



(51) International Patent Classification:

C12Q 1/10 (2006.01) *A61K 35/74* (2015.01)
C12Q 1/68 (2006.01) *A61P 1/16* (2006.01)

(21) International Application Number:

PCT/CN2014/082181

(22) International Filing Date:

15 July 2014 (15.07.2014)

(25) Filing Language:

English

(26) Publication Language:

English

(71) Applicant: **ZHEJIANG UNIVERSITY** [CN/CN]; QIN Nan, 38 Zheda Road, Xihu, Hangzhou, Zhejiang 310027 (CN).

(72) Inventors: **LI, Lanjuan**; QIN Nan, 38 Zheda Road, Xihu district, Hangzhou, Zhejiang 310027 (CN). **QIN, Nan**; QIN Nan, 38 Zheda Road, Xihu, Hangzhou, Zhejiang 310027 (CN).

(74) Agent: **HANGZHOU QIUSHI PATENT OFFICE CO., LTD**; LIN Songhai/Huanghongnian Science and Technology Complex Building/Room 215, 147 Yugu Road, Xihu District, Hangzhou, Zhejiang 310013 (CN).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— with international search report (Art. 21(3))

(54) Title: BIOMARKER FOR LIVER CIRRHOSIS AND USAGES THEREOF

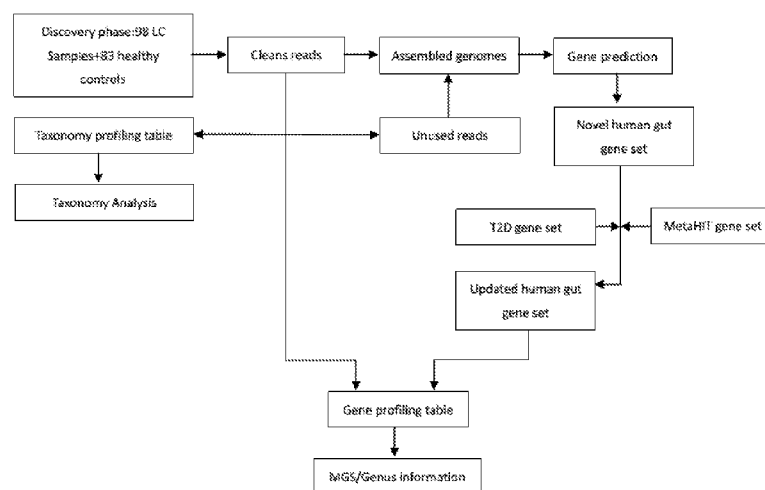


Figure 1

(57) Abstract: The faecal microbial communities and their functional composition are characterized by a whole gut microbiome-wild association study of stool samples from 98 liver cirrhosis patients and 83 healthy controls. *Veillonella* atypical ACS-134-V-Col7a is enriched in liver cirrhosis patients while *Bacteroides uniformis* ATCC 8492 and *Clostridiales* are enriched in controls. A biomarker for liver cirrhosis in human comprises *Veillonella* atypical ACS-134-V-Col7a and at least one bacterial strain over-represented in gut microbiota of healthy subject as compared to liver cirrhosis affected subjects, which comprises *Bacteroides uniformis* ATCC 8492 and *Clostridiales*. *Veillonella* atypical ACS-134-V-Col7a, *Bacteroides uniformis* ATCC 8492 and *Clostridiales* would be used for treating/preventing or diagnosing liver cirrhosis.

BIOMARKER FOR LIVER CIRRHOSIS AND USAGES THEREOF

FIELD OF THE INVENTION

The present invention relates to the field of biomedicine and biotech, specifically related to biomarker for liver cirrhosis and its applications.

BACKGROUND OF THE INVENTION

Liver cirrhosis (LC) is an advanced liver disease resulting from acute or chronic liver injury of any origin, including alcohol abuse, obesity and hepatitis virus infection. The prognosis for patients with decompensated liver cirrhosis is poor, and they frequently require liver transplantation¹. The liver interacts directly with the gut through the hepatic portal and bile secretion² systems. Enteric dysbiosis, especially the translocation of bacteria³ and their products^{4,5} across the gut epithelial barrier, is involved in the progression of liver cirrhosis. However, the phylogenetic and functional composition changes in the human gut microbiota that are related to this progression remain obscure⁵. Although some studies have revealed that alterations in the gut microbiota play an important role in complications of end-stage liver cirrhosis⁶ (such as spontaneous bacterial peritonitis⁷ and hepatic encephalopathy⁸) and the induction and promotion of liver damage in early-stage liver disease⁹ (such as alcoholic liver disease¹⁰ and non-alcoholic fatty liver disease¹¹), definitive associations between alterations in gut microbiota and liver pathology in humans are still lacking¹². Studies of liver cirrhosis patients¹³ and of mouse models for alcoholic liver disease¹⁰ have revealed a similar and substantial alteration in the gut microbiota, as measured by sequencing of 16S rRNA genes. How these phylogenetic alterations relate to changes in the functioning of the gut microbiota is unclear.

The role of gut microbiota in human health and disease¹⁴ has received unprecedented attention over the past few years with the rapid development of next-generation sequencing technologies. Several complex chronic diseases, such as obesity¹⁵⁻¹⁸, inflammatory bowel disease^{19,20}, diabetes mellitus²¹, metabolic syndrome²², symptomatic atherosclerosis²³ and non-alcoholic fatty liver disease¹⁰, have been associated to gut microbiota. A metagenomic study of 345 Chinese individuals with type-2 diabetes (T2D) identified 60,000 T2D-associated genes²⁴. The NIH Human Microbiome Project

(HMP) generated 3.5 Tb of metagenomic data from different anatomical sites among 242 healthy individuals and generated the largest human microbiome gene resource²⁵, which includes most of the genera, enzyme families and community configurations from the microbiota of healthy adults from Western countries²⁶. A quantitative metagenomics analysis^{27,28} of stool samples from Chinese liver cirrhosis patients and their healthy counterparts with the objective of improving the understanding of gut microbiota changes associated with liver cirrhosis was carried out.

SUMMARY

Our invention aims to provide additional knowledge of gut microbiota modifications in liver cirrhosis patients and to propose targeted biomarkers offering a non-invasive approach for early detection of the disease.

According to one embodiment of present disclosure, a biomarker for liver cirrhosis in a human comprising *Veillonella atypica* ACS-134-V-Col7a over-represented in gut microbiota of LC affected subjects as compared to healthy subjects and at least one bacterial strain over-represented in gut microbiota of healthy subject as compared to LC affected subjects.

Preferably, wherein the strain over-represented in gut microbiota of healthy subject as compared to LC affected subjects comprises *Bacteroides uniformis* ATCC 8492, and *Clostridiales*.

According to one embodiment of present disclosure, a method of treating/preventing liver cirrhosis in a human comprising administering to the human a therapeutically effective amount of a probiotic composition comprising one or more bacterial strains, wherein the composition

(i) stimulates growth and/or activity at least one of *Bacteroides uniformis* ATCC 8492, and *Clostridiales* which are over-represented in microbiota of healthy subjects as compared to LC affected subjects, and/or

(ii) inhibits growth and/or activity of *Veillonella atypica* ACS-134-V-Col7a which is over-represented in gut microbiota of LC affected subjects as compared to healthy subjects.

According to one embodiment of present disclosure, a probiotic composition, comprising one or more bacterial strains, wherein the composition

(i) stimulates growth and/or activity at least one of *Bacteroides uniformis* ATCC 8492, and *Clostridiales* which are over-represented in microbiota of healthy subjects as compared to LC

affected subjects, and/or

(ii) inhibits growth and/or activity of *Veillonella atypica ACS-134-V-Col7a* which is over-represented in gut microbiota of LC affected subjects as compared to healthy subjects.

According to one embodiment of present disclosure, a kit for diagnosing of liver cirrhosis comprising reagents for:

- (a) determining the amount of at least one bacterial strain of *Veillonella atypica ACS-134-V-Col7a*, *Bacteroides uniformis ATCC 8492*, and *Clostridiales*;
- (b) comparing the amount obtained in step (a) with a preset threshold.

According to one embodiment of present disclosure, a method for monitoring the efficacy of treatment of LC comprising the steps of:

- (a) determining the amount of at least one bacterial strain of *Veillonella atypica ACS-134-V-Col7a*, *Bacteroides uniformis ATCC 8492*, and *Clostridiales*;
- (b) comparing the amount obtained in step (a) with a preset threshold.

According to one embodiment of present disclosure, a method for identifying microbiota affected by LC disease in a human comprising;

- a) determining the metagenome of a diseased microbiota sample from the human;
- b) determining the metagenome of a control untreated/healthy sample;
- c) identifying under-represented and over-represented genes in the treated or diseased microbiota sample in relation to the control sample
- d) selecting one or more bacterial strains to compensate the expression of one or more genes identified in step c).

According to one embodiment of present disclosure, determining the metagenome of a diseased microbiota sample and a control untreated/healthy sample from the human; wherein determining the metagenome involves high- throughput sequencing;

Additional aspects and advantages of embodiments of present disclosure will be given in part in the following descriptions, become apparent in part from the following descriptions, or be learned from the practice of the embodiments of the present disclosure.

BRIEF DESCRIPTION OF THE FIGURES

Figure 1 illustrates diagram of the data analysis pipeline;

The study included a discovery and a validation phase. Volunteers for both phases were recruited in the same hospital. Both direct read mapping and *de novo* assembly were performed for each sample. Taxonomy profiling table was established for taxonomy analysis. A novel gut gene set was established, and annotated. Identification of the MGS, finding markers and validating markers is also shown.

Figure 2a, 2b illustrates taxonomic assignment of metagenomics species;

2a. MGS enriched in Chinese liver cirrhosis patients and healthy individuals. Species-level assignment was deduced from the best BlastN hits of genes from a given MGS at thresholds of the average of >95% identity and >90% overlap with genes from a sequenced genome. For MGS where these thresholds were not reached, an assignment was attributed at the lowest taxonomy level where at least 80% of the genes had the same best hit BlastP taxonomy; in all cases this criteria held true at higher taxonomic levels.

2b. taxonomic assignments of 58 species related to gut gene richness in a Danish cohort (Le Chatelier et al. 2013).

DETAILED DESCRIPTION OF THE INVENTION

These and other features, aspects, and advantages of the present invention will become better understood with regard to the following description, appended claims, and accompanying drawings. The following embodiments described by reference drawings are exemplary, which only used to explain the present invention, and not regarded as the limitations of the present invention.

Example 1 Construction of a liver cirrhosis gut microbial gene set and comparison with previous gene sets

All liver cirrhosis patients and healthy control adults were Han Chinese. In total, 123 liver cirrhosis patients and 114 healthy control adults were enrolled in our cohort. Our investigation included two phases. The first phase was a discovery phase in which 98 liver cirrhosis patients and 83 healthy controls were enrolled to characterize gut microbial compositional and functional changes between the two groups. The second phase was a validation phase, in which an additional 25 liver cirrhosis

patients and 31 controls were enrolled to validate the accuracy of the discovery phase findings. Total DNA was extracted from the faecal samples of 98 Chinese liver cirrhosis patients and 83 healthy Chinese controls and sequenced using the Illumina HiSeq 2000 (Illumina, San Diego, CA). This produced an average of 4.74 Gb (sd. ± 2.04 Gb) of high quality sequence for each sample, providing a total of 858 Gb of sequence data. The reads were assembled into contigs for all samples using the assembly software SOAPdenovo²⁹. Unassembled reads from 166 samples were pooled and the *de novo* assembly process was performed again for these reads (see Methods and Fig. 1). Finally, 61.68% of the total reads were used to generate 4.4 million contigs without ambiguous bases (minimum length of 500 bp). These contigs had a total length of 11.1 Gb, an average N50 length of 8,644 bp and ranged from 1,673 to 48,822bp.

To predict microbial genes for each of the 181 samples, we applied the methodology used in the MetaHIT human gut gene catalogue study³⁰. The MetaGene program predicted 13,371,697 open reading frames (ORFs) using a 100-bp cut-off for prediction. The total length of the predicted ORFs was 9,495,923,532 bp, representing 90.28% of the total length of the contigs. Among the ORFs, 1,047,885 (54.6%) were complete genes, while 869,808 (45.4%) were incomplete. A non-redundant “LC gene set” was established by removing redundant ORFs, defined as those sharing 95% identity over 90% of the shorter ORF length in pair-wise alignments. The final non-redundant liver cirrhosis gut gene set contained 2,688,468 ORFs, with an average length of 750bp and 42% of reads could be aligned to the gene catalogue.

We compared our LC gene set with the three other gut microbiota gene sets, MetaHIT³¹, HMP²⁵, and T2D²⁴. To facilitate this comparison, all genes were predicted from the original contigs using the same criteria. The MetaHIT catalogue contained 3,452,726 genes, HMP 4,768,112 genes, and T2D 2,148,029 genes. In total 674,131 genes were shared among all four catalogues. The LC, MetaHIT, HMP and T2D gene sets contained 794,647, 1,419,517, 2,620,096 and 623,570 unique genes, respectively.

Genes from the LC, T2D and MetaHIT gene catalogs were merged to create a non-redundant gene set for subsequent analyses. The HMP gut gene set was not included, as it contained Sanger, 454 or Illumina based 16S sequences, in addition to whole metagenomic data and it was generated from exclusively healthy individuals rather than from a disease cohort with accompanying healthy controls. The merged gene catalogue contained 5,382,817 genes, of which 797,690 were shared between all three catalogues. Of the genes in the LC gene set, 63.9% were also present in either one

or both of the remaining two, whereas 37.1% were unique. The MetaHIT and T2D sets contained 57.7% and 33.9% unique genes respectively. Large differences were also observed in the two gene sets derived from Chinese cohorts, the LC and T2D sets.

Example 2 Gut microbial species associated with liver cirrhosis

To investigate the relationship between the human gut metagenomes of healthy control individuals (n=83) and liver cirrhosis patients (n=96), we performed an association study for all genes from the merged gene set. Based on profiles of all 181 training samples, a Wilcoxon rank-sum test combined with Benjamini Hochberg adjustment for multiple testing was performed to identify differentially abundant genes. Using a very stringent significance threshold ($\text{fdr} < 0.0001$), significant differences were found for 75,245 genes between healthy and liver cirrhosis groups. Of these, 49,830 were more abundant in the liver cirrhosis patients and 25,415 in the healthy control group as determined using a rank-median test (**Methods**).

To explore and understand the microbial genes associated with liver cirrhosis we structured them into groups based on their abundance profiles²⁷. Briefly, genes of a given species are expected to have closely similar abundances in an individual. Species abundances vary greatly (12 to >1000-fold) across individuals in a cohort³¹, and so genes of a same species can be efficiently clustered by abundance co-variation. The resulting clusters are denoted metagenomic species (MGS) here. Similar approaches were previously used in other studies^{24,27,28,32}.

Of the 75,245 significantly different genes, a large majority (70%) clustered into 66 MGS (Table 1). It was shown previously that tracer genes are present only in cognate genomes and in all cognate genomes when more than one was sequenced, and can thus be used to measure abundance of a given MGS²⁷. The abundance of the 66 MGS was found to be significantly different in healthy and liver cirrhosis subjects. To confirm this finding, we further compared the abundance of the MGS in a validation cohort of 31 healthy and 21 liver cirrhosis subjects and found that 1 MGS of which were enriched in liver cirrhosis patients and 2 MGS in controls playing a important role in tracing found in the patients with liver cirrhosis.

Table 1. The list of LC-associated MGSs that could be assigned to previously known phylotypes

MGS	Genes	Taxonomic assignment	p
H_42	113	<i>Alistipes indistinctus</i> YIT 12060	7.43E-11
H_13	1304	<i>Bacteroides clarus</i> YIT 12056	1.49E-07
H_4	2970	<i>Bacteroides uniformis</i> ATCC 8492	1.79E-07
H_19	628	<i>Bilophila wadsworthia</i> 3_1_6	2.62E-07
H_32	283	<i>Coprococcus comes</i> SL71	1.65E-07
H_20	133	<i>Faecalibacterium cf. prausnitzii</i> KLE1255	2.00E-06
H_21	246	<i>Bacteroidales</i>	1.72E-07
H_24	258	<i>Bacteroidales</i>	3.49E-07
H_28	95	<i>Bacteroidales</i>	8.88E-07
H_11	429	<i>Clostridiales</i>	4.67E-08
H_12	375	<i>Clostridiales</i>	1.18E-07
H_15	330	<i>Clostridiales</i>	6.29E-08
H_2	623	<i>Clostridiales</i>	5.27E-07
H_3	561	<i>Clostridiales</i>	2.81E-08
H_38	140	<i>Clostridiales</i>	1.42E-07
H_40	51	<i>Clostridiales</i>	1.27E-08
H_7	2221	<i>Clostridiales</i>	9.74E-09
H_9	90	<i>Clostridiales</i>	5.89E-07
H_25	84	NA	2.87E-09
H_5	675	<i>Alistipes</i>	2.88E-08
H_8	174	<i>Clostridium</i>	5.23E-07
H_1	1052	<i>Dialister</i>	3.10E-09
H_43	143	<i>Eggerthella</i>	4.85E-09
H_10	2140	<i>Eubacterium</i>	1.46E-07
H_18	620	<i>Eubacterium</i>	7.90E-08
H_22	798	<i>Eubacterium</i>	6.54E-08
H_16	131	<i>Oscillibacter</i>	1.47E-07
H_36	216	<i>Parabacteroides</i>	2.82E-09
H_30	116	<i>Ruminococcaceae</i>	6.14E-11
H_33	759	<i>Ruminococcaceae</i>	1.15E-10
H_6	1888	<i>Subdoligranulum</i>	6.71E-08
H_14	112	<i>Lachnospiraceae</i>	2.22E-08
H_23	417	<i>Lachnospiraceae</i>	4.47E-08
H_34	72	<i>Lachnospiraceae</i>	9.27E-12
H_26	622	<i>Porphyromonadaceae</i>	8.63E-08
H_17	124	<i>Ruminococcaceae</i>	1.13E-05
H_29	201	<i>Ruminococcaceae</i>	3.07E-10
H_37	229	<i>Ruminococcaceae</i>	1.26E-07
L_40	115	<i>Lactobacillus</i>	3.30E-11
L_39	619	<i>Veillonella</i>	1.78E-08
L_55	119	<i>Veillonella atypica</i> KON	1.06E-16
L_59	157	<i>Veillonella sp. oral taxon 158 str. F0412</i>	2.17E-18
L_20	83	<i>Aggregatibacter segnis</i> ATCC 33393	1.22E-13
L_2	1489	<i>Bifidobacterium dentium</i> JCVIHMP022	1.17E-09
L_18	111	<i>Campylobacter sp. 10_1_50</i>	1.51E-19
L_13	859	<i>Clostridium perfringens</i> F262	1.80E-07

L_25	148	<i>Clostridium symbiosum</i> WAL-14673	4.18E-08
L_44	374	<i>Fusobacterium nucleatum</i> subsp. <i>animalis</i> F0419	1.75E-16
L_9	2751	<i>Haemophilus parainfluenzae</i> T3T1	1.38E-15
L_11	72	<i>Lactobacillus fermentum</i> ATCC 14931	6.81E-12
L_1	1242	<i>Lactobacillus mucosae</i> LM1	2.76E-12
L_3	1627	<i>Lactobacillus salivarius</i> CECT 5713	7.71E-12
L_10	1584	<i>Megasphaera micronuciformis</i> F0359	5.29E-18
L_8	1160	<i>Prevotella buccae</i> ATCC 33574	1.31E-12
L_5	74	<i>Ruminococcus gnavus</i> ATCC 29149	2.11E-07
L_12	1736	<i>Streptococcus anginosus</i> SK52 = DSM 20563	3.23E-16
L_32	125	<i>Streptococcus oralis</i> SK313	5.85E-14
L_15	2037	<i>Streptococcus parasanguinis</i> ATCC 903	5.41E-15
L_6	3307	<i>Streptococcus salivarius</i> JIM8777	3.34E-12
L_24	129	<i>Streptococcus</i> sp. 2 I 36FAA	3.54E-14
L_14	612	<i>Streptococcus vestibularis</i> F0396	5.90E-14
L_4	5755	<i>Veillonella atypica</i> ACS-134-V-Col7a	1.49E-19
L_19	914	<i>Veillonella dispar</i> ATCC 17748	4.79E-13
L_42	647	<i>Veillonella dispar</i> ATCC 17748	6.73E-18
L_7	2862	<i>Veillonella parvula</i> ATCC 17745	7.43E-19
L_17	678	<i>Veillonella</i> sp. oral taxon 158 str. F0412	5.35E-09

Only a small minority of the 38 MGS enriched in healthy individuals (15.8%) could be assigned species phylogenetic information by comparison with sequenced gut genomes using BlastN (95% identity and 90% overlap. Annotation to comparable taxonomic levels was observed for the 58 gut MGS analysed in the context of gene richness in a Danish cohort²⁷ (Fig.2b), reflecting a paucity of isolated and sequenced gut strains.

In sharp contrast with the MGS enriched in healthy subjects, an overwhelming majority of the MGS enriched in patients (24 of 28) could be assigned to a species. Such difference has a vanishingly low probability to be due to chance alone ($1.3e^{-21}$ by a χ^2 test, Fig. 2a) and indicates a highly modified gut microbial composition. The overall abundance of species enriched in patients reached very high levels, exceeding 5% in over a quarter of the patient group and approaching the extreme of 40%, whereas it was very low in healthy individuals, with a notable exception of 2 subjects.

Enrichment of 38 MGS in healthy people, at least two MGS has been verified, which in the following table (Table 2) as shown. 1 MGS of which were enriched in liver cirrhosis patients in the following table (Table 3) shows: *Veillonella atypical*, which can cause infections such as spontaneous bacterial peritonitis, liver abscesses, refractory respiratory tract infections, joint

infections acquired , endocarditis and bacteremia.

Table 2 MGS were enriched in controls

MGS	Taxonomic assignment
H_4	<i>Bacteroides uniformis</i> ATCC 8492
H_7	<i>Clostridiales</i>

Table 3 MGS was enriched in liver cirrhosis patients

MGS	Taxonomic assignment
L_4	<i>Veillonella atypica</i> ACS-134-V-Col7a

METHODS

Human faecal sample collection and DNA extraction

Each cirrhotic patient and healthy control subject provided a fresh stool sample that was delivered immediately from our hospital to the lab on ice bag using insulating polystyrene foam containers. In the lab it was divided into 5 aliquots of 200mg and immediately stored at -80°C. A frozen aliquot (200 mg) of each faecal sample was processed by phenol Trichloromethane DNA extraction method¹⁶ as previously described. DNA concentration was measured by nanodrop (Thermo Scientific) and its molecular size was estimated by agarose gel electrophoresis.

DNA library construction and sequencing

DNA libraries were constructed according to the manufacturer's instruction (Illumina). Same workflows from Illumina were used to perform cluster generation, template hybridization, isothermal amplification, linearization, blocking, denaturing and hybridization of the sequencing primers. Paired-end sequencing 2*100bp was performed for all libraries. The base-calling pipeline (*Casava 1.8.2 with parameters ---use-bases-mask y100n, I6n, Y100n, --mismatches 1, --adapter-sequence*) was used to process the raw fluorescent images and call sequences. The same insert size inferred by Agilent 2100 was used for all libraries (ranging from 275 to 450).

Quality control of reads

Reads that mapped to human genome together with their mated/paired reads were removed from each sample using BWA with parameters *-n 0.2*. Then quality control was preceded with following criteria: a) Reads containing more than 3 N bases were removed, b) Reads containing more than 50

bases with low quality (Q2) were removed, c) No more than 10 bases with low quality (Q2) or assigned as N in the tail of reads were trimmed. Sequences that lost their mated reads were considered as single reads and were used in the assembly procedure. Resulting filtered reads were considered for next step analysis.

***De novo* assembly of the Illumina short reads**

Considering that kmers with very low frequencies might arise from sequencing errors, they were not used in assembly by SOAPdenovo²⁹ (version1.05), which is based on *De brujin* graph construction. SOAPdenovo (version1.05) was used in Illumina short reads assembly with parameters *-d 1 -M 3*. Then we removed ambiguous bases from assembled scaffolds (this could divide one scaffold into multiple ones) and discarded scaffolds with length less than 500 bp. Finally we tested series of kmer values (from 31 to 59), then chose one with the longest N50 value for the remaining scaffolds. For each sample, we mapped clean data against scaffolds using SOAPalign version 2.21 with parameters *-u -2 -m 200*. Unused data from each sample were pooled and split into 4 parts (considering memory limit). Unused reads were repeatedly assembled with the same parameters but only one kmer value *-K 55* was chosen.

Gene prediction and non-redundant human gut gene set

MetaGeneMark³³ (prokaryotic GeneMark.hmm version 2.8) was used to predict ORFs in scaffolds without ambiguous bases. The non-redundant human gut gene set was built by pair-wise comparison of all the predicted ORFs using blat and the redundant ORFs were removed using a criterion of 95% identity over 90% of the shorter ORF length, which is consistent with the criterion used for the non-redundant European human gut gene set³¹ and T2D study²⁴. We checked the gaps and frames in the blat results. If there were gaps or the frames were different in the alignments result of two ORFs, the shorter one would not be removed as a redundancy. We used MetaGeneMark to predict genes in assembled contigs originally from MetaHIT and T2D study and merged these three gene sets into a single one with the above method.

Organism abundance profiling

SOAPalign 2.21 was used to align paired-end clean reads against reference genomes with parameters *-r 2 -m 200 -x 1000*. Reads with alignments on same reference genomes might be assigned into two types:

a) Unique reads (*U*): reads have alignments with only one genome; these reads were denoted as unique reads.

b) Multiple reads (M): reads have alignments with more than one genome, if these genomes come from one species; we denote these reads as unique reads. If they are from more than one species, we denote these reads as multiple reads.

For species S , if its abundance is $Ab(S)$, and it might have alignments with U unique reads and M multiple reads, computation is as follows.

$$Ab(S) = Ab(U) + Ab(M)$$

$$Ab(U) = U / l$$

$$Ab(M) = (\sum_{i=1}^M Co * \{M\}) / l$$

$Ab(U)$ and $Ab(M)$ are abundance of unique and multiple reads respectively, l is length of relative genome. For each multiple read, there is a species specific coefficient Co ; let us suppose one read in $\{M\}$ has alignments with N different species, then Co was calculated as follows.

$$Co = \frac{U}{\sum_{i=1}^N Ab(U)}$$

For these reads, we will add unique abundance of N species as denominator. Before we calculate abundance of species S , we had calculated $Ab(U)$ for all species as constants, if $Ab(U)$ of species S is 0, then Co will also be 0, and consecutively the abundance of species S is 0. Profiling table at genus level was generated by adding abundance of species into its genera. For some species that do not have a genus, they are denoted as unclassified genera.

Gene abundance profiling

Reads were aligned against the gene set by using SOAPalign³² with parameters “-r -m 200 -x 1000”. We counted gene’s abundance if both paired-end reads could be aligned on the same gene. If only one of the paired-end reads could be aligned on a gene, we aligned both reads against assembled contigs by checking if the previously not aligned read are in the non-translated region or not. If true, both reads will be validated for gene count, if not, then both reads were discarded.

When calculating abundance of genes, we used same strategy as for the organisms' abundance profiling. For a given gene G , its abundance is $Ab(G)$, and it might have alignments with U unique reads and M multiplereads, it goes as follows.

$$Ab(G) = Ab(U) + Ab(M)$$

$$Ab(U) = U / l$$

$$Ab(M) = (\sum_{i=1}^M Co * \{M\}) / l$$

$Ab(U)$ and $Ab(M)$ are abundance of unique and multiple reads respectively, l is length of gene G . For each multiple reads we calculate a specific coefficient Co for this gene, let us suppose one read with multiple $\{M\}$ alignments in N different genes, then Co was calculated as follows.

$$Co = \frac{U}{\sum_{i=1}^N Ab(U)}$$

For these reads, we will add unique abundance of N species as denominator.

Population stratification

Population stratification involved in our metagenomic data was corrected with modified EIGENSTART method shown as follows: firstly, singular value decomposition was carried to obtain axes of variation, where the number of significant axes was determined according to Tracy-Widom test at a significance level of $P < 0.05$; each axes was then replaced with the residuals of this axis from a regression to disease state; the corrected data was finally achieved by subtracting from original dataset the information associated with the residuals of each axis.

Gene count determination

Gene counts were computed essentially as described by Le Chatelier et al (2013). Briefly, data were downsized to adjust for sequencing depth and technical variability linked to different sequencing by randomly selecting 6.2 million of reads mapped to the merged gene catalog for each sample and then computing the mean number of genes over 30 random drawings. This was possible for all but 2 liver cirrhosis patients from the validation cohort (with not sufficient number of mapped reads), who were excluded from this analysis.

MetaGenomic Species (MGS)

We followed the approach described by Le Chatelier *et al.*, Nature 2013 to cluster genes from the current study into MetaGenomic Species (MGS). Briefly, in a first step pairwise Spearman correlation coefficient(ρ) of different genes was computed, using gene abundances across all individuals, and the genes correlated over a given threshold were clustered (a single-linkage clustering). To favour clustering specificity (that is, assigning only the genes of the same species to the same cluster) we used a rather high threshold ($\rho > 0.7$). To correct for the concomitant loss of sensitivity we carried out a second step, whereby the mean abundance signal of each cluster ≥ 50 genes was computed, using 50 most connected genes of a cluster, and the clusters that had a Spearman $\rho > .85$ were fused. This procedure was applied separately to the 49,830 genes more

abundant in liver cirrhosis patients and the 25,415 genes more abundant in healthy controls. 21,423 out of the 25,415 “healthy” genes fall into 43 clusters composed of 51 to 2702 genes after the first clustering step and 38 clusters of 51 to 2970 genes after the second step. 31,386 out of the 49,830 “liver cirrhosis” genes fall into 60 clusters of 51 to 3000 genes after the first clustering step and 28 clusters of 51 to 5755 genes after the second step.

To verify that the genes from a given cluster belong to the same genome and taxonomically annotate the MGS, we performed blastN and blastP analyses using a collection of 6,006 genomes (the available reference genomes from NCBI and the set of draft gastrointestinal genomes from the DACC of HMP and MetaHIT as of the August 3^d 2012 version). MGS were assigned to a given genome when >80% of its “tracer genes”²⁷ matched the same genome using BlastN, at a threshold of 95% identity over 90% of gene length. 6 “healthy” and 24 “liver cirrhosis” MGS could thus be assigned to the strain level (see Figure 2a and Table 1). The remaining MGS were annotated using blastP analysis and assigned to a given taxonomical level from genus to superkingdom level if >80% of its 50 tracer genes had the same level of assignment²⁷. All 36 remaining species but one could thus be assigned to a given genus, family or order. The quality of the clustering was thus validated by the homogenous annotation of its marker genes, which also held true for the whole MGS genes (data not shown). Abundance of the 66 MGS in each individual was computed using the 50 tracer genes.

To explore the origin of the species-level annotated MGS we constructed a reference catalogue, grouping 114 publicly available *Streptococcus* (57), *Fusobacterium* (26), *Lactobacillus* (16), *Veillonella* (12) and *Megasphaera* (3) genomes, mostly of oral (50) or gut (28) isolates. The 16 liver cirrhosis MGS that were assigned to the corresponding genera were compared to the genomes, using Blast and a score (T) was computed for each MGS, taking into account:

- (i) the proportion of genes above 95% identity & 90% coverage (Q);
- (ii) the average identity (R);
- (iii) the average coverage (S).
- (iv)

$$T=Q*R*S$$

A majority of the MGS enriched in liver cirrhosis patients (15/28) were of oral origin by this criterion whereas six were from gut or faeces, including a single species from ileum. To further

explore the origin of the LC enriched MGS we compared them by BlastN with the genes from three available ileum metagenomes³³ and failed to reveal identity beyond that found with sequenced genomes.

Although the present invention has been described in considerable detail with reference to certain preferred versions thereof, other versions are possible. Therefore, the spirit and scope of the appended claims should not be limited to the description of the preferred versions contained therein.

Reference

- 1 Fouts, D. E., Torralba, M., Nelson, K. E., Brenner, D. A. & Schnabl, B. Bacterial translocation and changes in the intestinal microbiome in mouse models of liver disease. *Journal of hepatology* **56**, 1283-1292, doi:10.1016/j.jhep.2012.01.019 (2012).
- 2 Cesaro, C. *et al.* Gut microbiota and probiotics in chronic liver diseases. *Digestive and liver disease : official journal of the Italian Society of Gastroenterology and the Italian Association for the Study of the Liver* **43**, 431-438, doi:10.1016/j.dld.2010.10.015 (2011).
- 3 Wiest, R. & Garcia-Tsao, G. Bacterial translocation (BT) in cirrhosis. *Hepatology* **41**, 422-433, doi:10.1002/hep.20632 (2005).
- 4 Nolan, J. P. The role of intestinal endotoxin in liver injury: a long and evolving history. *Hepatology* **52**, 1829-1835, doi:10.1002/hep.23917 (2010).
- 5 Gill, S. R. *et al.* Metagenomic analysis of the human distal gut microbiome. *Science* **312**, 1355-1359, doi:10.1126/science.1124234 (2006).
- 6 Garcia-Tsao, G. & Wiest, R. Gut microflora in the pathogenesis of the complications of cirrhosis. *Best practice & research. Clinical gastroenterology* **18**, 353-372, doi:10.1016/j.bpg.2003.10.005 (2004).
- 7 Wiest, R., Krag, A. & Gerbes, A. Spontaneous bacterial peritonitis: recent guidelines and beyond. *Gut* **61**, 297-310, doi:10.1136/gutjnl-2011-300779 (2012).
- 8 Bass, N. M. *et al.* Rifaximin treatment in hepatic encephalopathy. *The New England journal of medicine* **362**, 1071-1081, doi:10.1056/NEJMoa0907893 (2010).
- 9 Benten, D. & Wiest, R. Gut microbiome and intestinal barrier failure--the "Achilles heel" in hepatology? *Journal of hepatology* **56**, 1221-1223, doi:10.1016/j.jhep.2012.03.003 (2012).

- 10 Yan, A. W. *et al.* Enteric dysbiosis associated with a mouse model of alcoholic liver disease. *Hepatology* **53**, 96-105, doi:10.1002/hep.24018 (2011).
- 11 De Filippo, C. *et al.* Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 14691-14696, doi:10.1073/pnas.1005963107 (2010).
- 12 Cho, I. & Blaser, M. J. The human microbiome: at the interface of health and disease. *Nature reviews. Genetics* **13**, 260-270, doi:10.1038/nrg3182 (2012).
- 13 Chen, Y. *et al.* Characterization of fecal microbial communities in patients with liver cirrhosis. *Hepatology* **54**, 562-572, doi:10.1002/hep.24423 (2011).
- 14 Nelson, K. E. *et al.* A catalog of reference genomes from the human microbiome. *Science* **328**, 994-999, doi:10.1126/science.1183605 (2010).
- 15 Ley, R. E., Turnbaugh, P. J., Klein, S. & Gordon, J. I. Microbial ecology: human gut microbes associated with obesity. *Nature* **444**, 1022-1023, doi:10.1038/4441022a (2006).
- 16 Turnbaugh, P. J. *et al.* An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444**, 1027-1031, doi:10.1038/nature05414 (2006).
- 17 Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457**, 480-484, doi:10.1038/nature07540 (2009).
- 18 Ley, R. E. *et al.* Obesity alters gut microbial ecology. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 11070-11075, doi:10.1073/pnas.0504978102 (2005).
- 19 Lepage, P. *et al.* Twin study indicates loss of interaction between microbiota and mucosa of patients with ulcerative colitis. *Gastroenterology* **141**, 227-236, doi:10.1053/j.gastro.2011.04.011 (2011).
- 20 Garrett, W. S. *et al.* Enterobacteriaceae act in concert with the gut microbiota to induce spontaneous and maternally transmitted colitis. *Cell host & microbe* **8**, 292-300, doi:10.1016/j.chom.2010.08.004 (2010).
- 21 Wen, L. *et al.* Innate immunity and intestinal microbiota in the development of Type 1 diabetes. *Nature* **455**, 1109-1113, doi:10.1038/nature07336 (2008).
- 22 Vijay-Kumar, M. *et al.* Metabolic syndrome and altered gut microbiota in mice lacking Toll-like receptor 5. *Science* **328**, 228-231, doi:10.1126/science.1179721 (2010).

- 23 Karlsson, F. H. *et al.* Symptomatic atherosclerosis is associated with an altered gut metagenome. *Nature communications* **3**, 1245, doi:10.1038/ncomms2266 (2012).
- 24 Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**, 55-60, doi:10.1038/nature11450 (2012).
- 25 A framework for human microbiome research. *Nature* **486**, 215-221, doi:10.1038/nature11209 (2012).
- 26 Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207-214, doi:10.1038/nature11234 (2012).
- 27 Le Chatelier, E. *et al.* Richness of human gut microbiome correlates with metabolic markers. *Nature* **500**, 541-546, doi:10.1038/nature12506 (2013).
- 28 Cotillard, A. *et al.* Dietary intervention impact on gut microbial gene richness. *Nature* **500**, 585-588, doi:10.1038/nature12480 (2013).
- 29 Li, R. *et al.* De novo assembly of human genomes with massively parallel short read sequencing. *Genome research* **20**, 265-272, doi:10.1101/gr.097261.109 (2010).
- 30 Gautam, M., Chopra, K. B., Douglas, D. D., Stewart, R. A. & Kusne, S. Streptococcus salivarius bacteremia and spontaneous bacterial peritonitis in liver transplantation candidates. *Liver transplantation : official publication of the American Association for the Study of Liver Diseases and the International Liver Transplantation Society* **13**, 1582-1588, doi:10.1002/lt.21277 (2007).
- 31 Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59-65, doi:10.1038/nature08821 (2010).
- 32 Li, R. *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966-1967, doi:10.1093/bioinformatics/btp336 (2009)
- 33 Noguchi, H., Park, J. & Takagi, T. MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic acids research* **34**, 5623-5630, doi:10.1093/nar/gkl723 (2006).

CLAIMS

1. A biomarker for liver cirrhosis in a human comprising *Veillonella atypica* ACS-134-V-Col7a and at least one bacterial strain over-represented in gut microbiota of healthy subject as compared to LC affected subjects.
2. The biomarker composition of claim 1, wherein the at least one bacterial strain over-represented in gut microbiota of healthy subject as compared to LC affected subjects comprises *Bacteroides uniformis* ATCC 8492, and *Clostridiales*.
3. A method of treating/preventing liver cirrhosis in a human comprising administering to the human a therapeutically effective amount of a probiotic composition comprising one or more bacterial strains, wherein the composition
 - (i) stimulates growth and/or activity at least one of *Bacteroides uniformis* ATCC 8492, and *Clostridiales*, and/or
 - (ii) inhibits growth and/or activity of *Veillonella atypica* ACS-134-V-Col7a.
4. A probiotic composition, comprising one or more bacterial strains, wherein the composition
 - (i) stimulates growth and/or activity at least one of *Bacteroides uniformis* ATCC 8492, and *Clostridiales*, and/or
 - (ii) inhibits growth and/or activity of *Veillonella atypica* ACS-134-V-Col7a.
5. A kit for diagnosing of liver cirrhosis comprising reagents for:
 - (a) determining the amount of at least one bacterial strain of *Veillonella atypica* ACS-134-V-Col7a, *Bacteroides uniformis* ATCC 8492, and *Clostridiales*;
 - (b) comparing the amount obtained in step (a) with a preset threshold.
6. A method for monitoring the efficacy of treatment of LC comprising the steps of:
 - (a) determining the amount of at least one bacterial strain of *Veillonella atypica* ACS-134-V-Col7a, *Bacteroides uniformis* ATCC 8492, and *Clostridiales*;
 - (b) comparing the amount obtained in step (a) with a preset threshold.

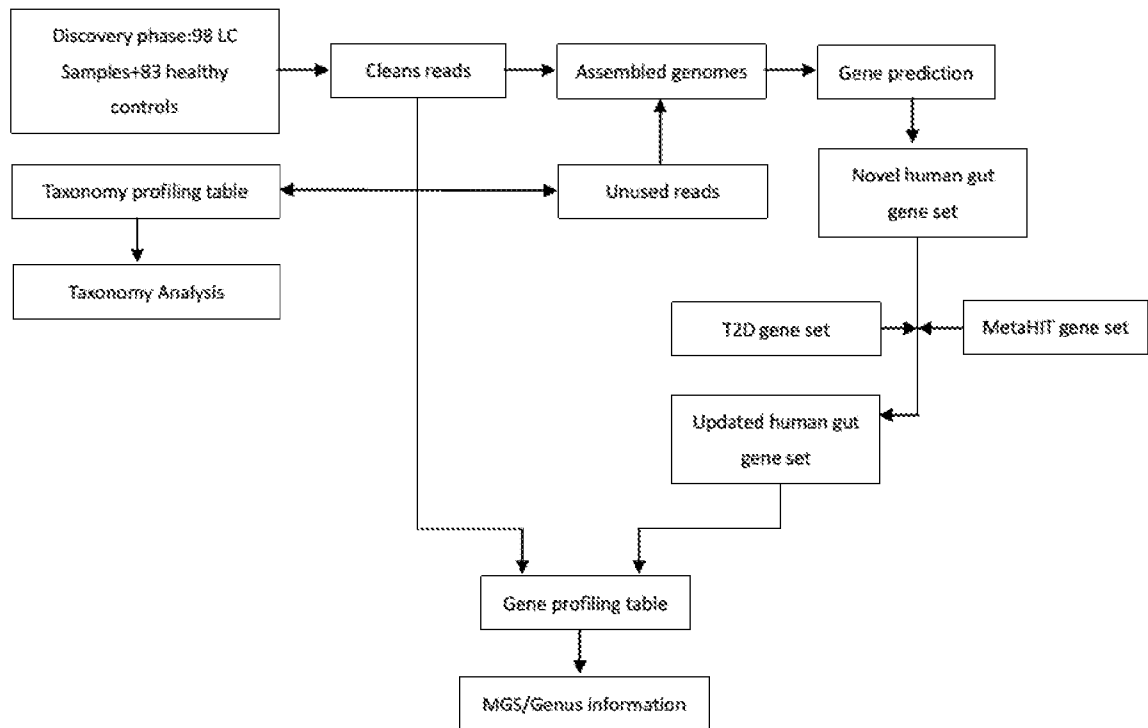


Figure 1

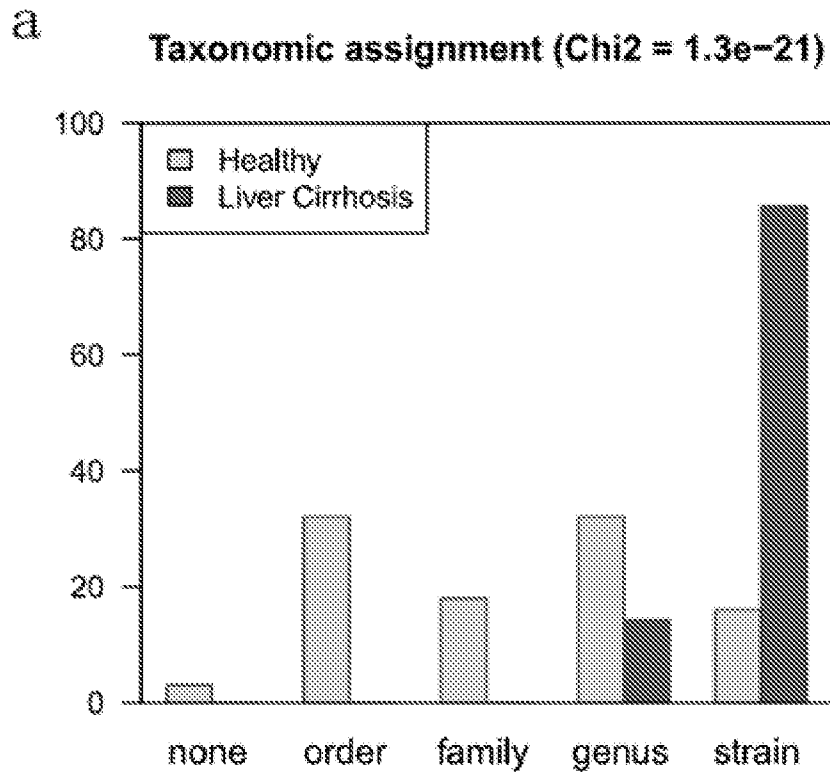


Figure 2a

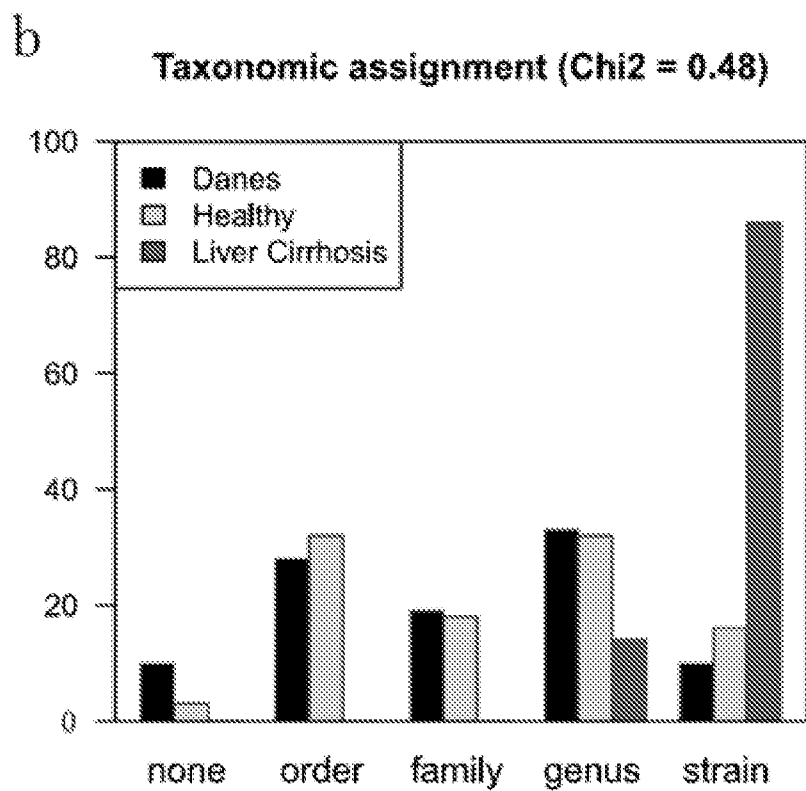


Figure 2b

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2014/082181

A. CLASSIFICATION OF SUBJECT MATTER

C12Q 1/10(2006.01)i; C12Q 1/68(2006.01)i; A61K 35/74(2015.01)i; A61P 1/16(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

C12Q1/-; A61K35/-; A61P1/-

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

WPI,EPODOC,CNPAT,CNKI,PubMed,ISI web of Knowledge,Elsevier Science Direct, SpringerLink: liver,hepatic,cirrhosis, hepatocirrhosis,Veillonella, atypica, V. atypica, ACS-134-V-Col7a, Bacteroid?? uniformis, B.uniformis, ATCC 8492, Clostridial+,clostridium, fusobacterium

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
E	CN 104195146 A (ZHEJIANG UNIVERSITY) 10 December 2014 (2014-12-10) claims 1-10	1-6
X	JP H04330016 A (RIKAGAKU KENKYUSHO) 18 November 1992 (1992-11-18) description, paragraphs [0006]-[0007], [0031], [0059]	3-4
X	CHEN, Y. et al. "Characterization of Fecal Microbial Communities in Patients with Liver Cirrhosis" <i>HEPATOLOGY</i> , Vol. vol.54, 31 December 2011 (2011-12-31), pages 562-572	3-6
X	WEI, X. et al. "Abnormal fecal microbiota community and functions in patients with hepatitis B liver cirrhosis as revealed by a metagenomic approach" <i>BMC Gastroenterology</i> , Vol. vol.13, 26 December 2013 (2013-12-26), ISSN: doi:10.1186/1471-230X-13-175, pages 1-8	3-6
A	US 2013288261 A1 (WALTER, J. ET AL) 31 October 2013 (2013-10-31) the whole document	1-6



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

29 March 2015

Date of mailing of the international search report

21 April 2015

Name and mailing address of the ISA/CN

STATE INTELLECTUAL PROPERTY OFFICE OF THE
P.R.CHINA(ISA/CN)
6,Xitucheng Rd., Jimen Bridge, Haidian District, Beijing
100088, China

Facsimile No. (86-10)62019451

Authorized officer

CHEN,Zhongwei

Telephone No. (86-10)62414108

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2014/082181

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☒ Claims Nos.: **3,6**
because they relate to subject matter not required to be searched by this Authority, namely:
 - [1] The subject matter of claims 3 and 6 relates to a method of treating/preventing liver cirrhosis or monitoring the efficacy of treatment of LC, and therefore does not warrant an international search according to the criteria set out in Rule 39.1(iv). However, the search has been carried out and based on the use of *Veillonella atypica* ACS-134-V-Col7a, *Bacteroides uniformis* ATCC 8492, or *Clostridiales* for manufacturing of a medicament.
2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:
3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2014/082181

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	104195146	A	10 December 2014	None	
JP	H04330016	A	18 November 1992	None	
US	2013288261	A1	31 October 2013	None	