Systematic and Applied Microbiology xxx (2017) xxx–xxx

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-

Please cite this article in press as: J. Li, et al., A metagenomic approach to dissect the genetic composition of enterotypes in Han Chinese and two Muslim groups, Syst. Appl. Microbiol. (2017), https://doi.org/10.1016/j.syapm.2017.09.006