneven distribution of the two enterotypes. However, the abundance of several genera was still significantly different among the three human populations when looking within each enterotype independently. For instance, although no genus showed a significant difference among the three populations within the B_type (Fig. S4B), the abundance of *Bacteroides* and *Sutterella* still differed significantly among populations within the P_type (Fig. S4C).

In order to elucidate further the differences between the enterotypes and the three human ethnic groups, alpha- (intra-individual) (Figs. 1 C, S5A and S5B) and beta- (inter-individual) (Figs. 1 D, S5C and S5D) diversity analyses were performed at the OTU level. As can be seen from the rarefaction curves, the P_type was characterized by a higher number of OTUs than the B_type (Fig. 1C). Inter-individual dissimilarity of the gut microbiota was calculated using the Sørensen index and was displayed as a heatmap diagram (Fig. 1D), which demonstrated that the microbiota within each enterotype was highly homogeneous. Furthermore, the beta-diversity was lower in the P_type compared to the B_type (P value < 2.2e-16) (Fig. 1E).

Functional profiling of the gut metagenome

To explore the functional profiles of the gut microbiome, metagenome shot-gun sequencing was performed and 13,428,860,552 raw pair-end reads (~1356 Gb) were obtained from the same 192 individuals (Table S2). After removing duplications, reads from human sources [18–20] and low quality reads with an average of 5.25 Gb high quality sequence data for each individual were obtained. SoapDenovo2.0 [22] was used to assemble an average of ~63,644 contigs per individual, with ~120,551 genes per individual being identified by MetaGeneMarker [41]. A total of 23,145,749 genes were obtained from all 192 individuals, from which 2,928,862 non-redundant genes were extracted [16]. Furthermore, a total of 2,476,725 genes (84.6%) matched with genes in the NCBI nr database [3]. Of these genes, 2,038,055 (69.6%) could be assigned [13] at the phylum level, 1,804,463 (61.6%) at the class level, 1,774,882 (60.6%) at the order level, 1,280,114 (43.7%) at the family level, and 1,124,250 (38.4%) at the genus level. When relating the shot-gun sequence data to each sample [21,31], the abundance distribution at the genus level could be reconstructed for each individual (Fig. S6A). This assignment based on the metagenomic data was highly consistent with the 16S rRNA gene data. The Pearson's correlation coefficient of the relative abundance of each genus across the samples for the two different sequence data sets was 0.9103 (P value < 2.2e-16) (Fig. S6B). In particular, the two enterotype clusters, after applying the same procedure to the NGS sequence data as to the 16S data, were highly consistent with the result based on 16S data, with only several individuals assigned differently (Fig. S6C).

A total of 1,009,933 out of 2,928,862 non-redundant genes (34.5%) were identified by KEGG catalogs, and the two enterotypes showed significant differences for the top 25 most abundant catalogs (Fig. 2A–C). The B_type gut microbiome in particular exhibited higher abundance of "Amino Acid Metabolism" related genes, while the P_type gut microbiome exhibited higher abundance of "Carbohydrate Metabolism" related genes. This was consistent with the common knowledge that the long-term diets of the *Bacteroides* enterotype are enriched for protein and animal fat, while the

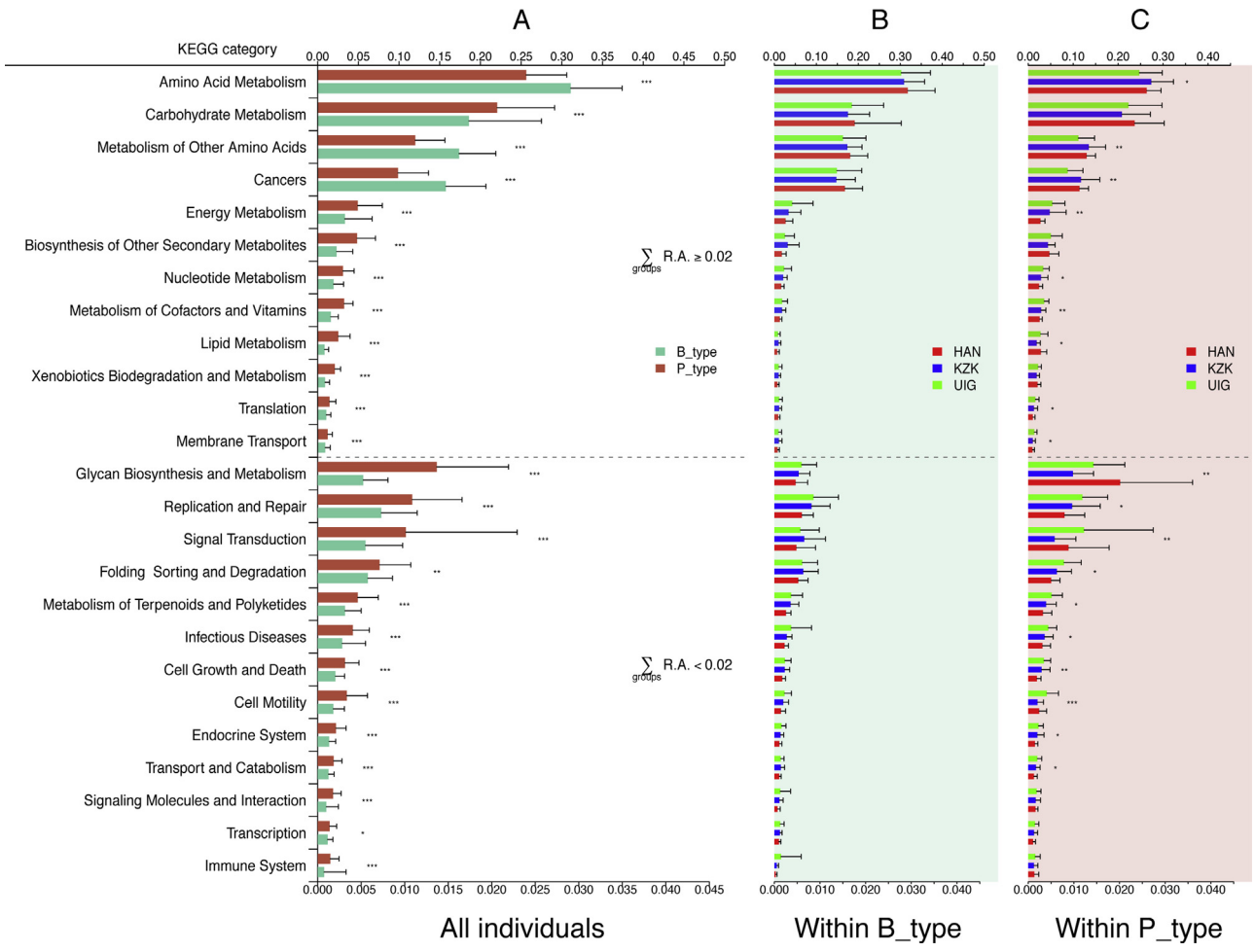


Fig. 2. Relative abundance of different functional categories present in the human gut microbiome.
Relative abundance (R.A.) of dominant genes identified in KEGG. (a) Comparison between the B_type and P_type for all individuals. (b) The three populations within the B_type. (c) The three populations within the P_type. The top 25 categories by R.A. for 192 individuals are displayed. In the figures, '*' represents $P < 0.05$, '**' represents $P < 0.01$, and '***' represents $P < 0.001$ after BH correction.

long-term diets of the *Prevotella* enterotype are enriched for carbohydrate [39].

In addition, further details for the functional gene differences between the two enterotypes were shown. For instance, based on the CAZy (Carbohydrate-Active enZymes) database, the B_type exhibited higher abundances at GH92 and GH20, which are associated with animal glycan, while the P_type exhibited significantly higher abundance at GH13, which is associated with starch and glycogen [15] (Fig. S7A–S7C). Based on the ARDB (Antibiotic Resistance Genes Database), the abundances of 14 catalogs were significantly different between the two enterotypes, with the B_type being characterized by the higher abundances of 12 catalogs (Fig. S8A–S8C).

Genetic landscape of the gut microbiome

The above analyses demonstrated the significant differences between the two enterotypes with respect to taxonomic composition and functional genes. Subsequently, in order to go further and reveal this differentiation on the basis of bacterial genomes (i.e. metagenomes), the genetic diversity patterns between the two enterotypes were investigated by mapping the high quality sequencing reads to the 1728 full bacterial reference genomes [33]. Taking the accumulated depth and coverage of the reference genomes as thresholds, the 111 most enriched bacterial strains were identified and subsequently treated as a core set of bacteria that represented a total of 54,874,539 SNPs of the reported bacterial genomes [20]. Using the stringent criteria of missing data less than 20% and minor allele frequencies larger than 0.02 at each locus, 15,304,848 filtered SNPs remained from the 111 reference genomes (Fig. 3 and Table S3). The depth of 111 reference genomes in all 192 individuals ranged from 606X (BACT_224: *Blautia hansenii*) to 10,389X (BACT_193: *Bacteroides dorei*) (i.e. the average depth ranged from 3.2X to 54.1X for different bacterial genomes per individual, as shown by the bottom two panels in Fig. 3). The density of filtered SNPs for the 111 reference genomes ranged from 0.14 SNPs per kb (BACT_534: *Escherichia* sp. 4_1_40B) to 152.3 SNPs per kb (BACT_545: *Faecalibacterium prausnitzii*), and more detailed information is given in Table S3.

The two commonly used measurements (i.e. the F_{ST} [38] and the *pN*/*pS* ratio [33]) were used to elucidate the genetic differentiations between the two enterotypes (the upper two panels of Fig. 3). The F_{ST} is frequently applied to estimate the genetic difference between/among groups, via the difference of allele frequency, and the *pN*/*pS* ratio, adapted from the *dN*/*dS* ratio, is indicative of selection, by testing if the ratio of the proportion of non-synonymous mutations (N) and the proportion of synonymous mutations (S) obviously deviates from 1. All F_{ST} values from the two enterotypes (red curve) were higher than the empirical top 5% F_{ST} values from 1000 permutated calculations for each reference genome. Furthermore, the average F_{ST} values calculated from

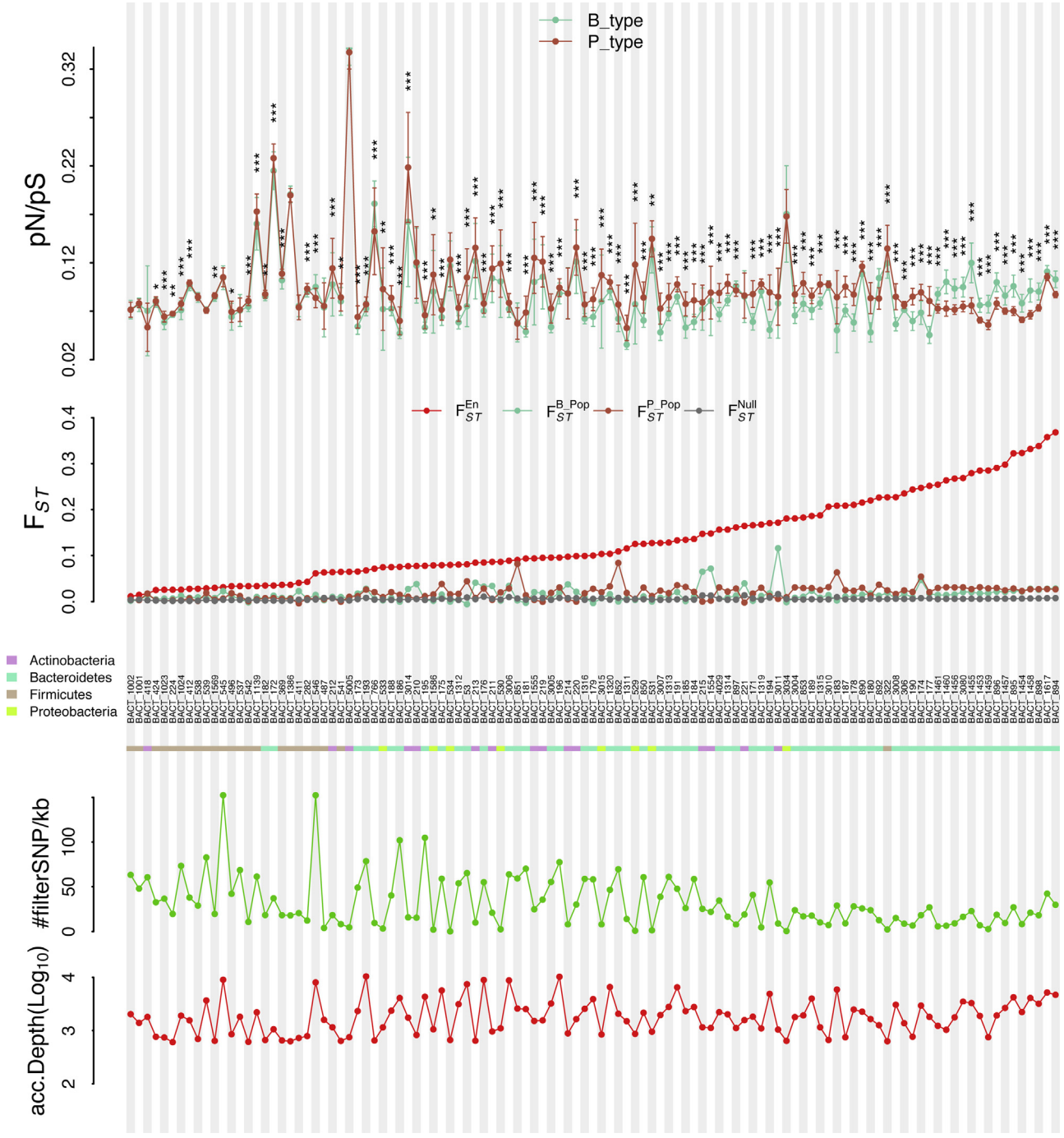


Fig. 3. Genetic variations of the 111 most dominant bacterial strains and distinction between the two enterotypes.
The genomic variation statistics are based on the 111 prevalent gut microbial strains from all 192 individuals, and the accumulated (over all individuals) base-pair depth and filtered SNP density are presented in the bottom two panels of the figure. The bacterial strain IDs on the x-axis were extracted from the HMP reference genome database and are ordered by the genomic F_{ST} values of each strain between the two enterotypes (F_{ST}^{En}). For comparison, the genomic F_{ST} values for the three populations within the B_type (F_{ST}^{Bpop}), P_type (F_{ST}^{Ppop}), and from a 5% cutoff of 1000 permutations (F_{ST}^{null}) are also presented in the F_{ST} panel. The mean and standard deviation of the genomic *pN*/*pS* ratio for each strain for both enterotypes were also calculated and tested by the Mann–Whitney U test. In the figures, '*' represents $P < 0.05$, '**' represents $P < 0.01$, and '***' represents $P < 0.001$ after BH correction.

Firmicutes were generally low (<0.1), while the highest F_{ST} values were mostly obtained from *Bacteroidetes*, which indicated that the two enterotypes exhibited significant genetic differences in many bacterial taxa across multiple bacterial phyla. Based on the *pN*/*pS* ratio, most bacterial genomes (i.e. 93 out of 111 bacterial reference genomes) were significantly different between the B_type and P_type. The majority of genomes with no significantly different *pN*/*pS* ratio belonged to *Firmicutes* and *Actinobacteria.* In contrast, the *pN*/*pS* ratio differences of the *Bacteroidetes* genomes were statistically highly significant (Fig. 3). Interestingly, the top 15 highest F_{ST} values between the two enterotypes were all present in different strains of *Bacteroides*, in which the *pN*/*pS* ratios obtained for the B_type were all higher than those obtained for the P_type. Notably, the 15 strains all belonged to the genus *Prevotella*. More detailed comparison of the *pN*/*pS* ratio calculated from each bacterial genome per individual is shown as a heatmap, but it again

closely matches the clustering of the two enterotypes (Fig. S9A). Furthermore, the PCA plot based on all filtered SNPs from the core bacterial genomes of all individuals and the NJ tree based on pairwise divergence distances of all filtered SNPs are presented in Figs. S9B and S9C, respectively. Collectively, the F_{ST} curve, *pN/pS* ratio heatmap, PCA plot, and NJ tree [12] revealed that there were significant genetic differentiations within bacterial genomes that differentiated the two enterotypes, with the analysis of the *pN/pS* ratios indicating that these significant genetic differences between the two enterotypes might have resulted from selection of certain environmental pressures.

To reveal the selection pressures shaping the genetic diversity in gut microbiota, the most differentiated genes between enterotypes were focused on. The non-redundant gene set (including 81,579 genes) was constructed from all the genes (133,431 genes) of the 111 reference genomes. By ranking, the average F_{ST} values of each gene from the non-redundant genes, the *pN/pS* ratios of the top 40 genes (0.05% of the total number of genes) with F_{ST} values larger than 0.706 are shown in Fig. 4. Most of these genes (i.e. 27 of the top 40 genes) belonged to *Bacteroides* genomes, 11 genes were from *Prevotella* genomes, only one gene (id_12781: 30S ribosomal protein S10) corresponded to a *Candidatus* genome, and one gene to a *Parabacteroides* genome. Interestingly, in all these 40 top genes, there were distinct genetic patterns between the B_type and P_type, and the genes identified from *Bacteroides* genomes showed very low *pN/pS* ratios in the individuals of the B_type, whereas the genes from the other bacterial genomes, especially *Prevotella*, showed very low *pN/pS* ratios in individuals of the P_type. On the contrary, some genes from *Bacteroides* genomes, and other genes from *Prevotella* genomes, showed *pN/pS* ratios larger than 1 in the individuals of the P_type and B_type, respectively (Fig. 4). Notably, a *pN/pS* ratio close to 0 is indicative of purifying selection, while a *pN/pS* ratio larger than 1 may indicate positive selection. Therefore, this implies that genes from different bacterial genomes are differentially selected in corresponding enterotypes. For instance, genes from *Bacteroides* (indicated by a light green bar on the right-hand side of Fig. 4) showed an overall *pN/pS* ratio close to 0 in the B_type, while a similar pattern was observed for genes from *Prevotella* (light red bar) in the P_type. Interestingly, based on the KEGG pathway analysis, these top 40 genes were mostly involved in amino acid metabolism (62.5% of the top 40 genes) and carbohydrate metabolism (17.5% of the top 40 genes), which was indicative that food, or diet, might play a role in shaping the observed metagenomic differentiation.

To elucidate further how the metagenomic differentiation between enterotypes was functionally related to diet preference, the protein products of the above highly differentiated genes were studied. For example, seven non-synonymous (NS) SNPs existed in the gene id_269210 (endo-1, 4-beta-mannosidase, *Prevotella bryantii*), which were correlated with four amino acid changes in the protein sequence (Fig. 5A). The haplotype network comprising these seven NS SNPs (Fig. 5B) confirmed the significance of the different haplotypes within the B_type and P_type, which were correlated to H_B: $G_{1063}G_{1065}C_{1073}T_{1074}C_{1076}A_{1077}A_{1079}$, and H_P: $A_{1063}T_{1065}G_{1073}C_{1074}A_{1076}C_{1077}C_{1079}$. By blasting to the Protein Data Bank (PDB) database, the closest homologous protein was β-mannosidase from the glycoside hydrolase family 5 (GH5, PDB id: 1UUQ_A) (Fig. 5C and D). The Glu^{330} within β-mannosidase acts as a catalytic nucleophile in the enzyme, which is spatially closely located to the four amino acid changes in the protein [5]. Therefore, the four amino acids with uncharged side chains ($G^{358}A^{361}A^{362}T^{363}$) in the enzyme more frequently found in the B_type should result in a different enzyme activity compared to the amino acids with electrically charged or polar side chains ($S^{358}G^{361}D^{362}N^{363}$) found in the P_type. More interestingly, it has been reported that β-mannosidase is an exo-acting glycoside hydrolase that plays a role in breaking down mannose-containing polysaccharides as carbon and energy sources, which are widely presented in the plant cell wall [5]. Previous studies have demonstrated that the *Prevotella* enterotype (or P_type in this study) was associated with long-term consumption of diets rich in plant polysaccharides [9,39]. The potentially functional differentiations of enzymes between the B_type and the P_type are further detailed in Fig. S10A–S10D for gene id_370523 (arabinose isomerase, *Bacteroides*), which is also involved in carbohydrate metabolism, and in Fig. S11A–S11D for gene id_350747 (leucyltransferase, *Prevotella ruminicola* 23), which is involved in amino acid metabolism. The above genes were randomly selected, which suggests that many more specific genes may exhibit significant differences between the two enterotypes.

Association between the host human genome and two dominant genera

The functional enrichment and genetic diversity analysis showed that the abundances of *Bacteroides* and *Prevotella* were strongly associated with the substrates of amino acids and carbohydrate. Next, the whole human genome was genotyped in order to explore [29] the potential association of host genetic effects with the two enterotypes. For *Bacteroides*, no SNPs were found with *P* values lower than 6×10^{-8} after Bonferroni adjustments (Fig. S12). The SNP with the lowest *P* value (5.937×10^{-07}) in association with *Bacteroides* abundance was rs730647 (chr22: 28250810), which is located in the *PITPNB* (phosphatidylinositol transfer protein) gene. However, as for *Prevotella*, one significant SNP rs878394 (chr1: 219073958) was found with a *P* value of 5.293e-08, which was not located within a gene but appeared adjacent to the *LYPLAL1* (lysophospholipase-like 1) gene reported to be associated with body fat distribution in Chinese [37] and Japanese [11] populations, as well as waist-hip ratio and insulin sensitivity in the Danish population [2] (Fig. 6A). Additionally, the box plot of abundance distributions of *Prevotella* in each genotype for rs878394 (Fig. 6B) demonstrated that the CC genotype of the human host corresponded to the highest *Prevotella* abundance, while the TT genotype corresponded to the lowest abundance. Such an association was also shown within each enterotype (Fig. 6C and D), although it was not significant in the B_type due to the low abundance of *Prevotella* in the data. This result suggested that the genetics of the human host might also play some role or roles in the formation of enterotypes.

Discussion

The human gut microbiome is a very complex ecosystem that might be influenced by many factors, such as diet, host health status, age, gender, height/weight, geographic environment, medication and host genetic structure [23]. Recent studies on the human microbiome have suggested that individuals can be classified into three distinct enterotypes representing a network of three distinct microbial community types with each one dominated by a particular genus [1]. While consensus largely exists that these enterotypes are the result of long-term dietary behavior [39], the concept itself has also been challenged, since other studies have proposed the existence of microbial community gradients rather than distinct enterotypes [17]. Hence, the possibility of a microbiome-based classification of human individuals is still subject to debate. In this current study, the data confirmed the intrinsic existence of the P and B enterotypes, with KZKs and UIGs (Muslim populations) dominated by the P_type and the HANs dominated by the B_type. This was consistent with a previous study showing that the P_type was mainly enriched in Buddhists and Muslims, while people with other

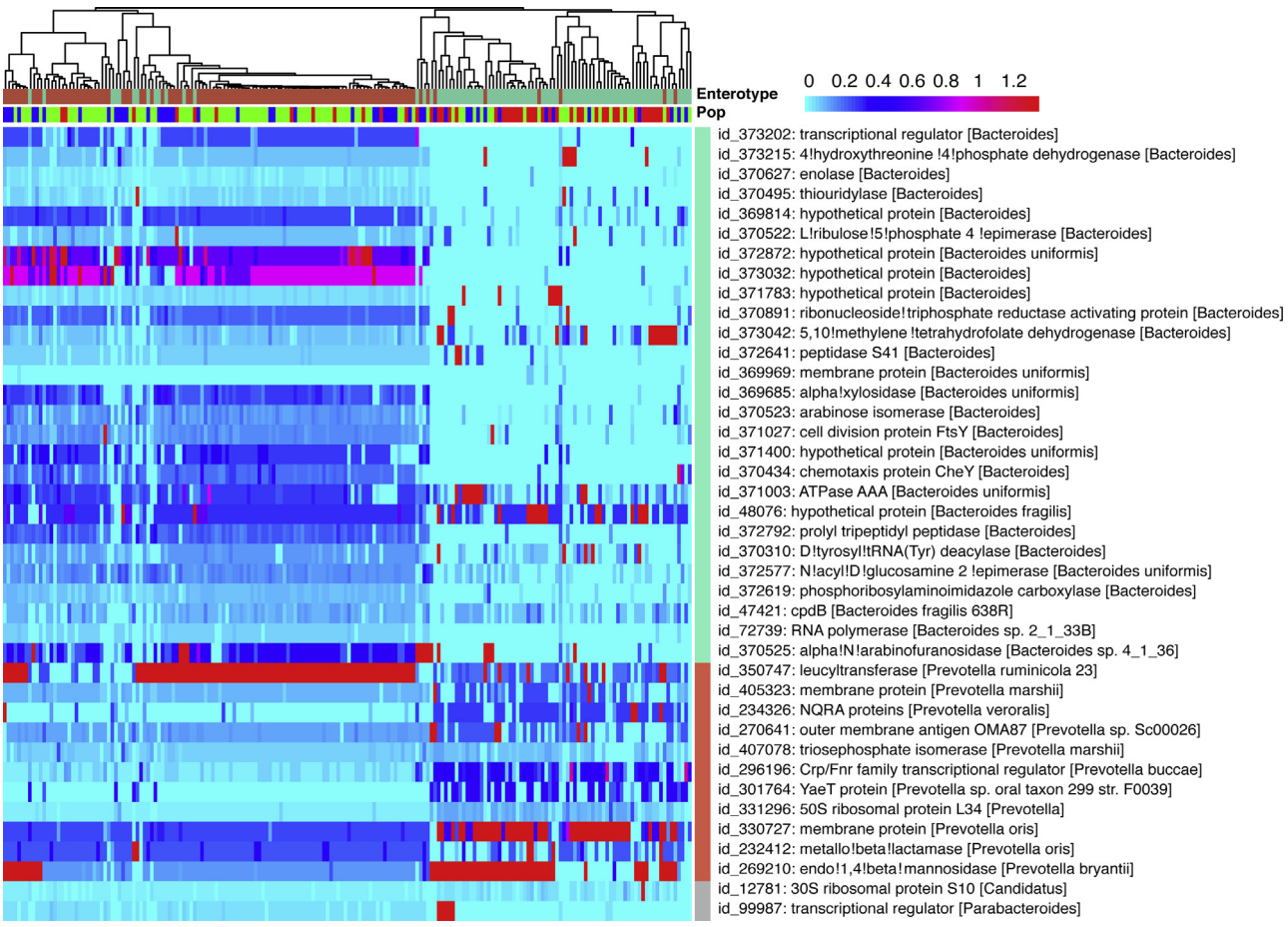


Fig. 4. Genetic differentiations between the two enterotypes on the basis of highly differentiated bacterial genes.
Heatmap plot of *pN*/*pS* ratios for 40 genes with the highest F_{ST} values between the two enterotypes. The *pN*/*pS* ratio is indicated by color according to the scale bar beside the heatmap, and the 40 genes were extracted by ranking the values of all non-redundant genes from the 111 prevalent gut microbial strains. Each column of the heatmap represents one individual, and all 192 individuals were hierarchically clustered according to the composition of the *pN*/*pS* ratios from the 40 genes. Each row represents one gene labeled and colored by the gene id (from the non-redundant gene set) and the corresponding protein name/taxon by blasting to the NCBI database.

religions were dominated by the B_type [24]. Therefore, although there was no precise information of food/nutrient uptake for the volunteers that took part in the study, the results still illustrated the link between diet and enterotype based on the knowledge of the diet difference between Muslims and Han Chinese.

The study indicated that both enterotypes were largely different in their taxonomic, as well as genetic, composition. This was supported by clear differences in functional gene enrichments, as well as by consistent polymorphisms among shared genes. Corresponding functional gene enrichment, as well as bacterial members within these enterotypes, were apparently subjected to distinct selective pressures, as indicated by different *pN*/*pS* ratios, with resulting amino acid changes in many enzymes involved in carbohydrate and amino acid metabolic transport. The genetic diversity patterns observed between the two enterotypes were probably sensitive to other confounding factors, such as age and different environments, and thus could be detected in the data because such factors were largely controlled by the study. In addition, the P_type is relatively rare, since it has a prevalence of less than 20% in the majority of populations, including Europeans, North Americans and East Asians, although metagenomic studies on other populations, such as Central Asians, are rarely conducted. Although the differentiation of different populations within enterotypes was obviously lower than the differences between enterotypes, several significant differences were still observed in the taxonomic compositions and the functional gene catalogs. Furthermore, the human genome-wide association study also showed that the abundance of the gut microbiota was highly correlated with the genotype of a specific locus within the human genome, although the responsible biological mechanism will need further evaluation in future studies.

Conclusion

In summary, considerable differentiations were identified between two enterotypes that were not only reflected in the taxonomic compositions but also in the composition of functional genes and genomic diversity patterns within microbial genomes. Furthermore, a link was established between the host genetic structure and enterotypes, which highlighted the importance of further research into the influence of host genetics on the gut microbiota and the interaction between the host genome and microbiome.

Data linking

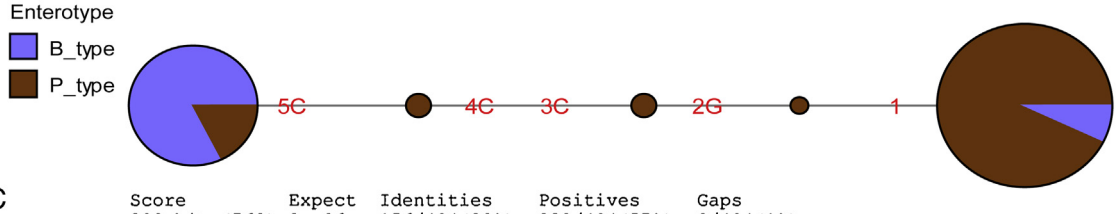
The 16S rRNA gene and metagenome sequencing data in this paper have been deposited in the National Omics Data Encyclopedia (NODE; http://www.biosino.org/node/index) under accession numbers NODEP00000052 and NODEP00000053, respectively.

A

id_269210: endo-1,4-beta-mannosidase [Prevotella bryantii]

Position	Allele	AA_pos	Pos_in_1UUQ	Amino Acid	f_B_type	f_P_type
1063	G/A	355	358	G/S	0.873/0.127	0.135/0.865
1065	G/T					
1073	C/G	358	361	A/G	0.840/0.160	0.121/0.879
1074	T/C					
1076	A/C	359	362	A/D/E	0.824/0.157/0.020	0.101/0.899/0
1077	A/C					
1079	A/C	360	363	T/N	0.824/0.176	0.083/0.917

B



C

```
Score         Expect  Identities    Positives     Gaps
300 bits(768) 2e-96   156/404(39%)  232/404(57%)  8/404(1%)

     id_269210 24  QIISCHDGKFWRGNQPYYYVGTNFWYGAILGSEGQGGNRARLKRELNHMQRLGINNLRIL 83
                     +  + G F    +PY   G N WY A LG+  + G+R RL +EL++++ +G+NNLR+L
1UUQ_A         23  HFVRVNGGHFELQGKPYVITGVNMWYAAYLGAPNEVGDRDRLAKELDNLKAIGVNNLRVL 82

     id_269210 84  VGSDGLRGVKTKVEPTLQVAPGVYNDTIMAGLDYLLAEMGKRNMVAVLYLNNSWEWSGGY 143
                     S+    + + V+P +    G Y++T++ GLDYLL E+ KR+M  VLY NN W+WSGG
1UUQ_A         83  AVSEKSE-INSAVKPAVTNGFGNYDETLLQGLDYLLVELAKRDMTVVLYFNNFWQWSGGM 141

     id_269210 144 GFYLEHAGEGKAPRPN-EDGYPAFMNFVSKFATNKKAHQLFYDYVKFILSRTNRYTGKRY 202
                     Y+          PN  + + AFM   + F  ++KA Q +   ++ I++R N   GK Y
1UUQ_A         142 TQYMAWIEGEPVQDPNVTNEWEAFMAKSASFYRSEKAQQEYRKTLEKIITRVNSINGKAY 201

     id_269210 203 IDDPTIMSWQIGNEPRAFSKKVLPAFAQ----WLSEASALIRKLDPNHLISIGSEGSWGC 258
                   +DD TIMSWQ+ NEPR  + +      Q    W+  A+A I+ LD +HL+S GSEG  G
1UUQ_A         202 VDDATIMSWQLANEPRPGNSQTTAEEKQIYIDWVHAAAAYIKTLDAHHLVSSGSEGEMGS 261

     id_269210 259 ENDEKVYEQICADKNIDYCNVHLWPYNWSWAKADSLIENLPRAKKNTKEYIDHHLIICKR 318
                    ND +V+    A  +IDY   H+W  NWSW       E  P A +  + Y+  H+ + K+
1UUQ_A         262 VNDMQVFIDAHATPDIDYLTYHMWIRNWSWFDKTKPAETWPSAWEKAQNYMRAHIDVAKQ 321

B type                                          G  AAT
P type                                          S  GDN
     id_269210 319 LQKPLVMEEYGYPRDGFKFDLEASTHARDSYYQYVFGLVAENATKGGLFAGCNFWGWGGE 378
                   L KPLV+EE+G  RD   + ++++T  RD+Y++ VF L+   +  +G   AG N W W G
1UUQ_A         322 LNKPLVLEEFGLDRDMGSYAMDSTTEYRDNYFRGVFELMLASLEQGEPSAGYNIWAWNGY 381

     id_269210 379 AKPKHEQ--WQVGDDYTGDPAQEAQGLNSVFVTDKSTLAIVKKY 420
                    +       WQ GDD+ GDP QE QG+  VF TD ST+AI+K++
1UUQ_A         382 GRTTRANYWWQEGDDFMGDPPQEEQGMYGVFDTDTSTIAIMKEF 425
```

D

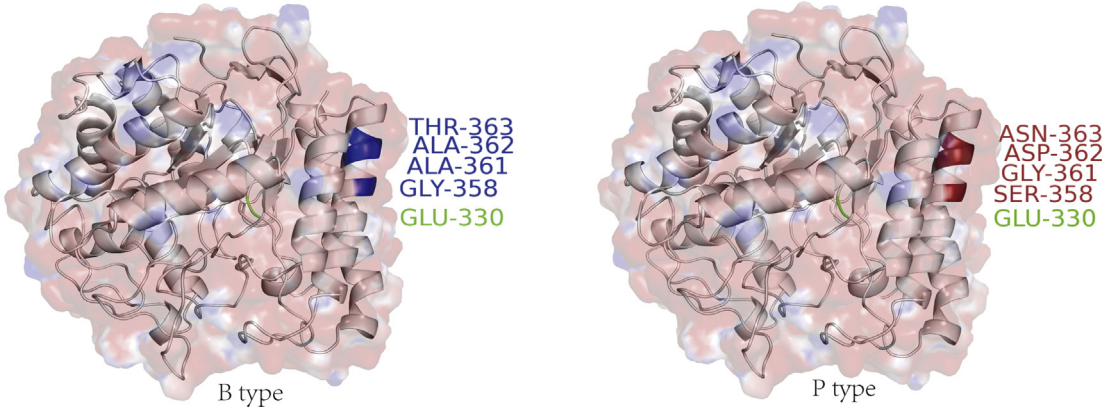


Fig. 5. Non-synonymous SNPs and associated amino acid changes for one selected gene/protein between the two enterotypes.
(a) Positions and types of seven non-synonymous SNPs and corresponding four amino acid changes in id_269210 (endo-1,4-beta-mannosidase, *Prevotella bryantii*). (b) Haplotype network based on the seven non-synonymous SNPs and all 192 individuals. (c) Alignment of the query protein (id_269210) and the target protein (PDB id: 1UUQ_A), in which the four significant amino acid changes between the two enterotypes are highlighted. (d) Three-dimensional structure of the protein (temple structure from PDB ID: 1UUQ_A) with the four significant amino acid changes highlighted. Additionally, the closely located amino acid GLU^{330} (marked in green) is shown, since it acts as a catalytic nucleophile in the enzyme.

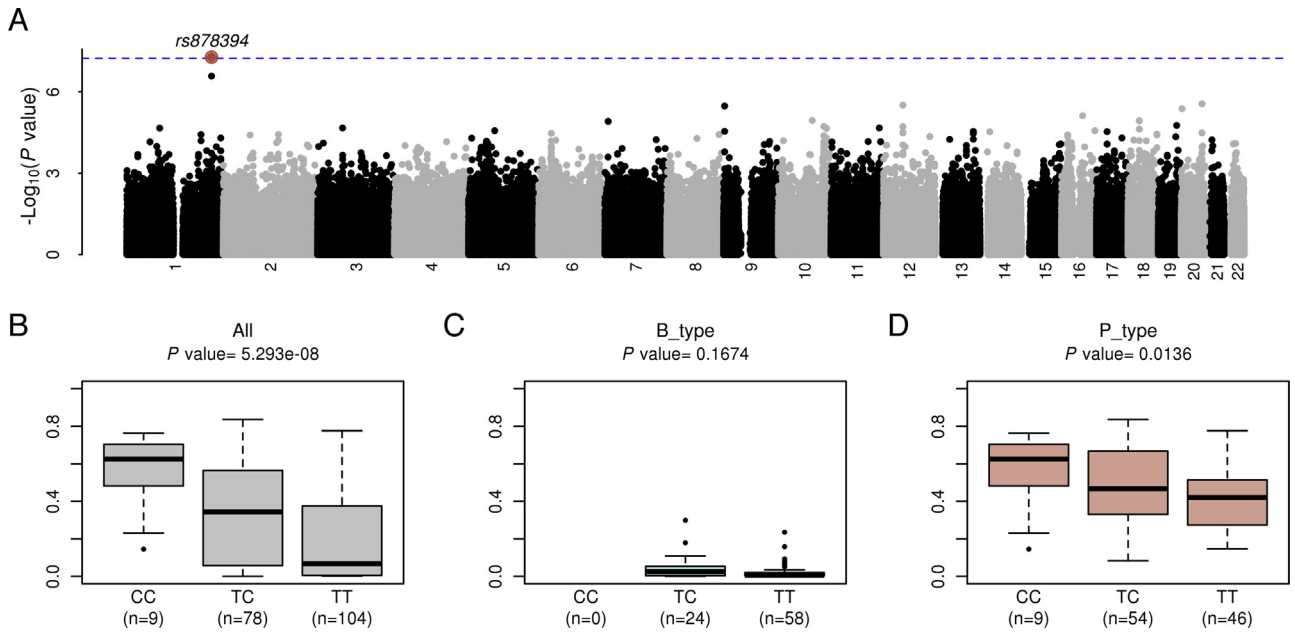


Fig. 6. Association between the human autosomal SNPs and the relative abundance of *Prevotella* in the gut microbiota.
(a) Manhattan plot for the log-transformed *P* values of all human autosomal SNPs tested for their association with the relative abundance of *Prevotella*. The blue line represents the *P* value cutoff (6×10^{-8}) for the associations. (b–d) Relative abundance of *Prevotella* corresponding to three genotypes of rs878394 for all 192 individuals (b), all individuals within the B_type (c), and all individuals within the P_type (d).

Funding

S.X. acknowledges financial support from the Strategic Priority Research Program (XDB13040100) and Key Research Program of Frontier Sciences (QYZDJ-SSW-SYS009) of the Chinese Academy of Sciences (CAS), the National Natural Science Foundation of China (NSFC) grants (91331204, 91731303, 31771388, and 31711530221), the National Science Fund for Distinguished Young Scholars (31525014), the National Key Research and Development Program (2016YFC0906403), and the Program of Shanghai Academic Research Leader (16XD1404700), whereas J.L. acknowledges financial support from NSFC grants (31370505 and 31670495). Y.L. acknowledges support from an NSFC grant (31501011) and a Science and Technology Commission of Shanghai Municipality grant (STCSM) (14YF1406800). H.L. acknowledges support from an STCSM grant (16YF1413900), and Y.G. acknowledges support from an NSFC grant (31260263). S.X. is a Max-Planck Independent Research Group Leader, and also gratefully acknowledges the support of the National Program for Top-notch Young Innovative Talents of the "*Wanren Jihua*" Project. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.syapm.2017.09.006.

References

[1] Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D.R., Fernandes, G.R., Tap, J., Bruls, T., Batto, J.-M., Bertalan, M., Borruel, N., Casellas, F., Fernandez, L., Gautier, L., Hansen, T., Hattori, M., Hayashi, T., Kleerebezem, M., Kurokawa, K., Leclerc, M., Levenez, F., Manichanh, C., Nielsen, H.B., Nielsen, T., Pons, N., Poulain, J., Qin, J., Sicheritz-Ponten, T., Tims, S., Torrents, D., Ugarte, E., Zoetendal, E.G., Wang, J., Guarner, F., Pedersen, O., de Vos, W.M., Brunak, S., Doré, J., Antolín, M., Artiguenave, F., Blottiere, H.M., Almeida, M., Brechot, C., Cara, C., Chervaux, C., Cultrone, A., Delorme, C., Denariaz, G., Dervyn, R., Foerstner, K.U., Friss, C., van de Guchte, M., Guedon, E., Haimet, F., Huber, W., van Hylckama-Vlieg, J., Jamet, A., Juste, C., Kaci, G., Knol, J., Lakhdari, O., Layec, S., Le Roux, K., Maguin, E., Mérieux, A., Melo Minardi, R., M'rini, C., Muller, J., Oozeer, R., Parkhill, J., Renault, P., Rescigno, M., Sanchez, N., Sunagawa, S., Torrejon, A., Turner, K., Vandemeulebrouck, G., Varela, E., Winogradsky, Y., Zeller, G., Weissenbach, J., Ehrlich, S.D., Bork, P. (2011) Enterotypes of the human gut microbiome. Nature 473 (7346), 174–180, http://dx.doi.org/10.1038/nature09944.

[2] Burgdorf, K.S., Gjesing, A.P., Grarup, N., Justesen, J.M., Sandholt, C.H., Witte, D.R., Jørgensen, T., Madsbad, S., Hansen, T., Pedersen, O. (2012) Association studies of novel obesity-related gene variants with quantitative metabolic phenotypes in a population-based sample of 6,039 Danish individuals. Diabetologia 55 (1), 105–113, http://dx.doi.org/10.1007/s00125-011-2320-4.

[3] Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L. (2009) BLAST+: architecture and applications. BMC Bioinform. 10, 421, http://dx.doi.org/10.1186/1471-2105-10-421.

[4] Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., Huttley, G.A., Kelley, S.T., Knights, D., Koenig, J.E., Ley, R.E., Lozupone, C.A., McDonald, D., Muegge, B.D., Pirrung, M., Reeder, J., Sevinsky, J.R., Turnbaugh, P.J., Walters, W.A., Widmann, J., Yatsunenko, T., Zaneveld, J., Knight, R. (2010) QIIME allows analysis of high-throughput community sequencing data. Nat. Methods 7 (5), 335–336, http://dx.doi.org/10.1038/nmeth.f.303.

[5] Dias, F.M.V., Vincent, F., Pell, G., Prates, J.A.M., Centeno, M.S.J., Tailford, L.E., Ferreira, L.M.A., Fontes, C.M.G.A., Davies, G.J., Gilbert, H.J. (2004) Insights into the molecular determinants of substrate specificity in glycoside hydrolase family 5 revealed by the crystal structure and kinetics of *Cellvibrio mixtus* mannosidase 5A. J. Biol. Chem. 279 (24), 25517–25526, http://dx.doi.org/10.1074/jbc.M401647200.

[6] Dusko Ehrlich, S., MetaHIT consortium, (2010) Metagenomics of the intestinal microbiota: potential applications. Gastroenterol. Clin. Biol. 34 (Suppl. 1), S23–S28, http://dx.doi.org/10.1016/S0399-8320(10)70017-8.

[7] Excoffier, L., Lischer, H.E.L. (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol. Ecol. Resour. 10 (3), 564–567, http://dx.doi.org/10.1111/j.1755-0998.2010.02847.x.

[8] Falony, G., Joossens, M., Vieira-Silva, S., Wang, J., Darzi, Y., Faust, K., Kurilshikov, A., Bonder, M.J., Valles-Colomer, M., Vandeputte, D., Tito, R.Y., Chaffron, S., Rymenans, L., Verspecht, C., De Sutter, L., Tigchelaar, E.F., Eeckhaudt, L., Fu, J., Henckaerts, L., Zhernakova, A., Wijmenga, C., Raes, J. (2016) Population-level analysis of gut microbiome variation. Science 352 (6285), 560–564, http://dx.doi.org/10.1126/science.aad3503.

[9] De Filippo, C., Cavalieri, D., Di Paola, M., Ramazzotti, M., Poullet, J.B., Massart, S., Collini, S., Pieraccini, G., Lionetti, P. (2010) Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. Proc. Natl. Acad. Sci. U. S. A. 107 (33), 14691–14696, http://dx.doi.org/10.1073/pnas.1005963107.

[10] Goodrich, J.K., Waters, J.L., Poole, A.C., Sutter, J.L., Koren, O., Blekhman, R., Beaumont, M., Van Treuren, W., Knight, R., Bell, J.T., Spector, T.D., Clark, A.G., Ley,

R.E. (2014) Human genetics shape the gut microbiome. Cell 159 (4), 789–799, http://dx.doi.org/10.1016/j.cell.2014.09.053.
[11] Hotta, K., Kitamoto, A., Kitamoto, T., Mizusawa, S., Teranishi, H., So, R., Matsuo, T., Nakata, Y., Hyogo, H., Ochi, H., Nakamura, T., Kamohara, S., Miyatake, N., Kotani, K., Itoh, N., Mineo, I., Wada, J., Yoneda, M., Nakajima, A., Funahashi, T., Miyazaki, S., Tokunaga, K., Masuzaki, H., Ueno, T., Chayama, K., Hamaguchi, K., Yamada, K., Hanafusa, T., Oikawa, S., Sakata, T., Tanaka, K., Matsuzawa, Y., Nakao, K., Sekine, A. (2013) Replication study of 15 recently published loci for body fat distribution in the Japanese population. J. Atheroscler. Thromb. 20 (4), 336–350 https//doi.org/10.5551/jat.14589.
[12] Hu, Y., Yang, X., Qin, J., Lu, N., Cheng, G., Wu, N., Pan, Y., Li, J., Zhu, L., Wang, X., Meng, Z., Zhao, F., Liu, D., Ma, J., Qin, N., Xiang, C., Xiao, Y., Li, L., Yang, H., Wang, J.J., Yang, R., Gao, G.F., Wang, J.J., Zhu, B. (2013) Metagenome-wide analysis of antibiotic resistance genes in a large cohort of human gut microbiota. Nat. Commun. 4, 2151, http://dx.doi.org/10.1038/ncomms3151.
[13] Huson, D.H., Auch, A.F., Qi, J., Schuster, S.C. (2007) MEGAN analysis of metagenomic data. Genome Res. 17 (3), 377–386, http://dx.doi.org/10.1101/gr.5969107.
[14] Jeffery, I.B., Claesson, M.J., O'Toole, P.W., Shanahan, F. (2012) Categorization of the gut microbiota: enterotypes or gradients? Nat. Rev. Microbiol. 10 (9), 591–592, http://dx.doi.org/10.1038/nrmicro2859.
[15] El Kaoutari, A., Armougom, F., Gordon, J.I., Raoult, D., Henrissat, B. (2013) The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. Nat. Rev. Microbiol. 11 (7), 497–504, http://dx.doi.org/10.1038/nrmicro3050.
[16] Kent, W.J. (2002) BLAT—the BLAST-like alignment tool. Genome Res. 12 (4), 656–664, http://dx.doi.org/10.1101/gr.229202.
[17] Knights, D., Ward, T.L., McKinlay, C.E., Miller, H., Gonzalez, A., McDonald, D., Knight, R. (2014) Rethinking enterotypes. Cell Host Microbe 16 (4), 433–437, http://dx.doi.org/10.1016/j.chom.2014.09.013.
[18] Li, H. (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27 (21), 2987–2993, http://dx.doi.org/10.1093/bioinformatics/btr509.
[19] Li, H., Durbin, R. (2010) Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics 26 (5), 589–595, http://dx.doi.org/10.1093/bioinformatics/btp698.
[20] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup, (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25 (16), 2078–2079, http://dx.doi.org/10.1093/bioinformatics/btp352.
[21] Li, R., Yu, C., Li, Y., Lam, T.W., Yiu, S.M., Kristiansen, K., Wang, J. (2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25 (15), 1966–1967, http://dx.doi.org/10.1093/bioinformatics/btp336.
[22] Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., Tang, J., Wu, G., Zhang, H., Shi, Y., Liu, Y., Yu, C., Wang, B., Lu, Y., Han, C., Cheung, D.W., Yiu, S.-M., Peng, S., Xiaoqian, Z., Liu, G., Liao, X., Li, Y., Yang, H., Wang, J., Lam, T., Wang, J. (2012) SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. Gigascience 1 (1), 18, http://dx.doi.org/10.1186/2047-217X-1-18.
[23] Moeller, A.H., Ochman, H. (2013) Factors that drive variation among gut microbial communities. Gut Microbes, 403–408, http://dx.doi.org/10.4161/gmic.26039.
[24] Nakayama, J., Watanabe, K., Jiang, J., Matsuda, K., Chao, S.-H., Haryono, P., La-Ongkham, O., Sarwoko, M.-A., Sujaya, I.N., Zhao, L., Chen, K.-T., Chen, Y.-P., Chiu, H.-H., Hidaka, T., Huang, N.-X., Kiyohara, C., Kurakawa, T., Sakamoto, N., Sonomoto, K., Tashiro, K., Tsuji, H., Chen, M.-J., Leelavatcharamas, V., Liao, C.-C., Nitisinprasert, S., Rahayu, E.S., Ren, F.-Z., Tsai, Y.-C., Lee, Y.-K. (2015) Diversity in gut bacterial community of school-age children in Asia. Sci. Rep. 5, 8397, http://dx.doi.org/10.1038/srep08397.
[25] Nava, G.M., Carbonero, F., Ou, J., Benefiel, A.C., O'Keefe, S.J., Gaskins, H.R. (2012) Hydrogenotrophic microbiota distinguish native Africans from African and European Americans. Environ. Microbiol. Rep. 4 (3), 307–315, http://dx.doi.org/10.1111/j.1758-2229.2012.00334.x.
[26] Patterson, N., Price, A.L., Reich, D. (2006) Population structure and eigenanalysis. PLoS Genet. 2 (12), 2074–2093, http://dx.doi.org/10.1371/journal.pgen.0020190.
[27] Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., Schloss, J.A., Bonazzi, V., McEwen, J.E., Wetterstrand, K.A., Deal, C., Baker, C.C., Di Francesco, V., Howcroft, T.K., Karp, R.W., Lunsford, R.D., Wellington, C.R., Belachew, T., Wright, M., Giblin, C., David, H., Mills, M., Salomon, R., Mullins, C., Akolkar, B., Begg, L., Davis, C., Grandison, L., Humble, M., Khalsa, J., Little, A.R., Peavy, H., Pontzer, C., Portnoy, M., Sayre, M.H., Starke-Reed, P., Zakhari, S., Read, J., Watson, B., Guyer, M. (2009) The NIH human microbiome project. Genome Res. 19 (12), 2317–2323, http://dx.doi.org/10.1101/gr.096651.109.
[28] Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., Reich, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38 (8), 904–909, http://dx.doi.org/10.1038/ng1847.
[29] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., Sham, P.C. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81 (3), 559–575, http://dx.doi.org/10.1086/519795.
[30] Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D.R., Li, J., Xu, J., Li, S.S.S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J.-M., Hansen, T., Le Paslier, D., Linneberg, A., Nielsen, H.B., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Li, S.S.S., Jian, M., Zhou, Y., Li, Y., Zhang, X., Li, S.S.S., Qin, N., Yang, H., Wang, J.J., Brunak, S., Doré, J., Guarner, F., Kristiansen, K., Pedersen, O., Parkhill, J., Weissenbach, J., MetaHIT Consortium, Bork, P., Ehrlich, S.D., Wang, J.J. (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464 (7285), 59–65, http://dx.doi.org/10.1038/nature08821.
[31] Qin, J., Li, Y., Cai, Z., Li, S.S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., Peng, Y., Zhang, D., Jie, Z., Wu, W., Qin, Y., Xue, W., Li, J., Han, L., Lu, D., Wu, P., Dai, Y., Sun, X., Li, Z., Tang, A., Zhong, S., Li, X., Chen, W., Xu, R., Wang, M., Feng, Q., Gong, M., Yu, J., Zhang, Y., Zhang, M., Hansen, T., Sanchez, G., Raes, J., Falony, G., Okuda, S., Almeida, M., LeChatelier, E., Renault, P., Pons, N., Batto, J.-M., Zhang, Z., Chen, H., Yang, R., Zheng, W., Li, S.S., Yang, H., Wang, J.J., Ehrlich, S.D., Nielsen, R., Pedersen, O., Kristiansen, K., Wang, J.J. (2012) A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature 490 (7418), 55–60, http://dx.doi.org/10.1038/nature11450.
[32] Rieder, R., Wisniewski, P.J., Alderman, B.L., Campbell, S.C. (2017) Microbes and mental health: a review. Brain Behav. Immun., http://dx.doi.org/10.1016/j.bbi.2017.01.016.
[33] Schloissnig, S., Arumugam, M., Sunagawa, S., Mitreva, M., Tap, J., Zhu, A., Waller, A., Mende, D.R., Kultima, J.R., Martin, J., Kota, K., Sunyaev, S.R., Weinstock, G.M., Bork, P. (2013) Genomic variation landscape of the human gut microbiome. Nature 493 (7430), 45–50, http://dx.doi.org/10.1038/nature11711.
[34] Shade, A., Gilbert, J.A. (2015) Temporal patterns of rarity provide a more complete view of microbial diversity. Trends Microbiol. 23 (6), 335–340, http://dx.doi.org/10.1016/j.tim.2015.01.007.
[35] Sundquist, A., Bigdeli, S., Jalili, R., Druzin, M.L., Waller, S., Pullen, K.M., El-Sayed, Y.Y., Taslimi, M.M., Batzoglou, S., Ronaghi, M., Amann, R., Ludwig, W., Schleifer, K., Rappé, M., Giovannoni, S., Tringe, S., Rubin, E., Anderson, B., Dawson, J., Jones, D., Wilson, K., Verhelst, R., Verstraelen, H., Claeys, G., Verschraegen, G., Delanghe, J., Van Simaey, L., De Ganck, C., Temmerman, M., Vaneechoutte, M., Goldenberg, R., Hauth, J., Andrews, W., Gravett, M., Novy, M., Rosenfeld, R., Reddy, A., Jacob, T., Turner, M., Sbarra, A., Selvaraj, R., Cetrulo, C., Feingold, M., Newton, E., Thomas, G., McGregor, J., Lawellin, D., Franco-Buff, A., Todd, J., Makowski, E., Andrews, W., Hauth, J., Goldenberg, R., Gomez, R., Romero, R., Cassell, G., Fortunato, S., Menon, R., Swan, K., Menon, R., Fortunato, S., Menon, R., Lombarda, S., Margulies, M., Egholm, M., Altman, W., Attiya, S., Bader, J., Bemben, L., Berka, J., Braverman, M., Chen, Y., Chen, Z., Dewell, S., Du, L., Fierro, J., Gomes, X., Godwin, B., He, W., Helgesen, S., Ho, C., Irzyk, G., Jando, S., Alenquer, M., Jarvie, T., Jirage, K., Kim, J.-B., Knight, J., Lanza, J., Leamon, J., Lefkowitz, S., Lei, M., Li, J., Lohman, K., Lu, H., Makhijani, V., McDade, K., McKenna, M., Myers, E., Nickerson, E., Nobile, J., Plant, R., Puc, B., Ronan, M., Roth, G., Sarkis, G., Simons, J., Simpson, J., Srinivasan, M., Tartaro, K., Tomasz, A., Vogt, K., Volkmer, G., Wang, S., Wang, Y., Weiner, M., Yu, P., Begley, R., Rothberg, J., Rogers, Y., Venter, J., Ronaghi, M., Karamohamed, S., Petterson, B., Uhlen, M., Nyren, P., Hyman, R., Fukushima, M., Diamona, L., Kumm, J., Giudice, L., Davis, R., Sogin, M., Morrison, H., Huber, J., Welch, D., Huse, S., Neal, P., Arrieta, J., Herndl, G., Kent, W., Cole, J., Chai, B., Farris, R., Wang, Q., Kulam, S., McGarrell, D., Garrity, G., Tiedje, J., DeSantis, T., Dubosarskiy, I., Murray, S., Andersen, G., Von Wintzingerode, F., Gobel, U., Stackebrandt, E., Monstein, H., Nikpour-Badi, S., Jonasson, J., Gray, M., Sankoff, D., Cedergren, R., Carey, J., Klebanoff, M., Hauth, J., Hillier, S., Thom, E., Ernest, J., Heine, R., Nugent, R., Fischer, M., Leveno, K., Wapner, R., Varner, M., Trout, W., Moawad, A., Sibai, B., Miodovnik, M., Dombrowski, M., O'Sullivan, M., Van Dorsten, J., Langer, O., Roberts, J., Klebanoff, M., Carey, J., Hauth, J., Hillier, S., Nugent, R., Thom, E., Ernest, J., Heine, R., Wapner, R., Trout, W., Moawad, A., Miodovnik, M., Sibai, B., Van Dorsten, J., Dombrowski, M., O'Sullivan, M., Varner, M., Langer, O., McNellis, D., Roberts, J., Leveno, K., Neefs, J., Van de Peer, Y., De Rijk, P., Goris, A., De Wachter, R. (2007) Bacterial flora-typing with targeted, chip-based pyrosequencing. BMC Microbiol. 7 (1), 108, http://dx.doi.org/10.1186/1471-2180-7-108.
[36] Waldor, M.K., Tyson, G., Borenstein, E., Ochman, H., Moeller, A., Finlay, B.B., Kong, H.H., Gordon, J.I., Nelson, K.E., Dabbagh, K., Smith, H. (2015) Where next for microbiome research? PLoS Biol. 13 (1), http://dx.doi.org/10.1371/journal.pbio.1002050.
[37] Wang, T., Ma, X., Peng, D., Zhang, R., Sun, X., Chen, M., Yan, J., Wang, S., Yan, D., He, Z., Jiang, F., Bao, Y., Hu, C., Jia, W. (2016) Effects of obesity related genetic variations on visceral and subcutaneous fat distribution in a Chinese population. Sci. Rep. 6, 20691, http://dx.doi.org/10.1038/srep20691.
[38] Weir, B.S., Cockerham, C.C. (1984) Estimating F-statistics for the analysis of population structure. Evolution (NY) 38 (6), 1358, http://dx.doi.org/10.2307/2408641.
[39] Wu, G.D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.Y., Keilbaugh, S.A., Bewtra, M., Knights, D., Walters, W.A., Knight, R., Sinha, R., Gilroy, E., Gupta, K., Baldassano, R., Nessel, L., Li, H., Bushman, F.D., Lewis, J.D. (2011) Linking long-term dietary patterns with gut microbial enterotypes. Science 334 (6052), 105–108, http://dx.doi.org/10.1126/science.1208344.
[40] Yuan, S., Cohen, D.B., Ravel, J., Abdo, Z., Forney, L.J. (2012) Evaluation of methods for the extraction and purification of DNA from the human microbiome. PLoS One 7 (3), e33865, http://dx.doi.org/10.1371/journal.pone.0033865.
[41] Zhu, W., Lomsadze, A., Borodovsky, M. (2010) *Ab initio* gene identification in metagenomic sequences. Nucleic Acids Res. 38 (12), e132, http://dx.doi.org/10.1093/nar/gkq275.