Appl Microbiol Biotechnol (2017) 101:423–435
DOI 10.1007/s00253-016-7910-0

ENVIRONMENTAL BIOTECHNOLOGY

HuMiChip2 for strain level identification and functional profiling of human microbiomes

Qichao Tu[1,2,3] · Jiabao Li[4] · Zhou Shi[3] · Yanfei Chen[5] · Lu Lin[1] · Juan Li[6] · Hongling Wang[2] · Jianbo Yan[2] · Qingming Zhou[6] · Xiangzhen Li[4] · Lanjuan Li[5] · Jizhong Zhou[3] · Zhili He[3]

Received: 3 June 2016 / Revised: 18 September 2016 / Accepted: 27 September 2016 / Published online: 12 October 2016

Abstract With the massive data generated by the Human Microbiome Project, how to transform such data into useful information and knowledge remains challenging. Here, with currently available sequencing information (reference genomes and metagenomes), we have developed a comprehensive microarray, HuMiChip2, for strain-level identification and functional characterization of human microbiomes. HuMiChip2 was composed of 29,467 strain-specific probes targeting 2063 microbial strains/species and 133,924 sequence- and group-specific probes targeting 157 key functional gene families involved in various metabolic pathways and host-microbiome interaction processes. Computational evaluation of strain-specific probes suggested that they were not only specific to mock communities of sequenced microorganisms and metagenomes from different human body sites but also to non-sequenced microbial strains. Experimental evaluation of strain-specific probes using single strains/species and mock communities suggested a high specificity of these probes with their corresponding targets. Application of HuMiChip2 to human gut microbiome samples showed the patient microbiomes of alcoholic liver cirrhosis significantly ($p < 0.05$) shifted their functional structure from the healthy individuals, and the relative abundance of 21 gene families significantly ($p < 0.1$) differed between the liver cirrhosis patients and healthy individuals. At the strain level, five *Bacteroides* strains were significantly ($p < 0.1$) and more frequently detected in liver cirrhosis patients. These results suggest that the developed HuMiChip2 is a useful microbial ecological microarray for both strain-level identification and functional profiling of human microbiomes.

Keywords HuMiChip2 · Microbial ecological microarray · Strain-level identification · Functional profiling · Human microbiome

Introduction

The human microbiome plays extremely important roles in human health, disease, nutrition, and antibiotic resistance, as revealed by extensive recent studies (Cho and Blaser 2012; Kau et al. 2011; Ley 2010; Peterson et al. 2009; Qin et al. 2012, 2014; Sommer et al. 2009; Turnbaugh et al. 2009). For example, studies have shown that several common human disorders and disease such as obesity (Turnbaugh et al. 2009), type 2 diabetes (Qin et al. 2012), and liver cirrhosis (Chen et al. 2011; Qin et al. 2014) are closely related with changed

Qichao Tu and Jiabao Li contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s00253-016-7910-0) contains supplementary material, which is available to authorized users.

✉ Jizhong Zhou
jzhou@ou.edu

✉ Zhili He
zhili.he@ou.edu

[1] Department of Marine Sciences, Ocean College, Zhejiang University, Zhejiang, China

[2] Zhoushan Municipal Center for Disease Control and Prevention, Zhoushan, Zhejiang, China

[3] Institute for Environmental Genomics, University of Oklahoma, Norman, OK, USA

[4] Chinese Academy of Sciences, Chengdu Institute of Biology, Chengdu, Sichuan, China

[5] The First Affiliated Hospital, Zhejiang University, Hangzhou, Zhejiang, China

[6] College of Agriculture, Hunan Agricultural University, Changsha, Hunan, China

gut microbiomes. In healthy individuals, metagenomic analysis suggests highly varied microbial taxa but relatively stable metabolic pathways and gene content among different individuals, as revealed by the Human Microbiome Project (The HMP Consortium 2012) and several other studies (Caporaso et al. 2011; Gevers et al. 2012; Turnbaugh et al. 2009). However, current efforts are mainly focusing on the taxonomic and functional levels as well as linkages between human microbiomes and various human disorders, and shotgun metagenome sequencing and 16S ribosomal RNA (rRNA) amplicon sequencing are mainly used in human microbiome studies. The accumulation of such sequencing data challenges us to translate them into useful information and knowledge.

Microarrays are one of metagenomic tools that utilize sequence data in research, clinic diagnosis, and environmental monitoring and detection (He 2014). For example, microbial ecological microarrays have been used to analyze various microbial communities at both functional and taxonomic levels, including human microbiomes (He et al. 2012a, b; Tu et al. 2014c). Until now, several types of microbial ecological microarrays, including HuMiChip (Tu et al. 2014a), HITChip (Rajilić-Stojanović et al. 2009), and HuGChip (Tottey et al. 2013), have been specifically developed to profile human microbiomes. Among these, HuMiChip is a functional gene array targeting 139 key functional gene families involved in various metabolic pathways and can be used for functional profiling of human microbiomes from different body sites (Tu et al. 2014a). HITChip and HuGChip are two microbial ecological microarrays composed of probes targeting 16S rRNA genes and can be used for taxonomic profiling of human gut microbiomes (Rajilić-Stojanović et al. 2009; Tottey et al. 2013). Notably, owing to the innovative explorative probes included on HuGChip, this microarray can also be used to detect microorganisms without reference 16S rRNA genes. However, due to the high conservative nature of 16S rRNA genes, none of these microbial ecological microarrays are able to identify and detect microorganisms at the strain/species level. This is an especially important issue because microbial strains and species are usually directly responsible for many human disorders and disease. One such well-known example is the microbial species *Escherichia coli* that the majority of *E. coli* strains are commensal and even beneficial to the human body, while a few of them are extremely pathogenic, such as the O157:H7 series (Kaper et al. 2004).

In order to analyze complex microbial communities at high resolutions such as strain level, we previously developed a *k*-mer-based algorithm to design strain-specific probes that can be used to construct microbial identification microarrays (Tu et al. 2013). These strain-specific probes/markers can also be used to identify microorganisms at the strain/species level in shotgun metagenomes, and an application of these markers to published metagenome datasets identified a series of microbial strains associated with type 2 diabetes and human obesity (Tu et al. 2014b). With the rapid accumulation of reference genomes of microorganisms isolated from human body, especially those generated by the Human Microbiome Project (Peterson et al. 2009), this algorithm provided an opportunity to construct a microbial ecological microarray for strain-level identification of human microbiomes.

In this study, we aimed to construct a more comprehensive microbial ecological microarray-termed HuMiChip2 for both strain/species-level identification and functional profiling of human microbiomes. Strain/species-specific probes were designed from more than 2000 reference genomes for strain/species-level identification of human microbiomes. Also, compared with HuMiChip which only included 322 bacterial genomes and 31 human gut shotgun metagenomes, we targeted more reference genomes (2063) and shotgun metagenomes (2.5 Gb assembled contigs from 14 different human body sites) in HuMiChip2 for functional gene families. In addition, HuMiChip2 is expected to provide an alternative metagenomic approach to shotgun metagenome and 16S rRNA gene amplicon sequencing for human microbiome studies with the ability to detect known microorganisms at the strain/species level.

Materials and methods

Data resources, probe design, and microarray fabrication

A total of 2063 sequenced microbial genomes and ~2.5 Gb assembled shotgun metagenome sequences from 14 different body sites were retrieved from the Human Microbiome Project Data Analysis and Coordination Center (http://hmpdacc.org). The sequenced microbial genomes were used for probe design of both functional gene families and strain-specific probes. The metagenome datasets were used for probe design of functional gene families. To insure the specificity of strain-specific probes against non-human microorganisms and the human genome, another 3327 sequenced microbial genomes and the human genome sequences were downloaded from the National Center for Biotechnology Information (NCBI) ftp site for specificity checking. A full list of targeted microbial genomes and assembled metagenomes could be found in Supplementary Tables S1 and S2, respectively.

Probe design for functional gene families was carried out using the same pipeline as described previously (Tu et al. 2014a). Briefly, hidden Markov models (HMM) for each functional gene family were built by HMMER program (Eddy 1998) using curated reference sequences retrieved from the KEGG database (Kanehisa et al. 2016). Protein and coding sequences for sequenced microbial genomes were extracted from downloaded GenBank files using an in-house developed PERL script. Gene prediction for assembled shotgun metagenomes was carried out by FragGeneScan (Rho et al.

2010). Protein sequences were searched against the HMM models with an *e* value cutoff of 1e^{-5}. Probe design for coding sequences was carried out by the CommOligo program (Li et al. 2005). Candidate probes were then searched against the whole database for specificity using the NCBI BLAST program. The best probes were selected for microarray fabrication.

Strain-specific probe design was carried out using a *k*-mer based algorithm published previously (Tu et al. 2013, 2014b). In this approach, non-redundant strain level *k*-mers were first extracted for all 5390 microbial strains. Non-redundant *k*-mers were also extracted for the human genome. Second, *k*-mers showing up in more than two microbial strains as well as all human genome *k*-mers were kept to build a non-specific *k*-mer database. Third, all 50-mers were then extracted from targeted human microorganisms and searched against the non-specific *k*-mer database. All mapped 50-mers were then discarded from further analysis. Fourth, the remaining candidate 50-mers were then BLAST searched against all 5390 microbial genomes and human genome with 85 % sequence identity and self-annealing (≤8 bp) properties. Remaining probes were then ranked according to sequence identity with non-targets, melting temperature, free energy, and GC content as described previously (Li et al. 2005). For each strain, up to 15 probes from locations as distant as possible in the genome were selected for microarray fabrication. For strains with less than 15 probes, all of qualified probes were selected. Selected probes were then uploaded to the Agilent eArray system. Fabrication of HuMiChip2 was carried out by Agilent Technologies (Santa Clara, CA, USA). The 4 × 180 K format microarray was used for fabrication in this study.

Specificity evaluation using individual strains/species and mock communities

To evaluate the specificity of strain-specific probes, we used single individual strains/species and also constructed a series of mock communities from 15 sequenced microbial strains, including *Acinetobacter baumannii* ATCC 17978, *A. baumannii* ATCC 19606, *A. baumannii* AYE, *Bacillus cereus* ATCC 10876, *B. cereus* ATCC 4342, *Lactobacillus rhamnosus* ATCC 21052, *L. rhamnosus* ATCC 8530, *L. rhamnosus* GG, *Lactobacillus ruminis* ATCC 25644, *Bifidobacterium adolescentis* ATCC 15703, *Bifidobacterium dentium* ATCC 27678, *Bifidobacterium longum* subsp. *infantis* ATCC 15697, *Prevotella buccae* ATCC 33574, *Prevotella buccalis* ATCC 35310, and *Prevotella oralis* ATCC 33269. For each microbial strain, 5 ng DNA was used for mock community construction. For each mock community, three replications were carried out. The procedure for labeling and HuMiChip2 hybridization was the same as described below.

Sampling, DNA extraction, purification, and quantification

Human fecal samples were collected at the First Affiliated Hospital of Zhejiang University. A total of 18 individuals were recruited for the study, among whom 9 were diagnosed with alcoholic cirrhosis and 9 were healthy individuals with alcohol abuse. All patients were provided with written informed consent, and research was approved by the First Affiliated Hospital of Zhejiang University ethics committee and Institutional Review Broad (IRB). More details for the recruited individuals were previously described (Chen et al. 2014).

Fecal samples were immediately frozen on collection and stored at −80 °C before analysis. A frozen aliquot (200 mg) of each fecal sample was added to a 2.0-ml screwcap vial containing 300 mg glass beads of 0.1 mm diameter (Sigma, St. Louis, MO, USA) and kept on ice until the addition of 1.4-ml ASL buffer from the QIAamp DNA Stool Mini Kit (Qiagen, Valencia, CA, USA). Samples were immediately subjected to beadbeating (45 s, speed 6.5) using a FastPrep machine (Bio 101, Morgan Irvine, CA, USA), prior to the initial incubation for heat and chemical lysis at 95 °C for 5 min. Subsequent steps of DNA extraction followed the QIAamp kit protocol for pathogen detection.

We used the absorbance ratios at A260/A280 and A260/A230 using spectrophotometry (NanoDrop 1000, Thermo Fisher Scientific, Wilmington, DE, USA) to evaluate DNA quality. Final DNA concentrations were quantified with the Pico-Green kit (Invitrogen, Carlsbad, CA, USA). Only DNA samples with A260/A280 >1.7 and A260/A230 >1.8 were used. The extracted whole community DNA for each sample was shipped to the University of Oklahoma (OU) for HuMiChip2 analysis.

Target labeling, hybridization, imaging, and data preprocessing

For each sample, 1 μg of DNA was labeled with the fluorescent dye Cy-3 (GE Healthcare, Vacaville, CA, USA) using random primers and the Klenow fragment of DNA polymerase I (Wu et al. 2006). Labeled DNA was then purified using a QIAquick Purification kit (Qiagen, Valencia, CA, USA), dried in a SpeedVac at 45 °C for 45 min (Thermo Savant, Holbrook, NY, USA). Dried DNA was then rehydrated with 13 μl of DNase/RNase-free distilled water, mixed completely, and centrifuged to collect all liquid at the bottom of the tube. A total of 42 μl of buffer, including 1× HI-RPM hybridization buffer, 1× aCGH blocking agent, 0.05 μg μl^{-1} Cot-1 DNA, 10 pM universal standard, and 10 % formamide (final concentrations), was added to each sample. After mixing completely by vortexing, the solution was spun down and incubated at 95 °C for 3 min, then incubated at 37 °C for 30 min.

The samples were then hybridized with HuMiChip2 at 67 °C for 24 h with a rotation at 20 rpm in an Agilent hybridization oven (Agilent Technologies, Inc., Santa Clara, CA, USA). The scanned images of hybridized HuMiChip2 were converted and extracted using the Agilent Feature Extraction 11.5 software (Agilent Technologies, Inc., Santa Clara, CA, USA) for further data analysis. Probe spots with coefficient of variance (CV) greater than 0.8 were removed. Probes with signal-to-noise ratio (SNR) less than 2 and signal intensities less than 500 were also removed. Microarray data was then normalized based on the total signal intensity of common oligonucleotide reference standard (CORS) probes (Liang et al. 2010).

Statistical analysis

We used three different non-parametric multivariate analysis methods, adonis (permutational multivariate analysis of variance using distance matrices) (Anderson 2001), anosim (analysis of similarities) (Clarke 1993), and multi-response permutation procedure (MRPP) (McCune et al. 2002), as well as principle coordinate analysis (PCoA) (Gower 1966), to measure and visualize the overall differences of the community functional gene structure between treatment and control samples. The significance of relative abundance differences between patients and healthy individuals for functional gene categories was evaluated by the response ratio analysis (Luo et al. 2006). For strain-specific probes, a threshold of 10 out of 15 probes being positively detected was set for positive calls of detected strains. Signal intensities of positive probes were then averaged and log-transformed to reflect the abundance of detected strains. Heat map analysis was used to visualize strain-specific probes across multiple samples. The Student's *t* test was used to estimate significance *p* values between healthy individuals and liver cirrhosis patients.

Availability HuMiChip2 is available through Glomics Inc. (Norman, OK, USA). Microarray data generated in this study are available under NCBI accession number GSE86162.

Results

An overall description of the HuMiChip2 architecture

HuMiChip2 was composed of two different types of probes, including strain/species-specific probes and functional gene probes, for the purposes of strain/species-level identification and functional profiling of human microbiomes, respectively. For strain/species-specific detection of human microbiomes, a total of 2063 sequenced microbial genomes from human body were recruited, including 3 archaeal strains, 28 eukaryotic strains, and 2032 bacterial strains (Table 1). Briefly, these strains covered 15 microbial phyla, 28 classes, 47 orders, 105 families, and 226 genera. For strains with more than 15 qualified probes, 15 probes more evenly distributed in the genome were selected according to their location on the genome. For strains with fewer than 15 qualified probes, all probes were selected. This resulted in a total of 29,467 strain-specific probes on HuMiChip2. At the phylum level, *Proteobacteria* and *Firmicutes* were the phyla with the most probes (11,560 and 11,077, respectively) and targeted strains (812 and 782, respectively), followed by *Actinobacteria* (225 strains and 3286 probes) and *Bacteroidetes* (155 strains and 2290 probes). At the class level, major taxonomic groups included *Bacilli* (556 strains and 7749 probes), *Gammaproteobacteria* (513 strains and 7432 probes), *Actinobacteria* (225 strains and 3286 probes), *Clostridia* (177 strains and 2595 probes), *Epsilonproteobacteria* (141 strains and 1874 probes), *Betaproteobacteria* (124 strains and 1775 probes), and *Bacteroidia* (110 strains and 1642 probes).

For microbial functional gene probes, a total of 157 gene families were selected as they play important roles in the human body. These 157 gene families were involved in at least 11 microbial functional processes (Table 2), including antibiotic resistance (18 gene families, 13,567 probes), amino acid metabolism (36 gene families, 30,037 probes), carbohydrate metabolism (27 gene families, 19,764 probes), energy metabolism (5 gene families, 8993 probes), glycan biosynthesis and metabolism (10 gene families, 8570 probes), lipid metabolism (5 gene families, 4435 probes), metabolism of non-essential amino acids (20 gene families, 18,460 probes), metabolism of cofactors and vitamins (16 gene families, 12,305 probes), metabolism of terpenoids and polyketides (5 gene families, 4505 probes), nucleotide metabolism (13 gene families, 10,584 probes), and translation (2 gene families, 2704 probes). These resulted in a total of 133,924 probes, including 94,387 sequence-specific probes and 39,537 group-specific probes, and covered 276,240 coding sequences.

Computational specificity evaluation

Computational evaluation of the specificity of strain/species-specific probes was performed as previously described (Tu et al. 2014b). Briefly, the specificity of strain-specific probes was reflected in the following aspects (Table 3). First, these probes were unique to currently sequenced microbial genomes included in the currently available database (~5400 microbial genomes). Second, all probes were searched against four metagenome datasets from mock communities of 21 strains, of which 16 were included in the genome list. As a result, 100 % true positives were found for evenly distributed mock communities and 75~87.5 % true positives were found for staggered mock communities. Third, the strain/species-

Table 1 Summary of strain-specific probes organized by taxonomic groups

Domain	Phylum	Class	No. of orders	No. of families	No. of genera	No. of strains	No. of probes
Bacteria	*Firmicutes*	*Bacilli*	2	13	20	556	7749
		Clostridia	1	9	35	177	2595
		Negativicutes	1	2	10	33	494
		Erysipelotrichia	1	1	9	16	239
		Sum	5	25	74	782	11,077
	Proteobacteria	*Gammaproteobacteria*	10	13	37	513	7432
		Epsilonproteobacteria	1	2	3	141	1874
		Betaproteobacteria	2	6	15	124	1775
		Alphaproteobacteria	5	8	13	30	420
		Deltaproteobacteria	1	1	2	4	59
		Sum	19	30	70	812	11,560
	Actinobacteria	*Actinobacteria*	3	18	32	225	3286
	Bacteroidetes	*Bacteroidia*	1	4	9	1642	1642
		Fusobacteria	1	2	2	27	389
		Flavobacteria	1	1	4	15	214
		Sphingobacteria	1	1	1	2	30
		Unclassified	N/A	N/A	N/A	1	15
		Sum	4	8	16	155	2290
	Spirochaetes	*Spirochaetia*	1	2	3	24	321
	Tenericutes	*Mollicutes*	1	1	2	17	228
	Chlamydia	*Chlamydia*	1	3	4	9	120
	Synergistetes	*Synergistia*	1	1	5	5	75
	Cyanobacteria	Unclassified	1	N/A	1	1	15
	Lentisphaerae	*Lentisphaeria*	1	1	1	1	15
	Verrucomicrobia	*Verrucomicrobiae*	1	1	1	1	15
Archaea	*Euryarchaeota*	*Methanobacteria*	1	1	2	3	45
Eukaryota	*Ascomycota*	*Eurotiomycetes*	2	5	6	14	210
		Dothideomycetes	1	1	1	1	15
		Saccharomycetes	1	2	2	3	45
		Sum	4	8	9	18	270
	Apicomplexa	*Aconoidasida*	1	1	1	3	45
		Coccidia	1	1	1	2	30
		Sum	2	2	2	5	75
	Basidiomycota	*Exobasidiomycetes*	1	1	1	2	30
		Tremellomycetes	1	1	1	1	15
		Sum	2	2	2	3	45
	Unclassified	Unclassified	N/A	2	2	2	30
Sum	15	28	47	105	226	2063	29,467

specific probes were also specific to newly sequenced microbial genomes. A total of 302 newly sequenced microbial strains were collected for the evaluation. The results suggested that ~67 % new genomes could not be targeted by these strain-specific probes, and that the ~25 % targeted genomes were identified by probes belonging to strains in the same species. Fourth, the strain-specific probes were also specific to the body sites where they were isolated as they were evaluated by nine shotgun metagenome datasets from different body sites. Probes targeting microbial strains isolated from a particular body sites were rarely found in metagenomes of other body sites. All these results suggested that all designed strain-specific probes in this study were highly specific to their targets and could be confidently applied to perform the microbial strain-level identification of human microbiomes.

Table 2 Summary of functional gene probes organized by functional processes

Microbial functional process	No. of gene families	No. of sequence-specific probes	No. of group-specific probes	No. of targeted CDS
Antibiotic	18	0	13,567	32,168
Amino acid metabolism	36	24,314	5723	59,392
Carbohydrate metabolism	27	15,322	4442	40,173
Energy metabolism	5	7115	1878	16,799
Glycan biosynthesis and metabolism	10	6010	2560	20,277
Lipid metabolism	5	3262	1173	8984
Metabolism of non-essential amino acids	20	14,420	4040	37,656
Metabolism of cofactors and vitamins	16	9700	2605	25,854
Metabolism of terpenoids and polyketides	5	3644	861	8398
Nucleotide metabolism	13	8523	2061	20,485
Translation	2	2077	627	6054
Sum[a]	157	94,387	39,537	276,240

[a] The total number of probes and covered coding sequences is based on non-redundant genes included in all pathways. Overlap of functional genes may occur among different functional processes

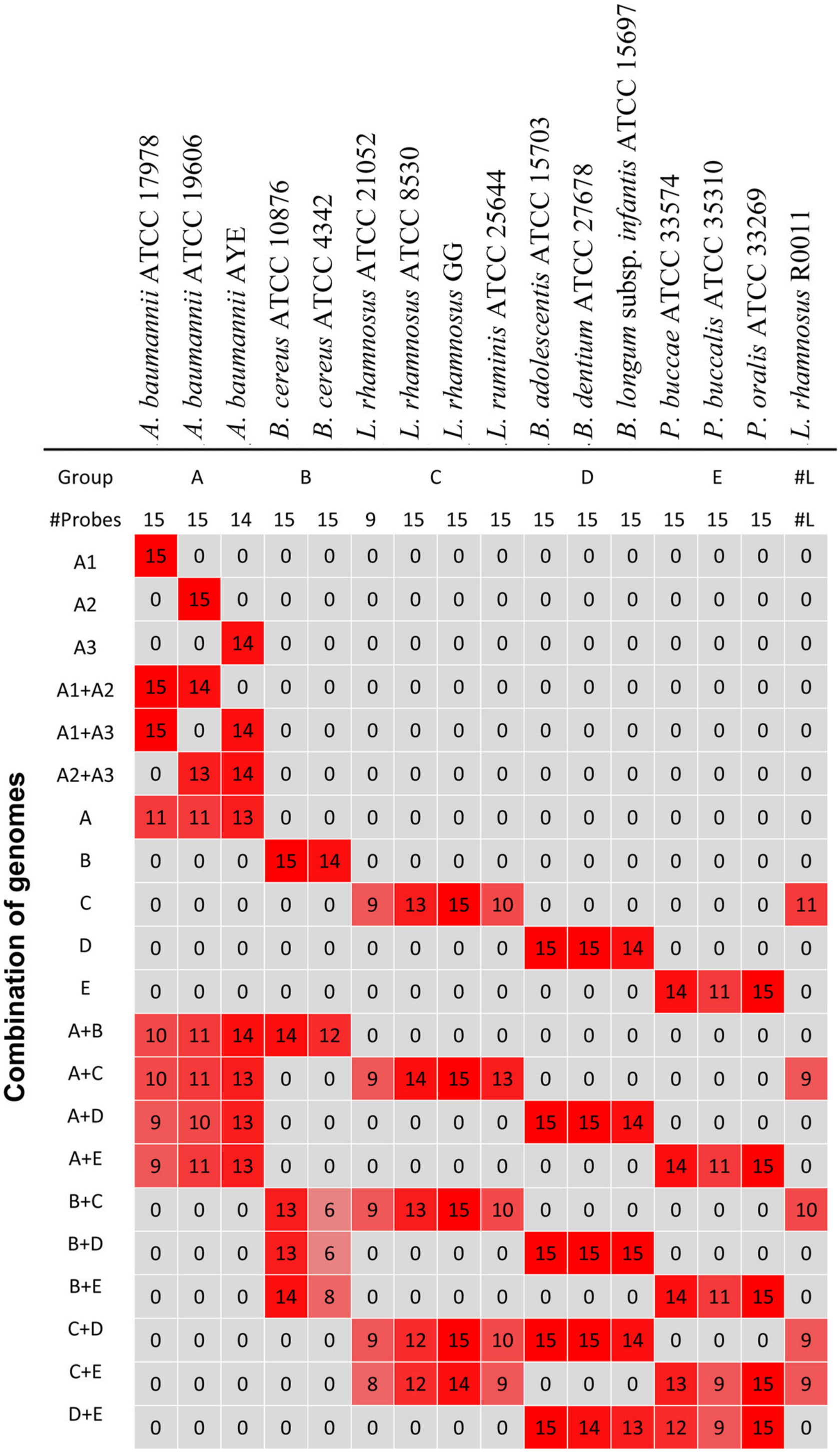
Experimental specificity evaluation using mock communities

In order to validate their performance in real case application, the specificity of strain-specific probes was also evaluated experimentally. The evaluation included an incremental complexity of microorganisms or mock communities using DNAs from (i) one single strain (three cases) and (ii) mock communities, including (a) two to multiple strains in a same species/genus (eight cases) and (b) multiple strains from two different species/genera (ten cases). A total of 15 microbial strains from *Acinetobacter* (3 strains), *Bacillus* (2 strains), *Lactobacillus* (4 strains), *Bifidobacterium* (3 strains), and *Prevotella* (3 strains) genera were selected for the evaluation. As a result, high specificity was observed for all combinations of genomes we tested (Fig. 1). Specifically, the following results were observed for the evaluation below.

(i) *Single-strain test.* We selected three strains of the species *A. baumannii* for this evaluation. DNA of each of the

Table 3 Summary of computational evaluation for strain-specific probes

Dataset	Dataset size	Methodology in brief	False-positives/negatives
Sequenced microbial genomes	5390 genomes	BLAST searching strain-specific probes against all non-target genomes and human genome	Not applicable, all selected strain-specific probes are highly specific to target genomes
Mock community metagenome	4 dataset (2 by Illumina and 2 by 454)	BLAST searching strain-specific probes against shotgun metagenomes	No false positives for even mock communities, 2~4 false negatives and 3 false positives (1 probe only) for staggered mock communities
Newly sequenced microbial genomes	302 recently sequenced microbial genomes	BLAST searching strain-specific probes against all 302 genomes	67.2 % strains cannot be targeted by current probes, 24.8 % were assigned to the same species, 4.6 % to the same genus
Metagenomes from different human body sites	Nine shotgun metagenomes from different body sites	BLAST searching strain-specific probes against all metagenomes	Probes targeting microbial strains from the gut, skin, and urogenital tract mainly (>95 %) hit metagenomes from corresponding body sites. Probes targeting microbial strains from oral and airways share hits with metagenomes from subgingival plaque, tongue dorsum, throat, and palatine tonsils



Combination of genomes	*A. baumannii* ATCC 17978	*A. baumannii* ATCC 19606	*A. baumannii* AYE	*B. cereus* ATCC 10876	*B. cereus* ATCC 4342	*L. rhamnosus* ATCC 21052	*L. rhamnosus* ATCC 8530	*L. rhamnosus* GG	*L. ruminis* ATCC 25644	*B. adolescentis* ATCC 15703	*B. dentium* ATCC 27678	*B. longum* subsp. *infantis* ATCC 15697	*P. buccae* ATCC 33574	*P. buccalis* ATCC 35310	*P. oralis* ATCC 33269	*L. rhamnosus* R0011
Group		A		B			C				D			E		#L
#Probes	15	15	14	15	15	9	15	15	15	15	15	15	15	15	15	#L
A1	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A2	0	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A3	0	0	14	0	0	0	0	0	0	0	0	0	0	0	0	0
A1+A2	15	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A1+A3	15	0	14	0	0	0	0	0	0	0	0	0	0	0	0	0
A2+A3	0	13	14	0	0	0	0	0	0	0	0	0	0	0	0	0
A	11	11	13	0	0	0	0	0	0	0	0	0	0	0	0	0
B	0	0	0	15	14	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	0	9	13	15	10	0	0	0	0	0	0	11
D	0	0	0	0	0	0	0	0	0	15	15	14	0	0	0	0
E	0	0	0	0	0	0	0	0	0	0	0	0	14	11	15	0
A+B	10	11	14	14	12	0	0	0	0	0	0	0	0	0	0	0
A+C	10	11	13	0	0	9	14	15	13	0	0	0	0	0	0	9
A+D	9	10	13	0	0	0	0	0	0	15	15	14	0	0	0	0
A+E	9	11	13	0	0	0	0	0	0	0	0	0	14	11	15	0
B+C	0	0	0	13	6	9	13	15	10	0	0	0	0	0	0	10
B+D	0	0	0	13	6	0	0	0	0	15	15	15	0	0	0	0
B+E	0	0	0	14	8	0	0	0	0	0	0	0	14	11	15	0
C+D	0	0	0	0	0	9	12	15	10	15	15	14	0	0	0	9
C+E	0	0	0	0	0	8	12	14	9	0	0	0	13	9	15	9
D+E	0	0	0	0	0	0	0	0	0	15	14	13	12	9	15	0

Fig. 1 Specificity evaluation of strain-specific probes using single genomes and mock communities. A total of 21 single genome and mock communities were generated for the evaluation. No false-positive detections could be observed for all types of mock communities except the ones with *Lactobacillus* strains, for which the strain *L. rhamnosus* R0011 was false positively detected. Groups A–E represent microbial species/genera involved in the test, including *A. baumannii*, *B. cereus*, *Lactobacillus*, *Bifidobacterium*, and *Prevotella*, respectively

three strains (each with three technical replicates) was hybridized with HuMiChip2 to evaluate the specificity. A total of 20 *A. baumannii* strains were targeted by HuMiChip2, thus the evaluation could exactly tell whether false positives or false negatives occurred among those closely related microbial strains. It was expected only the target strain had positive hybridization signals if those probes were strain-specific. Our results did show that no false positives or false negatives were observed. All three strains were well hybridized with all 44 probes targeting these three strains (15 probes for *A. baumannii* ATCC 17978 and 19606, and 14 probes for *A. baumannii* AYE) on HuMiChip2.

(iia) *Multi-strain in the same species/genus mock community test.* We also performed a multi-strain test encompassing DNAs from several strains in the same species/genus. A total of 87 strains belonging to these 5 genus were targeted by HuMiChip2, including 20 for *A. baumannii*, 39 for *B. cereus*, 6 for *L. rhamnosus*, 1 for *L. ruminis*, 2 for *B. adolescentis*, 3 for *B. dentium*, 12 for *B. longum*, 3 for *P. buccae*, and 1 for *P. oralis*. This test was expected to address the specificity of HuMiChip2 when multiple strains in the same species were present in the DNA. Our results showed no false-positive detection for four groups although it appeared to be more false negatives (with signal intensities <1000) detected for the three-strain group comparing with one- or two-strain group. For example, the number of false negatives detected for *A. baumannii* strains were respectively four (ATCC 17978), four (ATCC 17978), and one (AYE). We also found a false-positive detection for *L. rhamnosus* R0011, which could be potentially caused by incomplete genome sequences of some strains in *L. rhamnosus* (Fig. 1).

(iib) *Multi-species mock community test.* We finally mixed multiple strains from two species/genera as more complex mock communities to evaluate the specificity of HuMiChip2. A total of ten mock communities were generated by mixing strains from two species/genera, including *A. baumannii/B. cereus*, *A. baumannii/Lactobacillus*, *A. baumannii/Bifidobacterium*, *A. baumannii/Prevotella*, *B. cereus/Lactobacillus*, *B. cereus/Bifidobacterium*, *B. cereus/Prevotella*, *Lactobacillus/Bifidobacterium*, *Lactobacillus/Prevotella*, and *Bifidobacterium/Prevotella*. Similarly, a total of 87 strains belonging to these five genera were targeted by HuMiChip2. Again, no false positive was observed for all mock communities we tested except those with *Lactobacillus* strains, for which *L. rhamnosus* R0011 was detected even it was not added. Also, false negatives were detected for several strains, such as *A. baumannii* ATCC 17978 (6 out of 15), *A. baumannii* ATCC 19606 (5 out of 15), *B. cereus* ATCC 4342 (9 out of 15), *L. ruminis* ATCC 25644 (6 out of 15), and *P. buccalis* ATCC 35310 (6 out of 15). This indicated some potential competition of some probes on the HuMiChip2 by highly similar genome sequences from other strains. All the above experimental evaluations suggested the strain-specific probes on the HuMiChip2 were highly specific, as revealed by single genomes and mock communities.

Application of HuMiChip2 to characterize human gut microbiomes

To evaluate the performance of newly developed HuMiChip2 with real human microbiome samples, we used it to analyze the strain-level identification and functional profiling of human gut microbiomes (a total of 18 samples) with nine for each group—alcoholic cirrhosis patients and healthy individuals with alcohol abuse. In this analysis, we aimed to address the following two questions: (i) Do alcoholic abuse individuals and cirrhosis patients harbor specific microbial strains? (2) Do alcoholic abuse individuals and cirrhosis patients show different functional profiles as a result of alcoholic liver cirrhosis?

At the strain level, a total of 58 microbial strains were detected in at least one of 18 human fecal samples. Among these, 33 were found to be present in less than three samples. Of the 25 microbial strains detected in more than 3 samples, 18 were identified as *Bacteroides* strains, two as *Alistipes* strains, one as *Clostridium* strain, one as *Eubacterium* strain, one as *Fusobacterium* strain, one as *Parabacteroides* strain, and one as *Roseburia* strain (Fig. 2). Of them, five strains of *Bacteroides* including *Bacteroides* sp. D1, *Bacteroides* sp. 2_2_4, *Bacteroides ovatus* SD CMC 3f, *Bacteroides xylanisolvens* XB1A, and *Bacteroides* sp. 1_1_30 were significantly ($p < 0.1$) higher in liver cirrhosis patients. Another five strains, including *Fusobacterium mortiferum* ATCC 9817, *Eubacterium eligens* ATCC 27750, *Roseburia inulinivorans* DSM 16841, *Clostridium difficile* R20291, and *Bacteroides stercoris* ATCC 43183, though not significant, were more frequently detected in healthy individuals. This indicated a potential link between these microbial strains and human health, thus these strains can be potentially used as indicators for healthy status. The results also suggested that core microorganisms did not exist at the strain level among the 18 samples and *Bacteroides* strains seemed to be more commonly detected across different individuals, at least based on strain-specific probes targeting 2063 strains in this study.

At the functional gene level, a total of 15,318 probes were detected in at least one of 18 samples. The number of probes detected in each sample varied from 3262 to 8009, with an average of 5137 probes in alcoholic liver cirrhosis patients and 4233 probes in healthy individuals with alcoholic abuse. No significant difference ($p > 0.1$) was observed between healthy individuals and liver cirrhosis patients regarding the number of detected probes. Notably, the functional gene profiles between healthy individuals and liver cirrhosis patients were markedly different. PCoA analysis suggested a clear separation of healthy individuals with alcoholic abuse from alcoholic liver cirrhosis patients, except two samples (AB5 and CD15) being

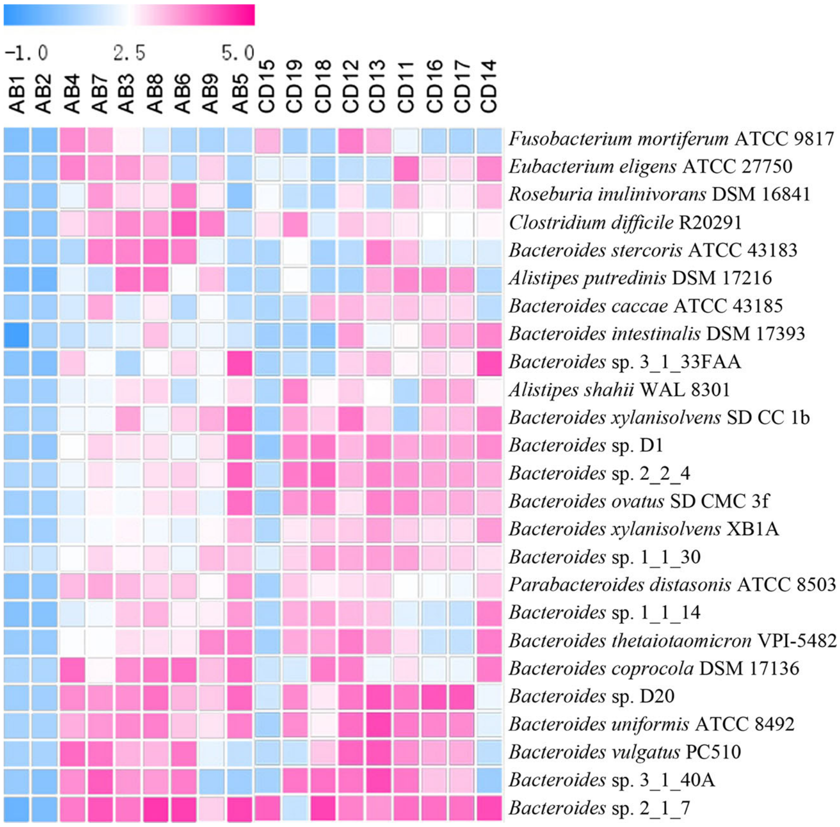


Fig. 2 Microbial strains detected in more than three human fecal samples as revealed by strain-specific probes. Normalized signal intensities were plotted. Five *Bacteroides* strains including *Bacteroides* sp. D1, *Bacteroides* sp. 2_2_4, *Bacteroides ovatus* SD CMC 3f, *Bacteroides xylanisolvens* XB1A, and *Bacteroides* sp. 1_1_30 were significantly higher in CD individuals with *p* value ≤0.1. *AB* healthy individuals with alcoholic abuse, *CD* liver cirrhosis patients

clustered to the other group (Fig. 3). The significantly different functional profiles between healthy individuals with alcoholic abuse and alcoholic liver cirrhosis patients was also verified by three non-parametric statistical methods, including MRPP ($\delta = 0.334$, $p = 0.002$), ADONIS ($F = 0.127$, $p = 0.002$), and ANOSIM ($R = 0.166$, $p = 0.009$).

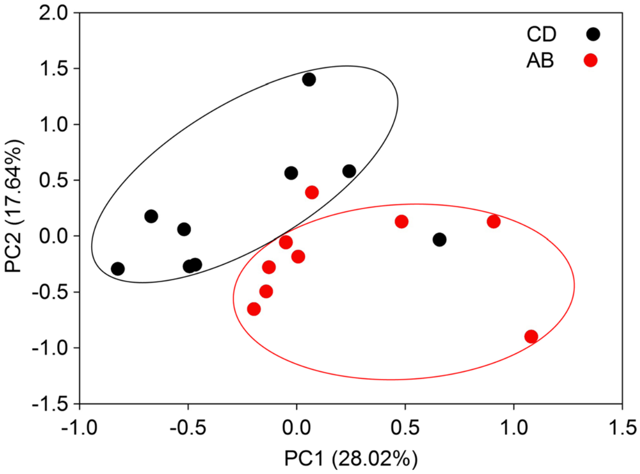


Fig. 3 Principle coordinate analysis of functional gene profiles between alcoholic cirrhosis patients and healthy individuals with alcohol abuse. A clear separation of liver cirrhosis patients from healthy individuals could be observed. *Red* represents healthy individuals with alcoholic abuse, and *black* represents alcoholic liver cirrhosis patients. *Ellipses* were added to better visualize the separation of liver cirrhosis patients from healthy individuals, as also suggested by three non-parametric statistical methods. *AB* healthy individuals with alcoholic abuse, *CD* liver cirrhosis patients

Further analysis showed that relative abundances of 21 functional gene families were significantly changed between healthy individuals with alcoholic abuse and alcoholic liver cirrhosis patients at the 90 % confidence interval as revealed by response ratio analysis (Fig. 4). Among these, four antibiotic resistance genes (*van*, *fosX*, ABC multidrug fungal gene, and MATE antibiotic gene) increased in the liver cirrhosis patients group. Of the five gene families related with amino acids metabolism, four of them (L-alanine dehydrogenase, PRAMP-cyclohydrolase, ATP phosphoribosyltransferase, and D-cystein desulfhydrase) increased and one (aspartate kinase) decreased in the liver cirrhosis patients group. For the three changed carbohydrate metabolism gene families, L-lactate dehydrogenase decreased and extracellular exopectate hydrolase and transaldolase increased in the liver cirrhosis patients group. Four gene families involved in glycan biosynthesis and metabolism changed in the liver cirrhosis patients group with three gene families (α- and β-mannosidase, and β-D-glucuronidase) increased and one gene family (*N*-acetylglucosamine acyltransferase) decreased. In addition, two gene families involved in energy metabolism, one in lipid metabolism, one in cofactor and vitamin metabolism, and one in nucleotide metabolism changed in the liver cirrhosis patients group. Of these, the β-ketoacyl-ACP synthase III related with lipid metabolism decreased in the liver cirrhosis patients group (Fig. 4).

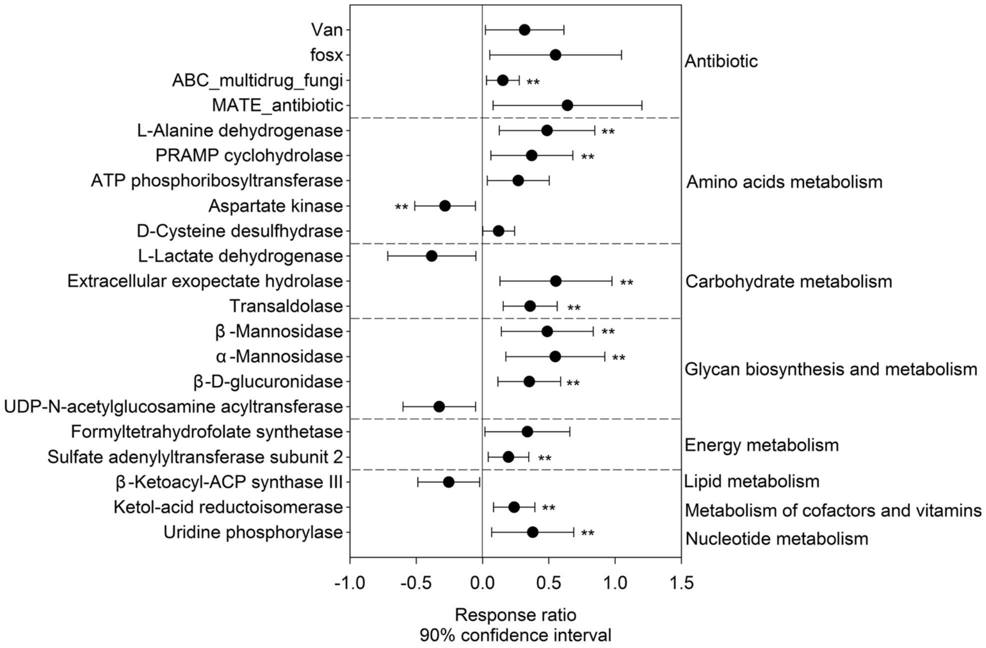


Fig. 4 Response ratio analysis of functional gene families between healthy individuals with alcoholic abuse and alcoholic liver cirrhosis patients. Genes with response ratio >0 indicated increased abundance under cirrhosis condition. The results suggested a series of significantly changed gene families as a result of liver cirrhosis. *Double asterisks* indicate significant changes at 95 % confidence interval

Discussion

Identifying microorganisms at the strain/species level and functionally characterizing human microbiomes are challenging due to the high diversity and complex microbial community composition and structure. In this study, by taking advantage of the mature functional gene array technology (He et al. 2012a, b; Tu et al. 2014a, c) and a *k*-mer-based algorithm for strain-specific probe design (Tu et al. 2013, 2014b), we developed the HuMiChip2 for both strain-level identification and functional profiling of human microbiomes.

Current human microbiome studies are mainly carried out by shotgun metagenome and 16S rRNA amplicon sequencing approaches, and have gained new insights into our understanding of the linkage between human microbiomes and human disorders, such as obesity (Turnbaugh et al. 2009), type 2 diabetes (Qin et al. 2012), liver cirrhosis (Chen et al. 2011; Qin et al. 2014), kwashiorkor (Smith et al. 2013), and periodontitis (Li et al. 2014). These efforts, together with the Human Microbiome Project (Peterson et al. 2009), generated valuable sequence datasets for human microbiome studies. However, transforming such large datasets into useful information and knowledge requires bioinformatics tools and metagenomics technologies like microarrays. Microbial ecological microarray is such a technology that can extract useful information from sequence data, quickly identify microbial taxa, and/or functionally profile microbial communities (He et al. 2012a, b).

Several types of microbial ecological microarrays suitable for human microbiome studies have been developed in the past years (Table 4). These include PhyloChip (DeSantis et al. 2007), HITChip (Rajilić-Stojanović et al. 2009), HuGChip (Tottey et al. 2013), and HuMiChip (Tu et al. 2014a). Among these, PhyloChip and HuMiChip/HuMiChip2 are more generic than HITChip and HuGChip that PhyloChip and HuMiChip/HuMiChip2 can be used to profile human microbiome from various body sites, while the latter two are specifically designed for human gut microbiomes. Compared with other available microarrays for the human microbiome, HuMiChip and HuMiChip2 are designed to profile selected functional gene families in the human microbiome. Although HuMiChip2 is equipped with strain-specific probes for more than 2300 strains, microbial ecological microarrays based on 16S rRNA genes (e.g., PhyloChip, HuGChip, and HITChip) hold the advantage for more comprehensive taxonomic profiling capability.

Strain-level identification of microorganisms in the environment is challenging. Current approaches in microbial ecology studies, e.g., 16S rRNA gene amplicon sequencing, can only confidently detect microbial taxa at the genus or sub-family level due to short and highly conserved regions, thus strain-level identification of microorganisms requires moving beyond single marker gene (e.g., 16S rRNA gene) to genome-wide analyses (Faith et al. 2015). In this study, we showed that HuMiChip2 could achieve a strain-level identification of microorganisms in human microbiomes owing to the advantage of strain-specific probes designed using a *k*-mer-based algorithm (Tu et al. 2013, 2014b), which can also be used for shotgun metagenome analysis. Similar to many other metagenomic approaches such as canSNPs (Karlsson et al. 2014), PathoScope (Hong et al. 2014), and Sigma (Ahn et al. 2015), such detection relies on reference genomes. Among various metagenomic samples such as human, soil, and water, the human microbiome has most reference genomes available,

Table 4 Comparison of microbial ecological microarrays suitable for human microbiome studies

	Targeted gene families	No. of probes	No. of targeted taxonomy	Main application
PhyloChip	16S rRNA genes	297,851	842 subfamilies	Taxonomic profiling of microbial communities in various environments
HuGChip	16S rRNA genes	4441	66 families	Taxonomic profiling of human gut microbiome with explorative capacity
HITChip	16S rRNA genes	4809	1140 phylotypes	Taxonomic profiling of human gut microbiome
HuMiChip	139 functional gene families	36,802	322 strains + metagenomes	Functional profiling of human microbiome
HuMiChip2	157 functional gene families + strain-specific probes	133,924	2063 strains + metagenomes	Functional profiling and strain/species-level identification of human microbiome

thus is especially suitable for strain-level identification using microarray technologies. Notably, with more reference genomes being generated by the scientific community, this technology can also be applied to analyze more complex metagenomes, such as soil and water microbial communities. Compared with shotgun metagenome sequencing for microbial identification, microarray technologies are expected to be more sensitive in detecting low abundant microorganisms (Zhou et al. 2015) as genes of interest and specific regions may only comprise a tiny portion of the genome and may not be well captured by sequencing technologies. Although this issue could be potentially solved by increasing the sequencing depth, the high cost to capture enough sequences for analysis would prevent many scientists from doing so.

Specificity is the most critical issue in microbial ecological microarrays, especially for strain-level identification of microorganisms. In this study, such specificity is first realized by strain-specific probe design and associated algorithms and criteria. This includes comparative genomic search for specific probes using *k*-mer-based ($k \leq 20$) algorithms against more than 5000 genomes with a sequence similarity cutoff of 80 % in addition to GC content, free energy, and annealing temperature. Second, we use multiple probes (up to 15 probes) per strain, which allows us to statistically analyze the strain-level detection, such as the number of probes (10 out of 15) required for reliable detection of certain strains. With such criteria, potential false positives from close relatives in the same species or genus could be effectively excluded. Third, when >15 probes are available, probes are selected from multiple locations in the genome to maximize even distribution. All those strategies guaranteed specific detection of microorganisms at the whole genome level, instead of a few specific fragments. Furthermore, to insure the high specificity of strain-specific probes, we performed both computational (Tu et al. 2013) and experimental evaluation for all and randomly selected probe sets, respectively. In general, very high specificity could be observed for strain-specific probes. For example, about 70 % of the 302 newly sequenced genomes could not be targeted by current strain-specific probes, and about 25 % could be assigned to the same species (Tu et al. 2013). Such specificity could also be observed against shotgun metagenome datasets that microbial strains isolated from a particular human body sites were mainly found in metagenomes of that body site, with a ratio of as high as >99 % in body sites such as human gut, skin, and urogenital tract (Tu et al. 2013). Experimental evaluation using single genome and mock communities showed no false positives for the majority randomly selected species, except for the species *L. rhamnosus* that the *L. rhamnosus* R0011 strain was detected as a false positive when other *L. rhamnosus* strains were included in the mock community. This issue may be hardly avoided as 9~11 out of 15 probes were detected for *L. rhamnosus* R0011 in the test. This is largely due to incomplete genome sequencing of some *L. rhamnosus* strains, and/or potential lateral gene transfer events among them (Soucy et al. 2015; Vos et al. 2015). Notably, no other false positives were detected in the test, suggesting an overall high specificity of HuMiChip2. False negatives, however, were also detected despite at a very low rate when multiple strains were mixed in the mock community. This may be largely due to potential competition in hybridization from those close relatives in the same species. This effect, to our best knowledge, is difficult to avoid although an increase of strain-specific probe number for each strain/species may be one of strategies. However, this issue could also be an exaggerated effect with mock communities, which are much less diverse than real samples.

Liver cirrhosis is the pathologic end stage of chronic liver disease as a result of several causes, such as obesity, hepatitis virus infection, and alcohol abuse (Qin et al. 2014). About 30 % alcoholics develop liver disease, and the reasons why certain individuals are more susceptible are not known (Bunout 1999; Diehl 1989). Many recent studies suggested a clear linkage between gut microbiome community changes and liver cirrhosis (Bajaj et al. 2014; Chen and Schnabl 2014; Chen et al. 2011, 2014; Qin et al. 2014). Here, HuMiChip2 was applied to perform both strain-level identification and the functional profiling of human gut microbiomes from alcoholic cirrhosis patients and healthy individuals with alcohol abuse. A significant shift of gut microbiomes in terms of functional composition and structure was observed between alcoholic liver cirrhosis patients and healthy individuals with alcohol abuse. Further analysis suggested that the changed gene families mainly belonged to functional categories such as antibiotic, amino acids metabolism, carbohydrate metabolism, glycan biosynthesis and metabolism, and energy metabolism. The increased relative abundances of these gene families are in agreement with the clinical observation that the metabolism of nutrition, carbohydrate, protein, and lipids are suppressed in liver cirrhosis patients (Bunout 1999). The gut microbiome, which habitat the internal human body in a symbiotic manner (Peterson et al. 2009), may help to restore the ability in these metabolic processes in human gut by increasing corresponding gene families. The results are also generally consistent with recent shotgun metagenomic studies, which suggested a series of marker genes to discriminate patients and healthy individuals (Qin et al. 2014). Interestingly, gene families involved in these categories were also regarded as marker genes in patients of type 2 diabetes (Qin et al. 2012). At the strain level, notably, we did not see a core set of microbial strains among patients or healthy individuals that could be linked with the disease, except that five *Bacteroides* strains were more frequently detected in the cirrhosis patients group. The low detection number of microbial strains could be due to the reason that a core microbiome may not exist at the organismal lineage level (Turnbaugh et al. 2009) as well as the high specificity of these strain-specific probes (Tu et al. 2013, 2014b). The results may provide better insights into the linkages between gut microbiome and human disorders like liver cirrhosis (Bajaj et al. 2014; Chen and Schnabl 2014; Chen et al. 2011, 2014; Qin et al. 2014; Smith et al. 2013).

In conclusion, we present a comprehensive microbial ecological microarray—HuMiChip2, for both strain-level identification and functional profiling of human microbiomes. HuMiChip2 targets more than 2000 sequenced microbial genomes, and 2.5 Gb assembled high-quality shotgun metagenomes from 14 different body sites. HuMiChip2 carries the ability of strain-level identification of human microbiomes. This makes HuMiChip2 a unique high throughput technology for quickly detecting known human microbial strains, especially pathogens in various human samples as well as environmental samples, providing novel insights into the epidemiology of human pathogens at the strain level. To our best knowledge, this is the first microbial ecological microarray equipped with strain-level detection ability. However, owing to the high variation of human microbiomes at the organismal lineage level (Turnbaugh et al. 2009), group-specific probes at sub-species level may need to be designed in the future to increase the number of detected strains, as well as for detecting novel microbial strains in human microbiome samples. In addition, HuMiChip2 can be used for functional profiling of key gene families in various pathways. Therefore, the newly developed HuMiChip2 presents a useful example in converting large volume of sequence data generated by the Human Microbiome Project (Peterson et al. 2009) into useful tools and knowledge, and can be widely used in human microbiomes studies. The same strategy may be used in other environmental microbiome studies when enough reference genomes are available for those ecosystems (soil, waters, and sediments).

Author contribution ZH, JZ, QT, and Lanjuan Li conceived and designed the study. Jiabao Li, QT, ZS, YC, Juan Li, QZ, and XL performed the research. QT, Jiabao Li, ZS, and Lu Lin analyzed the data. QT, ZH, HW, and JY wrote the manuscript. All authors read and approved the manuscript.

Compliance with ethical standards

Funding This work was supported by the Oklahoma Center for the Advancement of Science and Technology (OCAST) through the Oklahoma Applied Research Support (OARS) Project AR11-035, the Fundamental Research Funds for the Central Universities of China (Zhejiang University, 2016QNA4039, 2015QNA4044), and the Open Funding of Zhejiang Provincial Key Laboratory of Health Risk Factors for Seafood (201605), Zhoushan Center for Disease Control and Prevention.

Conflict of interests The authors declare that they have no conflict of interest.

Ethical approval All human individuals involved in this study were provided with written informed consent, and research was approved by the First Affiliated Hospital of Zhejiang University ethics committee and Institutional Review Broad (IRB). All procedures performed in studies involving human participants were in accordance with the ethical standards of the First Affiliated Hospital of Zhejiang University ethics committee and Institutional Review Broad (IRB) and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

References

Ahn TH, Chai J, Pan C (2015) Sigma: strain-level inference of genomes from metagenomic analysis for biosurveillance. Bioinformatics 31(2):170–177

Anderson MJ (2001) A new method for non-parametric multivariate analysis of variance. Austral Ecol 26(1):32–46

Bajaj JS, Heuman DM, Hylemon PB, Sanyal AJ, White MB, Monteith P, Noble NA, Unser AB, Daita K, Fisher AR (2014) Altered profile of human gut microbiome is associated with cirrhosis and its complications. J Hepatol 60(5):940–947

Bunout D (1999) Nutritional and metabolic effects of alcoholism: their relationship with alcoholic liver disease. Nutrition 15(7–8):583–589

Caporaso JG, Lauber CL, Costello EK, Berg-Lyons D, Gonzalez A, Stombaugh J, Knights D, Gajer P, Ravel J, Fierer N (2011) Moving pictures of the human microbiome. Genome Biol 12(5):R50

Chen P, Schnabl B (2014) Host-microbiome interactions in alcoholic liver disease. Gut Liver 8(3)

Chen Y, Yang F, Lu H, Wang B, Chen Y, Lei D, Wang Y, Zhu B, Li L (2011) Characterization of fecal microbial communities in patients with liver cirrhosis. Hepatology 54(2):562–572

Chen Y, Qin N, Guo J, Qian G, Fang D, Shi D, Xu M, Yang F, He Z, Van Nostrand JD (2014) Functional gene arrays-based analysis of fecal microbiomes in patients with liver cirrhosis. BMC Genomics 15(1):1

Cho I, Blaser MJ (2012) The human microbiome: at the interface of health and disease. Nat Rev Genet 13(4):260–270

Clarke KR (1993) Non-parametric multivariate analyses of changes in community structure. Aust J Ecol 18(1):117–143

DeSantis TZ, Brodie EL, Moberg JP, Zubieta IX, Piceno YM, Andersen GL (2007) High-density universal 16S rRNA microarray analysis reveals broader diversity than typical clone library when sampling the environment. Microb Ecol 53(3):371–383

Diehl AM (1989) Alcoholic liver disease. Med Clin North Am 73(4): 815–830

Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14(9): 755–763

Faith JJ, Colombel J-F, Gordon JI (2015) Identifying strains that contribute to complex diseases through the study of microbial inheritance. Proc Natl Acad Sci U S A 112(3):633–640

Gevers D, Knight R, Petrosino JF, Huang K, McGuire AL, Birren BW, Nelson KE, White O, Methé BA, Huttenhower C (2012) The Human Microbiome Project: a community resource for the healthy human microbiome. PLoS Biol 10(8):e1001377

Gower JC (1966) Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53(3–4):325–338

He Z (2014) Microarrays: current technology, innovations and applications. Caister Academic Press, Norfolk

He Z, Deng Y, Zhou J (2012a) Development of functional gene microarrays for microbial community analysis. Curr Opin Biotechnol 23(1):49–55

He Z, Van Nostrand JD, Zhou J (2012b) Applications of functional gene microarrays for profiling microbial communities. Curr Opin Biotechnol 23(3):460–466

Hong C, Manimaran S, Shen Y, Perez-Rogers JF, Byrd AL, Castro-Nallar E, Crandall KA, Johnson WE (2014) PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples. Microbiome 2(33):2049–2618

Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M (2016) KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res 44(D1):D457–D462

Kaper JB, Nataro JP, Mobley HLT (2004) Pathogenic *Escherichia coli*. Nat Rev Micro 2(2):123–140

Karlsson E, Macellaro A, Bystrom M, Forsman M, Frangoulidis D, Janse I, Larsson P, Lindgren P, Ohrman C, van Rotterdam B, Sjodin A, Myrtennas K (2014) Eight new genomes and synthetic controls increase the accessibility of rapid melt-MAMA SNP typing of *Coxiella burnetii*. PLoS One 9(1):e85417

Kau AL, Ahern PP, Griffin NW, Goodman AL, Gordon JI (2011) Human nutrition, the gut microbiome and the immune system. Nature 474(7351):327–336

Ley RE (2010) Obesity and the human microbiome. Curr Opin Gastroen 26(1):5–11

Li X, He Z, Zhou J (2005) Selection of optimal oligonucleotide probes for microarrays using multiple criteria, global alignment and parameter estimation. Nucleic Acids Res 33(19):6114–6123

Li Y, He J, He Z, Zhou Y, Yuan M, Xu X, Sun F, Liu C, Li J, Xie W (2014) Phylogenetic and functional gene structure shifts of the oral microbiomes in periodontitis patients. ISME J 8(9):1879–1891

Liang Y, He Z, Wu L, Deng Y, Li G, Zhou J (2010) Development of a common oligonucleotide reference standard for microarray data normalization and comparison across different microbial communities. Appl Environ Microbiol 76(4):1088–1094

Luo Y, Hui D, Zhang D (2006) Elevated CO_2 stimulates net accumulations of carbon and nitrogen in land ecosystems: a meta-analysis. Ecology 87(1):53–63

McCune B, Grace JB, Urban DL (2002) Analysis of ecological communities, vol 28. MjM software design, Gleneden Beach

Peterson J, Garges S, Giovanni M, McInnes P, Wang L, Schloss JA, Bonazzi V, McEwen JE, Wetterstrand KA, Deal C (2009) The NIH human microbiome project. Genome Res 19(12):2317–2323

Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, Liang S, Zhang W, Guan Y, Shen D (2012) A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature 490(7418):55–60

Qin N, Yang F, Li A, Prifti E, Chen Y, Shao L, Guo J, Le Chatelier E, Yao J, Wu L (2014) Alterations of the human gut microbiome in liver cirrhosis. Nature 513(7516):59–64

Rajilić-Stojanović M, Heilig HG, Molenaar D, Kajander K, Surakka A, Smidt H, De Vos WM (2009) Development and application of the human intestinal tract chip, a phylogenetic microarray: analysis of universally conserved phylotypes in the abundant microbiota of young and elderly adults. Environ Microbiol 11(7):1736–1751

Rho M, Tang H, Ye Y (2010) FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Res 38(20):e191–e191

Smith MI, Yatsunenko T, Manary MJ, Trehan I, Mkakosya R, Cheng J, Kau AL, Rich SS, Concannon P, Mychaleckyj JC, Liu J, Houpt E, Li JV, Holmes E, Nicholson J, Knights D, Ursell LK, Knight R, Gordon JI (2013) Gut microbiomes of Malawian twin pairs discordant for kwashiorkor. Science 339(6119):548–554

Sommer MO, Dantas G, Church GM (2009) Functional characterization of the antibiotic resistance reservoir in the human microflora. Science 325(5944):1128–1131

Soucy SM, Huang J, Gogarten JP (2015) Horizontal gene transfer: building the web of life. Nat Rev Genet 16(8):472–482

The HMP Consortium (2012) Structure, function and diversity of the healthy human microbiome. Nature 486(7402):207–214

Tottey W, Denonfoux J, Jaziri F, Parisot N, Missaoui M, Hill D, Borrel G, Peyretaillade E, Alric M, Harris HM (2013) The human gut chip "HuGChip", an explorative phylogenetic microarray for determining gut microbiome diversity at family level. PLoS One 8(5):e62544

Tu Q, He Z, Deng Y, Zhou J (2013) Strain/species-specific probe design for microbial identification microarrays. Appl Environ Microbiol 79(16):5085–5088

Tu Q, He Z, Li Y, Chen Y, Deng Y, Lin L, Hemme CL, Yuan T, Van Nostrand JD, Wu L (2014a) Development of HuMiChip for functional profiling of human microbiomes. PLoS One 9(3):e90546

Tu Q, He Z, Zhou J (2014b) Strain/species identification in metagenomes using genome-specific markers. Nucleic Acids Res 42(8):e67

Tu Q, Yu H, He Z, Deng Y, Wu L, Van Nostrand JD, Zhou A, Voordeckers J, Lee YJ, Qin Y (2014c) GeoChip 4: a functional gene-array-based high-throughput environmental technology for microbial community analysis. Mol Ecol Resour 14(5):914–928

Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP (2009) A core gut microbiome in obese and lean twins. Nature 457(7228):480–484

Vos M, Hesselman MC, Te Beek TA, van Passel MW, Eyre-Walker A (2015) Rates of lateral gene transfer in prokaryotes: high but why? Trends Microbiol 23(10):598–605

Wu L, Liu X, Schadt CW, Zhou J (2006) Microarray-based analysis of subnanogram quantities of microbial community DNAs by using whole-community genome amplification. Appl Environ Microbiol 72(7):4931–4941

Zhou J, He Z, Yang Y, Deng Y, Tringe SG, Alvarez-Cohen L (2015) High-throughput metagenomic technologies for complex microbial community analysis: open and closed formats. MBio 6(1):e02288–e02214