

(51) International Patent Classification:
C12Q 1/68 (2006.01)(21) International Application Number:
PCT/EP2015/066218(22) International Filing Date:
15 July 2015 (15.07.2015)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
14306150.5 15 July 2014 (15.07.2014) EP

(71) Applicant: INSTITUT NATIONAL DE LA RECHERCHE AGRONOMIQUE [FR/FR]; 147, rue de l'Université, 75007 Paris (FR).

(72) Inventors: EHRLICH, Stanislav, Dusko; 37, rue Charles de Gaulle, 91400 Orsay (FR). LE CHATELIER, épouse JANNIERE, Emmanuelle; 29bis rue André Hodebourg, 78114 Magny les Hameaux (FR). PRIFTI, Edi; 85, rue Mirabeau, 92160 Antony (FR).

(74) Agent: REGIMBEAU; Regimbeau, 20, rue de Chazelles, 75847 Paris Cedex 17 (FR).

(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, 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, ST, 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).

Declarations under Rule 4.17:

— of inventorship (Rule 4.17(iv))

Published:

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

— before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))

— with sequence listing part of description (Rule 5.2(a))

(54) Title: GUT BACTERIAL SPECIES IN HEPATIC DISEASES

(57) **Abstract:** An in vitro method for the diagnosis of liver cirrhosis in a subject and/or for assessing whether a subject is at risk of developing liver cirrhosis, comprising the following steps: a) determining from a biological sample of said subject the abundance of each of the bacterial gene clusters of a set of bacterial gene clusters, wherein each of said clusters consisting of 50 non-redundant and covariant bacterial genes belonging to the same genome, defined in Tables 1, wherein said set of bacterial gene clusters comprise at least 2 bacterial gene clusters chosen in the group consisting in H_42, H_13, H_32, H_4, H_19, H_20, H_30, H_33, H_36, H_1, H_43, H_5, H_22, H_6, H_18, H_16, H_10, H_8, H_34, H_29, H_14, H_23, H_26, H_37, H_17, H_7, H_40, H_3, H_11, H_15, H_12, H_38, H_21, H_24, H_2, H_9, H_28, H_25, L_18, L_4, L_7, L_10, L_42, L_44, L_12, L_9, L_15, L_24, L_14, L_32, L_20, L_19, L_8, L_1, L_6, L_11, L_3, L_2, L_17, L_25, L_13, L_5, L_59, L_55, L_40, and L_39 from table 1. b) comparing the obtained abundances with at least one reference value, c) determining the diagnosis and/or risk of developing liver cirrhosis of said subject from said comparison.



GUT BACTERIAL SPECIES IN HEPATIC DISEASES

SUMMARY OF THE INVENTION

This invention is related to the field of metabolic disorders characterization. In particular, it relates to the use of specific bacteria, which are present in the gut of
5 human subjects and in their faeces, as a marker for the diagnosis of liver cirrhosis, as well as a predictive marker of a risk to develop the disease in the future. The invention thus pertains to noninvasive *in vitro* methods for diagnosing liver cirrhosis, and for assessing the risk of developing liver cirrhosis. The invention further provides methods for assessing the prognosis of liver cirrhosis.

10

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
15 frequently require liver transplantation. The liver interacts directly with the gut through the hepatic portal and bile secretion systems. Enteric dysbiosis, especially the translocation of bacteria and their products across the gut epithelial barrier, is involved in the progression of liver cirrhosis.

The gold standard for diagnosis of cirrhosis is liver biopsy, through a percutaneous,
20 transjugular, laparoscopic, or fine-needle approach.

Yet, liver biopsy can cause significant complications such as pneumothorax, bleeding, or puncture of the biliary tree. In addition, despite its specificity and sensitivity,

liver biopsy is nevertheless limited by inter-observer variation amongst pathologists, fibrosis staging systems and sampling errors.

In order to find alternative approaches to liver biopsy, or at least circumvent its application, physicians have looked for markers to estimate the liver fibrosis stages
5 and inflammation grades noninvasively.

Medical history, physical examination and blood tests can suggest liver cirrhosis. In that regards, the best predictors of cirrhosis are the presence of ascites (i.e., the accumulation of fluid in the peritoneal cavity), platelet count below 160,000/mm³, spider angiomas (swollen blood vessels found slightly beneath the skin surface).

10 However, those symptoms generally reveal liver fibrosis as large, rather than liver cirrhosis *per se*. Needle liver biopsy for diagnosis remains then compulsory to ascertain the diagnosis, in particular in cases of coexisting disorders (such as for example human immunodeficiency virus [HIV] and hepatitis C virus infection, or alcoholic liver disease and hepatitis C), or overlapping syndrome (such as for example
15 primary biliary cirrhosis with autoimmune hepatitis).

There is therefore a need to develop reliable non-invasive tests to efficiently identify patients suffering from liver cirrhosis or at risk of developing such disease.

In this perspective, the role of gut microbiota as an indicator of liver fibrosis, and more precisely of liver cirrhosis, has been investigated. A few studies have shown an
20 association between the gut microbiota and liver complications such as cirrhosis and other liver injuries. However, while enteric dysbiosis is involved in the progression of liver cirrhosis, the phylogenetic and functional composition changes in the human gut microbiota that are related to this progression remain obscure.

Hence, despite those discoveries, no biomarkers have emerged that can effectively be used to distinguish cirrhosis patients from healthy subjects, nor to give an indication of the prognosis of this disease.

The present invention provides specific markers associated with the gut microbiota, and in particular gut bacterial species, which can be used in noninvasive approaches for detection of liver cirrhosis, assessment of the risk to develop the disease, or even its prognosis over time.

The methods of the invention can further be used as a pre-screening, thus allowing physicians to narrow down the patients' population before definitive testing by liver biopsy. Various treatments with antibiotics or probiotics are being developed, in particular in case of complications (reviewed by Quigley et al 2013, in J. Hepatology 58,1020-27); the methods of the invention may also be used to monitor the effect of a given treatment on the gut microbiome.

FIGURE LEGEND

Figure 1. Differentially abundant metagenomic species between liver cirrhosis patients (n=123) and healthy individuals (n=114).

Abundance of 50 'tracer' genes for each species in the discovery ($n_{\text{patients}}=98$, $n_{\text{healthy}}=83$) and the validation cohort ($n_{\text{patients}}=25$, $n_{\text{healthy}}=31$) is shown. Genes are in rows, abundance is indicated by grayscale gradient (white, not detected; dark grey, most abundant); the enrichment significance is indicated (q indicates the Mann-Whitney test p-values corrected for multiple testing using Benjamini Hochberg method). Individuals, shown in columns are ordered by increasing total abundance of

patient-enriched species. Correlation of the species abundance and patients' clinical parameters in the discovery cohort are indicated by the shade of gray; intensity reflects the level of correlation; correlation is positive and negative for the LC-enriched and control-enriched species, respectively. MELD, CTP, TB, PT, INR, Crea & Alb stand respectively for **Model for End-stage Liver Disease**, **Child-Turcotte-Pugh score**, **Total Bilirubin**, **Prothrombin Time test**, **International Normalized Ratio** describing coagulation of the blood in liver cirrhosis patients, **Creatinine** and **Albumin** levels, respectively.

Figure 2. Differentially abundant metagenomic species between liver cirrhosis patients (n=123) and healthy individuals (n=114).

Patient clinical parameters for the lowest (LPA, n=24) and highest (HPA, n=24) patient-enriched species abundance. P-values indicate the significance of the difference by Mann-Whitney test except for MELD where Student test was used. MELD, CTP, TB, PT, INR, Crea & Alb stand respectively for **Model for End-stage Liver Disease**, **Child-Turcotte-Pugh score**, **Total Bilirubin**, **Prothrombin Time test**, **International Normalized Ratio** describing coagulation of the blood in liver cirrhosis patients, **Creatinine** and **Albumin** levels, respectively.

Figure 3. Massive changes of the gut microbiome in liver cirrhosis.

A: Gene counts are significantly reduced in the liver cirrhosis (LC) patients relative to healthy individuals. B: Most of the 38 species enriched in healthy individuals have no species-level taxonomic assignment; those that are originate from the gut. C: Species enriched in patients are largely assigned to a species level; they are most frequently of oral origin.

Figure 4. Massive changes of the gut microbiome in liver cirrhosis.

Abundance of patient-enriched species (n=28) in liver cirrhosis patients (n=98) and healthy individuals (n=83). The relative abundance of each patient-enriched species was computed as a sum of the abundances of all the genes assigned to it divided by the sum of the abundances of all gut microbial genes in each patient, which is equal to 1 in the normalized dataset. Bar length indicate the relative abundance of a given species. Shade of gray denotes different species. Individuals are ordered by the total patient-enriched species abundance; Low patient abundance (LPA) and High patient abundance (HPA) quartiles (n=24) are indicated by vertical lines.

Figure 5. A. Best prediction disease models N1 to N10 according to the discovery cohort. B. AUC using the best N6 model for the discovery cohort. c. AUC using the same model in the validation cohort.

Figure. 6. Profiles of the “severity MGS” in the discovery and validation cohorts.

Barcodes of the 21 severity MGS were performed using 50 marker genes and the 237 samples of the discovery (A) and validation B) cohorts. Samples are ordered according to cohorts, patient or healthy status and increasing CTP. Vertical bars separate controls and patients.

Figure 7. Predictions based on MELD severity score

A. Best prediction models N1 to N12 selected on the discovery cohort and tested on the validation one. B. Best N4 model trained on the discovery cohort. C. Performance of the same model on the validation cohort. MELD score is shown in function of the MGS based score.

Figure. 8. MELD scores in the cohort.

Vertical line denotes the threshold of high severity of the disease.

Figure. 9. Best prediction MELD high severity model

A. Best prediction models N1 to N10 trained on the discovery cohort to discriminate high severity patients (MELD>15) and tested on the validation cohort. B. Best N7 model trained on the discovery cohort C. Performance of the same model on the validation cohort. In gray are depicted 1000 AUC of the bootstrapped cohort of 90% of the cohort. The confidence interval is shown in the legend. The model is ("H_16", "H_22", "H_42", "SEV_4", "SEV_13", "SEV_15").

Figure. 10. Predictions based on CTP severity score

A. Best prediction models N1 to N12 trained on the discovery cohort and tested on the validation cohort. B. Best N7 model trained on the discovery cohort. C. Performance of the model on the discovery cohort. CTP score is shown as a function of the MGS based score.

Figure. 11. CTP scores in the cohort.

Vertical line denotes a clinically relevant threshold of the disease severity.

Figure. 12. Best prediction CTP severity models using discrete scores

A. Best prediction models N1 to N10 trained on the discovery cohort to discriminate patients with CTP >6 and tested on the validation cohort. B. Best N5 model trained on the discovery cohort C. Performance of the same model on the validation cohort. In gray are depicted 1000 AUC of the bootstrapped cohort of 90% of the cohort. The confidence interval is shown in the legend. The model is ("H_19", "SEV_24", "L_59", "H_42", "H_33").

DETAILED DESCRIPTION OF THE INVENTION

The inventors have found that the abundances of specific gut bacterial genes in faeces samples significantly correlate with the presence of liver cirrhosis in human subjects. They have found that among those bacterial genes, specific sets of genes varied in the same proportions among individuals. The inventors have determined 66
5 bacterial gene clusters based on this co-variance, which can be used to significantly discern subjects having liver cirrhosis, or likely to develop such disease, from healthy subjects.

It is the relative abundance of those bacterial gene clusters, rather than their presence or absence in the total gut microbial metagenome (and hence the faeces)
10 of a subject, which is the most relevant for establishing a diagnosis of liver cirrhosis or assessing whether a subject is at risk of developing liver cirrhosis. Hence, the applicants identified bacterial gene clusters (or metagenomic species) which are identified as most abundant in the healthy individuals of the cohort or in the patients with liver cirrhosis, respectively designated "H_..." and "L_...".

15 In addition, the Inventors have surprisingly determined a set of specific bacterial gene clusters, designated SEV_... clusters, which they found was significantly associated with the severity of the disease. Those specific bacterial gene clusters may therefore be used as biomarkers to estimate the severity, and thus the outcome of the disease. The method of the invention thus solely requires measuring the
20 abundance of specific bacterial genes directly in a faeces a sample, rather than on determining the presence/ absence of phylogenetically defined bacteria. It thus provides a simple and pragmatic approach to diagnosing liver cirrhosis. The bacterial genes of the invention have been identified from metagenomic shotgun sequences, according to techniques that are considered routine in the field of gut bacterial
25 metagenome analysis. Those techniques have notably been described in Liu et al.

(BMC genomics, 12(S2):S4, 2011), Arumugam et al. (Nature, 473(7346):174-80, 2011) or Qin et al. (Nature, 490(7418):55-60, 2012). Consequently, bacterial gene clusters grouping covariant genes (i.e., genes which relative abundances in the sample varied accordingly, with a significant correlation) have been determined by the inventors
5 according to a methodology described in the experimental part of the present application.

The bacterial gene clusters of the invention can thus advantageously be used instead of phylogenetically known bacterial genomes for convenience of analysis.

Moreover, the inventors have determined specific sets of 2, 3, 4, 5 or 6 bacterial
10 gene clusters with an increased correlation with the liver cirrhosis status of a subject, and which enable a particularly sensitive and specific diagnosis of liver cirrhosis, as evidenced by the AUC obtained with said combinations.

Based on the bacterial gene clusters identified, the inventors have additionally found that that it is possible to accurately diagnose whether a subject suffers from liver
15 cirrhosis or assess whether he is at risk of developing liver cirrhosis by measuring the abundance of a limited number specific bacterial gene clusters in a faeces sample. Indeed, they found that bacterial gene clusters corresponding to bacterial strains from the species *Streptococcus anginosus*, *Veillonella atypica*, *Veillonella dispar*, *Veillonella sp. oral taxon*, and *Haemophilus parainfluenzae*, are significantly more
20 abundant in liver cirrhosis subjects rather than in healthy subjects. This is particularly surprising as all of said bacteria are usually considered bacteria of oral origin, and were not expected to be found in overabundance in the gut of human subjects.

Therefore, the abundance of these gut bacterial strains of oral origin can be used to significantly discern subjects who are developing liver cirrhosis, or are likely to develop such a disease, from healthy subjects.

Thus, a first object of the invention is an *in vitro* method for the diagnosis of liver cirrhosis in a subject and /or for assessing whether a subject is at risk of developing liver cirrhosis, comprising the following steps:

- a) determining from a biological sample of said subject the abundance of each of the bacterial gene clusters of a set of bacterial gene clusters comprising at least 2 bacterial gene clusters chosen in the group consisting in H_42, H_13, H_32, H_4, H_19, H_20, H_30, H_33, H_36, H_1, H_43, H_5, H_22, H_6, H_18, H_16, H_10, H_8, H_34, H_29, H_14, H_23, H_26, H_37, H_17, H_7, H_40, H_3, H_11, H_15, H_12, H_38, H_21, H_24, H_2, H_9, H_28, H_25, L_18, L_4, L_7, L_10, L_42, L_44, L_12, L_9, L_15, L_24, L_14, L_32, L_20, L_19, L_8, L_1, L_6, L_11, L_3, L_2, L_17, L_25, L_13, L_5, L_59, L_55, L_40, and L_39 from table 1,
- b) comparing the obtained abundances with at least one reference value,
- c) determining the diagnosis and/or risk of developing liver cirrhosis of said subject from said comparison.

In one embodiment the present invention refer to an *in vitro* method for the diagnosis of liver cirrhosis in a subject and /or for assessing whether a subject is at risk of developing liver cirrhosis, comprising the following steps:

- a) determining from a biological sample of said subject the abundance of each of the bacterial gene clusters of a set of bacterial gene clusters,

wherein the said clusters consist of 50 non-redundant and covariant bacterial genes belonging to the same genome,

wherein the said set of bacterial gene clusters comprise at least 2 bacterial gene clusters chosen in the group consisting in H_42, H_13, H_32, H_4, H_19, H_20, H_30, H_33, H_36, H_1, H_43, H_5, H_22, H_6, H_18, H_16, H_10, H_8, H_34, H_29, H_14, H_23, H_26, H_37, H_17, H_7, H_40, H_3, H_11, H_15, H_12, H_38, H_21, H_24, H_2, H_9, H_28, H_25, L_18, L_4, L_7, L_10, L_42, L_44, L_12, L_9, L_15, L_24, L_14, L_32, L_20, L_19, L_8, L_1, L_6, L_11, L_3, L_2, L_17, L_25, L_13, L_5, L_59, L_55, L_40, and L_39 from table 1.

- b) comparing the obtained abundances with at least one reference value,
- c) determining the diagnosis and/or risk of developing liver cirrhosis of said subject from said comparison.

By “**subject**”, it is herein referred to a vertebrate, preferably a mammal, and most preferably a human.

By “**comprising**”, it is intended to mean that the invention encompasses the recited features, and may further comprise additional features.

By “**consisting in**”, it is intended to mean that the invention does only encompass the recited features and does not comprise additional features.

By “**biological sample**”, it is herein referred to any sample that may be isolated from a subject, including, without limitation, faeces, mucus, in particular colonic or oral mucus, sputum, or urine. In the context of the present invention, the biological fluid sample is preferably a faeces sample, also called stool sample.

By “**Liver cirrhosis**” it is referred to a medical condition resulting from of chronic liver disease and characterized by replacement of liver tissue by fibrosis, scar tissue and regenerative nodules (lumps that occur as a result of a process in which damaged tissue is regenerated) leading to loss of liver function.

5 As used hereafter, “**diagnosing**” a disease or a condition in a subject means to identify or to detect that the said subject is actually suffering from said disease or said condition. In particular, it is hereby contemplated that “**diagnosing liver cirrhosis**” in a subject means identifying or detecting that said subject has **liver cirrhosis**, as opposed to other liver diseases.

10 By “**the subject has a risk of developing a disease**”, it is hereby meant that the subject has more than 50%, preferably more than 60% and more preferably more than 75% of risk of suffering from said disease in the future.

As used herein, the term “**gene**” refers to genetic information coded into a nucleic acid molecule. It is composed of nucleic acid, preferably DNA, which may code for a
15 polypeptide or for an RNA chain of a given organism. More specifically, a gene is a region of a genome, which is associated with regulatory regions, transcribed regions, and /or other functional sequence regions within the said genome. The genes which are referred to in this invention are preferably “**bacterial genes**”, i.e., they correspond to a region of the genome of a bacterium.

20 By “**bacterial gene cluster**”, “**gene cluster**” or a “**cluster**”, it is herein referred to a set of bacterial genes. Preferably, the bacterial gene cluster comprises covariant bacterial genes, wherein the abundance of each of the genes varies in the same proportion compared with the abundance of the other genes in the same cluster among different individual samples. In other words, a bacterial gene cluster

according to the invention is a cluster of bacterial gene sequences which abundance levels in samples from distinct subjects are statistically linked rather than being randomly distributed. For simplicity purposes, the acronym MGS (standing for Meta Genomic Species), is also used herein. Genes of the microbiome can be ascribed to a bacterial gene cluster by several statistical methods known to the person skilled in the art. Preferably, a statistical method for testing covariance is used for testing whether two genes belong to the same cluster. To this end, the skilled person may use non-parametrical measures of statistical dependence, such as the Spearman's rank correlation coefficient for example. Most preferably, a bacterial gene cluster according to the invention comprises gut bacterial genes and that is determined by the method used in Le Chatelier et al. (*Nature*; 500:541-546; 2013) or in Cotillard et al. (*Nature*; 500: 585-588; 2013) for identifying metagenomic linkage groups.

In the context of the invention, the terms “**bacterial species**”, refer to a taxonomical reference defined by specific features and that can be used to classify bacteria which harbor such features. Thus, in the context of the invention, bacteria are said to belong to a bacterial species when they harbor genomic features, such as genomic sequences, corresponding to said bacterial species.

It is well known in the art that bacterial species comprise several strains, the genome of which may vary. Yet, within a bacterial population, each bacterial strain may be present in a specific proportion. As a result, the abundance of the genes related to each of said bacteria strain may differ, thus leading to the identification of several bacterial gene clusters which all relate to the same bacterial species. It is thus to be understood that different gene clusters according to the invention may be, in certain cases, assigned to the same bacterial species.

The person skilled in the art will immediately understand that, in the context of the invention, individual bacteria can be assigned to a bacterial strain and species based on the percentage of identity of their genes with the genes of bacterial gene clusters corresponding to said bacterial strain and species. Accordingly, in the context of the invention, bacterial gene clusters will be assigned to a bacterial strain according to the invention when it comprises genes that share at least 90 %, at least 95 %, at least 96 %, at least 97 %, at least 98 %, at least 99 %, or 100 % identity with a majority (>50%) of the genes of said cluster.

The assignment of the bacterial gene clusters of the invention to taxonomically defined bacterial strains and/or bacterial species is indicated in Table 1 (when applicable).

MGS	Assignment method*	Average identity	Average alignment	Assigned genes, %	Taxonomic level	Taxonomic assignment
H_42	Blast N	99,5	99,9	100	strain	Alistipes indistinctus YIT 12060
H_13	Blast N	99,4	99	100	strain	Bacteroides clarus YIT 12056
H_32	Blast N	99,3	96,1	94	strain	Coprococcus comes SL71
H_4	Blast N	99,5	98,6	100	strain	Bacteroides uniformis ATCC 8492
H_19	Blast N	99,6	99,2	100	strain	Bilophila

						wadsworthia 3_1_6
H_20	Blast N	98,6	99,7	100	strain	Faecalibacterium cf. prausnitzii KLE1255
H_30	Blast P	97	93,4	100	genus	Ruminococcaceae
H_33	Blast P	70,5	87,2	82	genus	Ruminococcaceae
H_36	Blast P	79,3	99,2	80	genus	Parabacteroides
H_1	Blast P	74,5	95,8	86	genus	Dialister
H_43	Blast P	62,9	87,9	84	genus	Eggerthella
H_5	Blast P	68,4	95,2	96	genus	Alistipes
H_22	Blast P	99,6	98,8	100	genus	Eubacterium
H_6	Blast P	75,4	94,8	84	genus	Subdoligranulum
H_18	Blast P	99,2	99,2	98	genus	Eubacterium
H_16	Blast P	76,1	94,7	86	genus	Oscillibacter
H_10	Blast P	69,1	94,7	86	genus	Eubacterium
H_8	Blast P	58,6	97,3	92	genus	Clostridium
H_34	Blast P	54,6	94,8	90	family	Lachnospiraceae
H_29	Blast P	63,6	96	84	family	Ruminococcaceae
H_14	Blast P	62,6	94	94	family	Lachnospiraceae
H_23	Blast P	57,9	88,7	84	family	Lachnospiraceae
H_26	Blast P	58,1	94,7	82	family	Porphyromonadac eae

H_37	Blast P	62,7	90,8	86	family	Ruminococcaceae
H_17	Blast P	55,7	95	86	family	Ruminococcaceae
H_7	Blast P	65,2	92,3	98	order	Clostridiales
H_40	Blast P	56,9	92,2	96	order	Clostridiales
H_3	Blast P	55,9	92,4	98	order	Clostridiales
H_11	Blast P	53,5	92,7	82	order	Clostridiales
H_15	Blast P	66,2	94,8	96	order	Clostridiales
H_12	Blast P	61,5	97,6	96	order	Clostridiales
H_38	Blast P	50,9	87,4	94	order	Clostridiales
H_21	Blast P	48,6	96,9	94	order	Bacteroidales
H_24	Blast P	62,6	96,1	98	order	Bacteroidales
H_2	Blast P	56,3	95	90	order	Clostridiales
H_9	Blast P	57,8	95,1	96	order	Clostridiales
H_28	Blast P	60,5	95,8	96	order	Bacteroidales
H_25	Blast P	NA	NA	NA	NA	NA
L_18	Blast N	98,2	99,7	100	strain	Campylobacter sp. 10_1_50
L_4	Blast N	98	99,8	100	strain	Veillonella atypica ACS-134- V-Col7a
L_7	Blast N	98,4	100	100	strain	Veillonella sp. 6_1_27
L_10	Blast N	98,7	99,9	100	strain	Megasphaera

						micronuciformis F0359
L_42	Blast N	90,9	96,4	100	strain	Veillonella dispar ATCC 17748
L_44	Blast N	98,9	65,1	100	strain	Fusobacterium nucleatum subsp. animalis F0419
L_12	Blast N	97,7	99,1	100	strain	Streptococcus anginosus SK52 = DSM 20563
L_9	Blast N	98,2	100	100	strain	Haemophilus parainfluenzae T3T1
L_15	Blast N	96,8	99,6	100	strain	Streptococcus parasanguinis ATCC 903
L_24	Blast N	97,7	99,4	100	strain	Streptococcus sp. 2_1_36FAA
L_14	Blast N	98,5	99,8	100	strain	Streptococcus vestibularis F0396
L_32	Blast N	97,5	99,7	94	strain	Streptococcus oralis SK313
L_20	Blast N	98,1	99,8	100	strain	Aggregatibacter segnis ATCC 33393

L_19	Blast N	98,1	99,9	100	strain	Veillonella dispar ATCC 17748
L_8	Blast N	99,2	100	100	strain	Prevotella buccae ATCC 33574
L_1	Blast N	98,3	100	100	strain	Lactobacillus mucosae LM1
L_6	Blast N	98,7	99,9	100	strain	Streptococcus salivarius SK126
L_11	Blast N	99,5	98,4	98	strain	Lactobacillus fermentum 28-3- CHN
L_3	Blast N	99,5	99,8	100	strain	Lactobacillus salivarius CECT 5713
L_2	Blast N	99,5	99,9	100	strain	Bifidobacterium dentium JCVIHMP022
L_17	Blast N	96,5	99,9	100	strain	Veillonella sp. oral taxon 158 str. F0412
L_25	Blast N	99,6	99,9	100	strain	Clostridium symbiosum WAL- 14673
L_13	Blast N	98,9	99,8	100	strain	Clostridium perfringens F262

L_5	Blast N	99,5	99,9	100	strain	Ruminococcus gnavus ATCC 29149
L_59	Blast P	94,4	98	98	genus	Veillonella
L_55	Blast P	93,1	98,2	96	genus	Veillonella
L_40	Blast P	72	94,1	88	genus	Lactobacillus
L_39	Blast P	89,6	99,6	96	genus	Veillonella

Table 1

- The performance of a diagnosis method or test is typically measured by the Area Under the Curve (AUC). Area Under the Curve, also called area under the ROC (Receiver Operating Characteristic) curve, is a measure of a classifier/test performance across all possible values of the thresholds. The higher the AUC, the better the performance of the test. A receiver operating characteristic (ROC), or simply ROC curve, is a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold is varied. It is created by plotting the fraction of true positives out of the positives (TPR = true positive rate) vs. the fraction of false positives out of the negatives (FPR = false positive rate), at various threshold settings. TPR is also known as sensitivity, and FPR is one minus the specificity or true negative rate. The sensitivity of a method is the proportion of actual positives which are correctly identified as such, and can be estimated by the area under the ROC (Receiver Operating Characteristic) curve, also called AUC.
- The use of multiple markers, and in particular as linear combinations thereof, would be expected by a person skilled in the art to increase the confidence of a given test.

Yet, the inventors have discovered that specific combinations of bacterial gene clusters correlate with a particularly important increase of the sensitivity of the test.

The inventors have thus determined the specific combinations of bacterial gene clusters that correlate with important AUCs. The more advantageous combinations of

- 5 2, 3, 4 and 5 bacterial gene clusters and their respective AUC are indicated in tables 2, 3, 4 and 5 respectively.

COMBINATION	AUC
L_7 and H_12	0,902
L_7 and L_25	0,899
L_7 and H_43	0,896
L_7 and H_3	0,894
L_19 and L_7	0,894
L_4 and L_25	0,893
L_7 and H_9	0,893
L_7 and H_34	0,892
L_4 and L_7	0,892
L_7 and L_9	0,891
L_7 and H_22	0,888
L_7 and H_5	0,888
L_7 and H_21	0,888
L_7 and H_17	0,887
L_7 and H_11	0,886

L_42	and	L_7	0,885
L_7	and	H_40	0,885
L_7	and	H_24	0,884
L_4	and	H_12	0,884
L_7	and	H_33	0,884
L_7	and	H_42	0,882
L_7	and	H_38	0,882
L_7	and	H_14	0,881
L_7	and	L_15	0,881
L_9	and	L_25	0,881
L_7	and	H_2	0,880
L_55	and	L_7	0,880
L_7	and	H_25	0,880
L_4	and	H_33	0,880
L_7	and	H_26	0,879
L_7	and	H_37	0,878
L_7	and	H_10	0,878
L_4	and	H_43	0,877
L_7	and	H_29	0,877
L_7	and	L_13	0,877
L_7	and	L_3	0,877
L_7	and	H_8	0,877

L_7	and	L_17	0,877
L_7	and	L_12	0,876
L_7	and	H_32	0,875
L_7	and	L_39	0,875
L_32	and	L_7	0,875
L_7	and	H_13	0,875
L_7	and	L_40	0,874
L_7	and	L_1	0,874
L_7	and	H_36	0,874
L_7	and	H_1	0,873
L_59	and	L_7	0,873
L_7	and	L_14	0,873
L_4	and	H_26	0,872
L_7	and	L_11	0,872
L_7	and	H_23	0,872
L_7	and	L_20	0,872
L_7	and	L_2	0,871
L_4	and	H_37	0,871
L_9	and	L_3	0,870
L_4	and	H_9	0,870

Table 2: advantageous combinations of 2 bacterial gene clusters and their AUC

Thus, preferably, the set of bacterial gene clusters of the invention comprises or consists in L_7 and H_12 ; L_7 and L_25 ; L_7 and H_43 ; L_7 and H_3 ; L_19 and L_7 ; L_4 and L_25 ; L_7 and H_9 ; L_7 and H_34 ; L_4 and L_7 ; L_7 and L_9 ; L_7 and H_22 ; L_7 and H_5 ; L_7 and H_21 ; L_7 and H_17 ; L_7 and H_11 ; L_42 and L_7 ;

5 L_7 and H_40 ; L_7 and H_24 ; L_4 and H_12 ; L_7 and H_33 ; L_7 and H_42 ; L_7 and H_38 ; L_7 and H_14 ; L_7 and L_15 ; L_9 and L_25 ; L_7 and H_2 ; L_55 and L_7 ; L_7 and H_25 ; L_4 and H_33 ; L_7 and H_26 ; L_7 and H_37 ; L_7 and H_10 ; L_4 and H_43 ; L_7 and H_29 ; L_7 and L_13 ; L_7 and L_3 ; L_7 and H_8 ; L_7 and L_17 ; L_7 and L_12 ; L_7 and H_32 ; L_7 and L_39 ; L_32 and L_7 ; L_7 and H_13 ; L_7 and L_40

10 ; L_7 and L_1 ; L_7 and H_36 ; L_7 and H_1 ; L_59 and L_7 ; L_7 and L_14 ; L_4 and H_26 ; L_7 and L_11 ; L_7 and H_23 ; L_7 and L_20 ; L_7 and L_2 ; L_4 and H_37 ; L_9 and L_3 or L_4 and H_9.

COMBINATION	AUC
L_19 and L_7 and H_12	0,921
L_4 and L_7 and L_25	0,920
L_7 and L_25 and H_12	0,920
L_4 and L_25 and H_12	0,918
L_19 and L_7 and L_25	0,918
L_7 and L_25 and H_43	0,916
L_7 and H_43 and H_12	0,916
L_19 and L_7 and H_43	0,916
L_4 and L_7 and H_12	0,915
L_7 and L_25 and H_3	0,915

L_42	and	L_7	and	L_25	0,915
L_42	and	L_7	and	H_12	0,915
L_19	and	L_7	and	H_3	0,914
L_4	and	L_7	and	H_43	0,914
L_7	and	L_25	and	H_37	0,913
L_7	and	L_15	and	H_12	0,913
L_7	and	L_25	and	H_9	0,913
L_55	and	L_7	and	H_12	0,912
L_7	and	L_25	and	H_24	0,912
L_4	and	L_25	and	H_43	0,912
L_7	and	L_25	and	H_34	0,912
L_4	and	L_7	and	H_17	0,911
L_7	and	L_25	and	L_3	0,911
L_19	and	L_7	and	H_34	0,911
L_7	and	H_38	and	H_12	0,911
L_4	and	L_7	and	H_1	0,911
L_4	and	L_17	and	L_25	0,911
L_7	and	H_21	and	H_12	0,910
L_42	and	L_7	and	H_43	0,910
L_4	and	L_25	and	H_37	0,910
L_19	and	L_7	and	H_9	0,910
L_4	and	H_33	and	H_12	0,910

L_4 and L_7 and H_11	0,910
L_7 and H_43 and H_9	0,910

Table 3: advantageous combinations of 3 bacterial gene clusters and their AUC

Yet preferably, the set of bacterial gene clusters according to the invention comprises or consists in L_19 and L_7 and H_12; L_4 and L_7 and L_25; L_7 and L_25 and H_12; L_4 and L_25 and H_12; L_19 and L_7 and L_25; L_7 and L_25 and H_43; L_7 and H_43 and H_12; L_19 and L_7 and H_43; L_4 and L_7 and H_12; L_7 and L_25 and H_3; L_42 and L_7 and L_25; L_42 and L_7 and H_12; L_19 and L_7 and H_3; L_4 and L_7 and H_43; L_7 and L_25 and H_37; L_7 and L_15 and H_12; L_7 and L_25 and H_9; L_55 and L_7 and H_12; L_7 and L_25 and H_24; L_4 and L_25 and H_43; L_7 and L_25 and H_34; L_4 and L_7 and H_17; L_7 and L_25 and L_3; L_19 and L_7 and H_34; L_7 and H_38 and H_12; L_4 and L_7 and H_1; L_4 and L_17 and L_25; L_7 and H_21 and H_12; L_42 and L_7 and H_43; L_4 and L_25 and H_37; L_19 and L_7 and H_9; L_4 and H_33 and H_12; L_4 and L_7 and H_11 or L_7 and H_43 and H_9.

COMBINATION	AUC
L_19 and L_7 and L_25 and H_12	0,934
L_4 and L_7 and L_25 and H_12	0,934
L_19 and L_7 and L_25 and H_43	0,934
L_4 and L_7 and L_25 and H_43	0,934
L_4 and L_17 and L_25 and H_12	0,932
L_42 and L_7 and L_25 and H_12	0,932
L_19 and L_7 and H_43 and H_12	0,931

L_42 and L_7 and L_25 and H_43	0,931
L_19 and L_7 and L_25 and H_3	0,930
L_19 and L_7 and L_25 and H_37	0,930
L_4 and L_7 and L_25 and H_17	0,930
L_4 and L_7 and L_25 and H_3	0,930
L_7 and L_25 and H_43 and H_12	0,930
L_4 and L_42 and L_7 and L_25	0,930

Table 4: advantageous combinations of 4 bacterial gene clusters and their AUC

More preferably, the set of bacterial gene clusters of the invention comprises or consists in L_19 and L_7 and L_25 and H_12; L_4 and L_7 and L_25 and H_12; L_19 and L_7 and L_25 and H_43; L_4 and L_7 and L_25 and H_43; L_4 and L_17 and L_25 and H_12; L_42 and L_7 and L_25 and H_12; L_19 and L_7 and H_43 and H_12; L_42 and L_7 and L_25 and H_43; L_19 and L_7 and L_25 and H_3; L_19 and L_7 and L_25 and H_37; L_4 and L_7 and L_25 and H_17; L_4 and L_7 and L_25 and H_3; L_7 and L_25 and H_43 and H_12 or L_4 and L_42 and L_7 and L_25.

COMBINATION	AUC
H_12 and H_43 and L_19 and L_25 and L_7	0,944
H_12 and H_37 and L_19 and L_25 and L_7	0,942
H_12 and L_17 and L_19 and L_25 and L_42	0,942
H_12 and L_19 and L_25 and L_42 and L_7	0,942
H_12 and H_43 and L_25 and L_4 and L_7	0,942

H_12 and H_43 and L_25 and L_42 and L_7	0,941
H_12 and L_17 and L_25 and L_4 and L_40	0,941
H_37 and H_43 and L_19 and L_25 and L_7	0,941
H_12 and H_37 and L_25 and L_42 and L_7	0,940
H_12 and L_25 and L_4 and L_42 and L_7	0,940
H_3 and H_43 and L_19 and L_25 and L_7	0,940
H_43 and L_19 and L_25 and L_42 and L_7	0,940
H_12 and L_17 and L_25 and L_4 and L_42	0,940
H_43 and L_25 and L_4 and L_42 and L_7	0,940
H_12 and H_38 and L_19 and L_25 and L_7	0,940

Table 5: advantageous combinations of 5 bacterial gene clusters and their AUC

Even more preferably, the set of bacterial gene clusters of the invention comprises or consists in H_12 and H_43 and L_19 and L_25 and L_7; H_12 and H_37 and L_19 and L_25 and L_7; H_12 and L_17 and L_19 and L_25 and L_42; H_12 and L_19 and L_25 and L_42 and L_7; H_12 and H_43 and L_25 and L_4 and L_7; H_12 and H_43 and L_25 and L_42 and L_7; H_12 and L_17 and L_25 and L_4 and L_40; H_37 and H_43 and L_19 and L_25 and L_7; H_12 and H_37 and L_25 and L_42 and L_7; H_12 and L_25 and L_4 and L_42 and L_7; H_3 and H_43 and L_19 and L_25 and L_7; H_43 and L_19 and L_25 and L_42 and L_7; H_12 and L_17 and L_25 and L_4 and L_42; H_43 and L_25 and L_4 and L_42 and L_7 or H_12 and H_38 and L_19 and L_25 and L_7.

Moreover, the combinations of 6 bacterial gene clusters corresponding to the bacterial clusters:

- i. H_12 and H_37 and L_17 and L_19 and L_2 and L_25 ,and

ii. H_12 and H_43 and L_17 and L_19 and L_25 and L_42;

have an AUC of 0,946 and 0,951 respectively.

Yet more preferably, the set of bacterial gene clusters of the invention comprises or consists in H_12 and H_37 and L_17 and L_19 and L_2 and L_25 or H_12 and H_43 and
5 L_17 and L_19 and L_25 and L_42.

Interestingly, the inventors have additionally found that that it is possible to accurately diagnose whether a subject suffers from liver cirrhosis or assess whether he is at risk of developing liver cirrhosis by measuring the abundance of a limited number of specific bacterial gene clusters in a faeces sample.

10 They found in particular that the bacterial gene clusters L_4, L_42, L_12, L_9, L_19 and L_17 were found to be significantly more abundant in liver cirrhosis subjects than in healthy subjects.

Those clusters have been taxonomically assigned to the described bacterial species *Streptococcus anginosus*, *Veillonella atypica*, *Veillonella dispar*, *Veillonella sp. oral*
15 *taxon*, and *Haemophilus parainfluenzae*.

Thus, preferably, the set of bacterial gene clusters according to the invention comprises or consists in bacterial clusters corresponding to the bacterial species *Streptococcus anginosus*, *Veillonella atypica*, *Veillonella dispar*, *Veillonella sp. oral*
taxon, and *Haemophilus parainfluenzae*.

20 Thus, in an embodiment, said set of bacterial gene clusters comprises or consists in L_4, L_42, L_12, L_9, L_19 and L_17

In an embodiment, the said set of bacterial gene clusters comprises or consists in the bacterial strains corresponding to the bacterial gene clusters L_4, L_12, L_9, L_19 and L_17.

In another embodiment, the said set of bacterial gene clusters comprises or consists
5 in L_4, L_42, L_12, L_9 and L_17.

In another embodiment, the said set of bacterial gene clusters comprises or consists in L_4, L_42, L_12, L_9, L_19 and L_17

The inventors have found that most of the bacterial gene clusters that were significantly overabundant in liver cirrhosis subjects could be assigned to bacterial
10 strains of oral origin. Those bacterial strains correspond to the bacterial gene clusters L_4, L_10, L_42, L_44, L_12, L_9, L_15, L_14, L_32, L_20, L_19, L_8, L_6, L_11, L_3, L_2, and L_17.

It is quite surprising to find an overabundance of these types of bacterial strains in the gut, and the more so as overabundance of oral bacteria in the gut microbiome
15 has never been correlated to liver cirrhosis. Those bacterial strains, and the corresponding bacterial gene clusters, can advantageously be used as markers of liver cirrhosis.

Thus, in an embodiment, said set of bacterial gene clusters comprises or consists in bacterial gene clusters chosen in the group consisting in L_4, L_10, L_42, L_44,
20 L_12, L_9, L_15, L_14, L_32, L_20, L_19, L_8, L_6, L_11, L_3, L_2, and L_17.

In another embodiment, the said set of bacterial gene clusters comprises or consists in L_4, L_10, L_42, L_44, L_12, L_9, L_15, L_14, L_32, L_20, L_19, L_8, L_6, L_11, L_3, L_2, and L_17.

Specific bacterial genes may be present in biological sample at a very low rate, thus making it difficult to assess their abundance. In that regard, the person skilled in the art will understand that trying to measure the abundance of particularly rare genes may lead to unreliable results due to technical limitations.

- 5 However, the inventors have found that, among the 66 bacterial gene clusters of the invention, 28 of them were overabundant in liver cirrhosis subjects, while 38 of them were actually underabundant in said subjects, compared to healthy controls.

The 28 bacterial gene clusters overabundant in liver cirrhosis subjects are: L_18, L_4, L_7, L_10, L_42, L_44, L_12, L_9, L_15, L_24, L_14, L_32, L_20, L_19, L_8, L_1, L_6,
10 L_11, L_3, L_2, L_17, L_25, L_13, L_5, L_59, L_55, L_40, and L_39.

It will be advantageous to use preferably those gene clusters as markers, in order to facilitate the implementation of the method of the invention, and further, increase the confidence of the test.

Thus, in an embodiment, said set of bacterial gene clusters comprises or consists in
15 bacterial gene clusters chosen in the group consisting in L_18, L_4, L_7, L_10, L_42, L_44, L_12, L_9, L_15, L_24, L_14, L_32, L_20, L_19, L_8, L_1, L_6, L_11, L_3, L_2, L_17, L_25, L_13, L_5, L_59, L_55, L_40, and L_39.

In another embodiment, the said set of bacterial gene clusters comprises or consists in L_18, L_4, L_7, L_10, L_42, L_44, L_12, L_9, L_15, L_24, L_14, L_32, L_20, L_19,
20 L_8, L_1, L_6, L_11, L_3, L_2, L_17, L_25, L_13, L_5, L_59, L_55, L_40, and L_39.

In another embodiment, the said set of bacterial gene clusters comprises or consists in L_18, L_4, L_7, L_10, L_42, L_44, L_12, L_9, L_15, L_24, L_14, L_32, L_20, L_19, L_8, L_1, L_6, L_11, L_3, L_2, L_17, L_25, L_13, L_5, L_59, L_55, L_40, and L_39.

Nevertheless, in order to obtain the more reliable result possible and limit the uncertainties due to the techniques of detection of the bacterial gene clusters, the person skilled in the art may use all of the 66 bacterial gene clusters.

In another embodiment, the set of bacterial gene clusters of the invention comprises
5 or consists in H_42, H_13, H_32, H_4, H_19, H_20, H_30, H_33, H_36, H_1, H_43, H_5, H_22, H_6, H_18, H_16, H_10, H_8, H_34, H_29, H_14, H_23, H_26, H_37, H_17, H_7, H_40, H_3, H_11, H_15, H_12, H_38, H_21, H_24, H_2, H_9, H_28, H_25, L_18, L_4, L_7, L_10, L_42, L_44, L_12, L_9, L_15, L_24, L_14, L_32, L_20, L_19, L_8, L_1, L_6, L_11, L_3, L_2, L_17, L_25, L_13, L_5, L_59, L_55, L_40, and L_39

10 Moreover, for each of these gut bacterial gene clusters, the inventors have determined a set of 50 non redundant genes that can be usefully used as tracer genes. By “**tracer gene**” or “**tracer**”, it is herein referred to those non-redundant and covariant genes from one bacterial gene cluster which are the most connected or inter-correlated. The use of non-redundant genes is particularly useful for
15 determining the abundance a gut bacterial gene cluster in a biological sample (Le Chatelier et al.; *Nature*; 500:541-546; 2013)

As a result, the method of the invention does not require either the phylogenetic identification of bacterial strains, or the sequencing of their entire genome. Hence, the method of the invention can be implemented using simple and cost effective
20 approaches such as PCR and PCR related technics. Further, bacterial strains can be rationally identified with a lesser risk of error.

Preferably, determining the abundance of a bacterial gene cluster from table 1 is performed by determining the number of copies of the genes of sequences SEQ ID No 1 to 3300, such as detailed in table 6

MGS	SEQ ID of non-redundant genes of interest.		
H_1	1	to	50
H_10	51	to	100
H_11	101	to	150
H_12	151	to	200
H_13	201	to	250
H_14	251	to	300
H_15	301	to	350
H_16	351	to	400
H_17	401	to	450
H_18	451	to	500
H_19	501	to	550
H_2	551	to	600
H_20	601	to	650
H_21	651	to	700
H_22	701	to	750
H_23	751	to	800
H_24	801	to	850
H_25	851	to	900
H_26	901	to	950
H_28	951	to	1000
H_29	1001	to	1050
H_3	1051	to	1100

H_30	1101	to	1150
H_32	1151	to	1200
H_33	1201	to	1250
H_34	1251	to	1300
H_36	1301	to	1350
H_37	1351	to	1400
H_38	1401	to	1450
H_4	1451	to	1500
H_40	1501	to	1550
H_42	1551	to	1600
H_43	1601	to	1650
H_5	1651	to	1700
H_6	1701	to	1750
H_7	1751	to	1800
H_8	1801	to	1850
H_9	1851	to	1900
L_1	1901	to	1950
L_10	1951	to	2000
L_11	2001	to	2050
L_12	2051	to	2100
L_13	2101	to	2150
L_14	2151	to	2200
L_15	2201	to	2250
L_17	2251	to	2300

L_18	2301	to	2350
L_19	2351	to	2400
L_2	2401	to	2450
L_20	2451	to	2500
L_24	2501	to	2550
L_25	2551	to	2600
L_3	2601	to	2650
L_32	2651	to	2700
L_39	2701	to	2750
L_4	2751	to	2800
L_40	2801	to	2850
L_42	2851	to	2900
L_44	2901	to	2950
L_5	2951	to	3000
L_55	3001	to	3050
L_59	3051	to	3100
L_6	3101	to	3150
L_7	3151	to	3200
L_8	3201	to	3250
L_9	3251	to	3300

Table 6

Since the methods of the invention rely on the determination of specific genes within a sample, it can advantageously be performed directly on DNA, and in particular on gut microbial DNA, extracted from said sample, for practicability purposes.

In an embodiment of the invention, the method further comprises a step of extracting bacterial DNA, in particular gut microbial DNA, from the biological sample. Hence, in an embodiment, the biological sample is a gut microbial DNA sample.

5 There are several ways to obtain samples of the said subject's gut microbial DNA (Sokol et al., *Inflamm. Bowel Dis.*, 14(6): 858-867, 2008). For example, it is possible to prepare mucosal specimens, or biopsies, obtained by coloscopy. However, coloscopy is an invasive procedure which is ill-defined in terms of collection procedure from study to study. Likewise, it is possible to obtain biopsies through
10 surgery. However, even more than coloscopy, surgery is an invasive procedure, which effects on the microbial population are not known. Preferred is the fecal analysis, a procedure which has been reliably been used in the art (Bullock et al., *Curr Issues Intest Microbiol.*; 5(2): 59-64, 2004; Manichanh et al., *Gut*, 55: 205-211, 2006; Bakir et al., *Int J Syst Evol Microbiol*, 56(5): 931-935, 2006; Manichanh et al., *Nucl. Acids Res.*, 36(16): 5180-5188, 2008; Sokol et al., *Inflamm. Bowel Dis.*, 14(6): 858-867, 2008). An example of this procedure is described in the Methods section of the Experimental Examples. Faeces contain about 10^{11} bacterial cells per gram (wet weight) and bacterial cells comprise about 50 % of fecal mass. The microbiota of the faeces represents primarily the microbiology of the distal large bowel. It is thus
20 possible to isolate and analyze large quantities of microbial DNA from the faeces of an individual. By "gut microbial DNA", it is herein understood the DNA from any of the resident bacterial communities of the human gut. The term "gut microbial DNA" encompasses both coding and non-coding sequences; it is in particular not restricted to complete genes, but also comprises fragments of coding sequences. Fecal analysis
25 is thus a non-invasive procedure, which yields consistent and directly-comparable results from patient to patient.

By “**the abundance of a bacterial gene clusters**”, it is herein referred to the abundance of the genes from said cluster in the tested sample, i.e by the gene abundance of genes from said cluster.

It will be immediately apparent to the person skilled in the art that the abundance of
5 a bacterial gene cluster, whether absolute or relative, can easily be determined by
quantifying the abundance of one or several of genes from said cluster in the sample.
Indeed, since the bacterial genes of the invention are grouped together within a
bacterial gene cluster because of their covariance, one would expect that the
relative abundance of each of the bacterial genes within a given cluster are similar.
10 Examples of methods that can be used to calculate the bacterial gene cluster
abundance in a given sample are detailed in the experimental part of the present
application. By “**gene abundance**”, it is herein referred to the absolute or relative
number of copies of said gene in the samples. “Absolute amount” (or “absolute
abundance”) of a gene designates the total number of copies of said gene in a define
15 volume of the tested sample, whereas “relative amount” (or “relative abundance”)
of a gene designates the total number of copies of said gene relative to the total
amount of genes or alternatively the total number of copies of said gene relative to
the amount of a single reference gene or preferably a combination of reference
genes present in the tested sample.
20 In an embodiment, determining the abundance of a bacterial gene cluster is
performed by determining the abundance of at least one bacterial gene from said
bacterial gene cluster.

Depending on the size of the sample and of the occurrence of the bacterial genes of
interest, certain bacterial genes may be difficult to detect in a sample. The skilled
25 person would thus easily conceive that, to increase the confidence of the results, it

is advantageous to determine the abundance of a bacterial gene cluster by determining the average abundance of several bacterial genes from said bacterial gene cluster of interest. Preferably, when several genes are used to determine the abundance of a bacterial gene cluster, the abundance of said bacterial gene cluster
5 corresponds to the average (i.e. the arithmetic mean) of the abundances of the bacterial genes tested.

In another embodiment, determining the abundance of a bacterial gene cluster is performed by determining the abundance of at least 5, at least 10, at least 20, at least 30, at least 40 or at least 50 bacterial genes from said bacterial gene cluster.

10 In the context of the invention, the inventors have defined, for each bacterial gene cluster of interest, as set of 50 sequences corresponding to non-redundant genes found within said cluster, which references are indicated in table 6.

In a preferred embodiment, determining the abundance of a bacterial gene cluster is performed by determining the number of copies of at least 5, at least 10, at least 20,
15 at least 30, at least 40 or at least 50 bacterial genes indicated in table 6 from said bacterial gene cluster.

Thus, determining the abundance of a bacterial gene cluster may be performed using any technique appropriate for quantifying nucleic acids sequences, which include *inter alia* hybridization with a labelled probe, PCR-based techniques, sequencing,
20 and all other methods known to the person of skills in the art. In a preferred embodiment, PCR-based techniques are used to determine the abundance of at least one bacterial gene. In a preferred embodiment, the abundance of the bacterial genes of the invention is determined by quantitative PCR (qPCR).

Representative methods for hybridization with a labelled probe include Northern blotting and *in situ* hybridization (Parker & Barnes, *Methods in Molecular Biology* 106:247-283 (1999)).

Examples of PCR-based techniques according to the invention include techniques
5 such as, but not limited to, quantitative PCR (Q-PCR), such as in particular quantitative polymerase chain reaction (qPCR) (Held et al., *Genome Research* 6:986-994 (1996)), and Rapid Quantitative PCR-Based methods (such as described in Noble et al., *Appl. Environ. Microbiol.*, 76(22): 7437-7443, 2010), reverse-transcriptase polymerase chain reaction (RT-PCR), quantitative reverse-transcriptase PCR (QRT-
10 PCR), rolling circle amplification (RCA) or digital PCR. PCR-based techniques further include for instance. Preferably, the PCR technique used quantitatively measures starting amounts of DNA, cDNA, or RNA. These techniques are well known and easily available technologies for those skilled in the art and do not need a precise description.

15 Sequencing methods include for instance sequencing by ligation, pyrosequencing, sequencing-by-synthesis, ion proton sequencing, nanopore sequencing or single-molecule sequencing. Sequencing also includes PCR-Based techniques, such as for example quantitative PCR or emulsion PCR. In the context of the invention, sequencing also includes PCR-Based techniques, such as for example quantitative PCR
20 or emulsion PCR. Optionally, prior to sequencing, DNA is fragmented, for example by restriction nuclease.

Sequencing is performed on the entire DNA contained in the biological sample, or on portions of the DNA contained in the biological sample. It will be immediately clear to the skilled person that the said sample contains at least a mixture of bacterial
25 DNA and of human DNA from the host subject. However, though the overall bacterial

DNA is likely to represent the major fraction of the total DNA present in the sample, each bacterial gene cluster may only represent a small fraction of the total DNA present in the sample.

To overcome this difficulty, the skilled person can use a method that allows the quantitative genotyping of sequences obtained from the biological sample with high precision. In one embodiment of this approach, the precision is achieved by analysis of a large number (for example, millions or billions) of polynucleotides. Furthermore, the precision can be enhanced by the use of massively parallel DNA sequencing, such as, but not limited to that performed by the Illumina Genome Analyzer platform (Bentley et al. Nature; 456: 53-59, 2008), the Roche 454 platform (Margulies et al. Nature; 437: 376-380, 2005), the ABI SOLiD platform (McKernan et al., Genome Res; 19: 1527-1541, 2009), the Helicos single molecule sequencing platform (Harris et al. Science; 320: 106-109, 2008), real-time sequencing using single polymerase molecules (Science; 323: 133-138, 2009), Ion Torrent sequencing (WO 2010/008480; Rothberg et al., Nature, 475: 348-352, 2011) and nanopore sequencing (Clarke J et al. Nat Nanotechnol.; 4: 265-270, 2009).

When the skilled person relies on sequencing methods to detect the abundance of certain bacterial genes, the information collected from sequencing is used to determine the number of copies of nucleic acid sequences of interest via bioinformatics procedures. For example, in an embodiment, the nucleic acid sequences of said bacterial gene cluster in the gut bacterial DNA sample are identified in the global sequencing data by comparison with the nucleic acid sequences SEQ ID No.1 described herein. Preferably, the nucleic acid sequences of said bacterial gene cluster in the gut bacterial DNA sample are identified in the global sequencing data by comparison with the nucleic acid sequences referred to in

Table 1. This comparison is advantageously based on the level of sequence identity with the sequences SEQ ID described herein.

Thus, a nucleic acid sequence displaying at least 90 %, at least 95 %, at least 96 %, at least 97 %, at least 98 %, at least 99 %, or 100 % identity with at least one of the
5 nucleic acid sequences SEQ ID No. 1 to 3700 is identified as a sequence comprised in one of the bacterial gene cluster of the invention.

A "percentage of sequence identity" is determined by comparing two optimally aligned sequences over a comparison window, where the portion of the peptide or polynucleotide sequence in the comparison window may comprise additions or
10 deletions (i.e., gaps) as compared to the reference sequence (which does not comprise additions or deletions) for optimal alignment of the two sequences. The percentage is calculated by determining the number of positions at which the identical amino-acid residue or nucleic acid base occurs in both sequences to yield the number of matched positions, dividing the number of matched positions by the
15 total number of positions in the window of comparison and multiplying the result by 100 to yield the percentage of sequence identity.

The definition of sequence identity given above is the definition that would use one of skill in the art. The definition by itself does not need the help of any algorithm, said algorithms being helpful only to achieve the optimal alignments of sequences,
20 rather than the calculation of sequence identity. From the definition given above, it follows that there is only one well defined value for the sequence identity between two compared sequences, which value corresponds to the value obtained for the best or optimal alignment.

The percentage of sequence identity between 2 amino acid sequences may be determined by using with default parameters the Blastp 2.0 program provided by the National Center for Biotechnology Information (<http://www.ncbi.nlm.nih.gov/blast/>; Tatusova et al., "Blast 2 sequences - a new tool for comparing protein and nucleotide
5 sequences", FEMS Microbiol. Lett. 174 : 247-250), which is habitually used by the inventors and in general by the skilled man for comparing and determining the identity between two sequences.

Identity between amino acid or nucleic acid sequences can be determined by comparing a position in each of the sequences which may be aligned for the purposes
10 of comparison. When a position in the compared sequences is occupied by the same amino acid or nucleotide, then the sequences are identical at that position. A degree of identity between amino acid sequences is a function of the number of identical amino acid sequences that are shared between these sequences. A degree of sequence identity between nucleic acids is a function of the number of identical
15 nucleotides at positions shared by these sequences.

To determine the percentage of identity between two amino acid sequences or two nucleic acid sequences, the sequences are aligned for optimal comparison. For example, gaps can be introduced in the sequence of a first amino acid sequence or a first nucleic acid sequence for optimal alignment with the second amino acid
20 sequence or second nucleic acid sequence. The amino acid residues or nucleotides at corresponding amino acid positions or nucleotide positions are then compared. When a position in the first sequence is occupied by the same amino acid residue or nucleotide as the corresponding position in the second sequence, the molecules are identical at that position.

The percentage of identity between the two sequences is a function of the number of identical positions shared by the sequences. Hence % identity = number of identical positions / total number of overlapping positions X 100.

In this comparison, the sequences can be of the same length or may be of different
5 lengths.

Optimal alignment of sequences may be conducted by a global homology alignment (i.e. an alignment of all amino acids or nucleotides of each sequence to be compared), such as by the global homology alignment algorithm of Needleman and Wunsch (1972), by computerized implementations of this algorithm or by visual
10 inspection. The best alignment (i.e., resulting in the highest percentage of identity between the compared sequences) generated by the various methods is selected.

In other words, the percentage of sequence identity is calculated by comparing two optimally aligned sequences, determining the number of positions at which the identical amino acid or nucleotide occurs in both sequences to yield the number of
15 matched positions, dividing the number of matched positions by the total number of positions and multiplying the result by 100 to yield the percentage of sequence identity.

Once the abundance of step a) has been obtained, it may be compared with at least one reference value.

20 By “**comparing**” the abundance of bacterial strains with a reference value, it is meant to compare either the abundance of each gene of the bacterial gene cluster separately with said reference value, or to compare the abundance of the bacterial gene cluster as herein defined with said reference value.

It is well known in the art that the different techniques available to detect and quantify nucleic acid molecules may have different limitations. When comparing the abundance of bacterial gene cluster with a reference value, the person skilled in the art may apply mathematical and statistical methods known in the art in order to
5 compensate for such limitations.

By "**reference value**" (or "**control value**"), it is herein referred to a specific value or dataset that can be used to identify patients associated with a specific outcome (e.g., liver cirrhosis). This reference or control value can be a predetermined value, or may correspond to a value obtained by determining the abundance of the
10 bacterial gene clusters of the invention in a reference sample.

Preferably, said reference or control value is obtained from samples from subject or pool of subjects having being diagnosed unambiguously as healthy. In the context of the invention, a healthy subject is a subject that does not suffer from liver cirrhosis, nor is at risk of developing liver cirrhosis.

15 The reference value according to the invention can for example be a single cut-off value, such as a median or mean.

The inventors have shown that some bacterial gene clusters are significantly less abundant in subjects having or at risk to develop liver cirrhosis than in healthy subjects, while, conversely, other bacterial gene clusters are significantly more
20 abundant in subjects having or at risk to develop liver cirrhosis rather than in healthy subjects.

In particular, the bacterial gene clusters H_42, H_13, H_32, H_4, H_19, H_20, H_30, H_33, H_36, H_1, H_43, H_5, H_22, H_6, H_18, H_16, H_10, H_8, H_34, H_29, H_14, H_23, H_26, H_37, H_17, H_7, H_40, H_3, H_11, H_15, H_12, H_38, H_21, H_24,

H_2, H_9, and H_28 are less abundant in subjects having or at risk to develop liver cirrhosis, rather than in healthy subjects.

On the contrary, the bacterial gene clusters L_18, L_4, L_7, L_10, L_42, L_44, L_12, L_9, L_15, L_24, L_14, L_32, L_20, L_19, L_8, L_1, L_6, L_11, L_3, L_2, L_17, L_25,
5 L_13, L_5, L_59, L_55, L_40, and L_39 are more abundant in subjects having or at risk to develop liver cirrhosis, rather than in healthy subjects.

After comparing the abundance of bacterial gene clusters in a faeces sample of a tested subject to a reference value, if said reference value is the abundance of the same bacterial gene cluster in faeces sample(s) of a healthy or several healthy
10 subjects, the person skilled in the art may conclude that the tested subject has or is at risk of developing liver cirrhosis when:

The ratio of the abundance of H_42 H_13, H_32, H_4, H_19, H_20, H_30, H_33, H_36, H_1, H_43, H_5, H_22, H_6, H_18, H_16, H_10, H_8, H_34, H_29, H_14, H_23, H_26, H_37, H_17, H_7, H_40, H_3, H_11, H_15, H_12, H_38, H_21, H_24, H_2, H_9, or
15 H_28 in a faeces sample of a tested subject on said reference value is inferior to 1;
or

The ratio of the abundance of L_18, L_4, L_7, L_10, L_42, L_44, L_12, L_9, L_15, L_24, L_14, L_32, L_20, L_19, L_8, L_1, L_6, L_11, L_3, L_2, L_17, L_25, L_13, L_5, L_59, L_55, L_40, or L_39 in a faeces sample of a tested subject on said reference
20 value is superior to 1.

The method of diagnosis of the invention enables the person skilled in the art a precise determination of the health status of the subject, enabling a specific treatment to be tailored to the needs of the patient. The prior determination of the diagnosis of liver cirrhosis with the method of the invention may thus be followed by

the indication and/or the administration of an appropriate treatment or of therapeutic measures.

Thus the present invention also relates to a method for designing a treatment for a subject, said method comprising:

- 5 a) determining that the subject is developing or at risk of developing liver cirrhosis with a method according to the invention, and
- b) designing a therapeutic treatment.

According to the invention, step b) of designing a therapeutic treatment may be followed by a step c) of administering said treatment.

- 10 Treatments for patients with cirrhosis are well known from the person skilled in the art, and include for instance the use of antibiotics, in particular of rifaximin. Therapeutic measures and treatments for patients with cirrhosis have been detailed in Garcia-Tsao et al.

- 15 The person skilled in the art may then determine the prognosis liver cirrhosis of the tested subject.

The inventors have further found that the relative abundance of specific bacterial strains could be used to evaluate the prognosis of subjects suffering from liver cirrhosis.

- 20 Indeed, the inventors found that the bacterial strains corresponding to the bacterial gene clusters L_4, L_10, L_42, L_44, L_12, L_9, L_15, L_14, L_32, L_20, L_19, L_8, L_6, L_11, L_3, L_2, and L_17, which all correspond to bacteria of oral origin, were particularly abundant in subject who had a poor prognosis, as defined by the Child-Pugh-Turcotte (CPT) and the Model of End-Stage Liver Disease (MELD) scores. More

precisely, they have found that the prognosis worsen with the increase in relative abundance of the sum of the bacterial gene clusters L_4, L_10, L_42, L_44, L_12, L_9, L_15, L_14, L_32, L_20, L_19, L_8, L_6, L_11, L_3, L_2, and L_17, compared with the total quantity of bacterial genes in the sample.

- 5 Those bacterial strains can then advantageously be used as markers for the evaluating the outcome of liver cirrhosis in a subject.

Therefore, another object of the invention is an *in vitro* method for the prognosis of liver cirrhosis for a subject, comprising the following steps:

- a) determining from a biological sample of said subject the abundance of the
10 bacterial strains L_4, L_10, L_42, L_44, L_12, L_9, L_15, L_14, L_32, L_20, L_19, L_8, L_6, L_11, L_3, L_2, and L_17,
b) comparing the obtained abundance with at least one reference value,
c) determining the prognosis liver cirrhosis of said subject from said comparison.

As used hereafter, “**prognosing**” a disease or a condition in a subject hereby means
15 to predict the likely outcome of one's current standing. In particular, a prognosis of a disease in a subject may include estimating survival rate of said subject, due to said disease within a given time frame. A prognosis of a disease in a subject may include estimating the risk to develop complications, in particular severe complications.

As used hereafter, “**prognosing liver cirrhosis**” in a subject means estimating
20 survival rate of said subject, due to liver cirrhosis within a given time frame, and/or estimating the risk to develop complications, in particular severe complications such as for instance portal hypertension, esophageal varices or gastric varices, hepatic encephalopathy, jaundice.

The Child-Pugh-Turcotte (CPT) and the Model of End-Stage Liver Disease (MELD) scores are commonly used in the medical field of liver disease, and can be briefly summarized as follows.

The Child-Pugh-Turcotte (CPT) score employs five clinical measures of liver disease, which are summarized in Table 7. Each measure is scored 1-3, with 3 indicating most severe derangement.

Measure	1 point	2 points	3 points
Total bilirubin, $\mu\text{mol/l}$ (mg/dl)	<34 (<2)	34-50 (2-3)	>50 (>3)
Serum albumin, g/dl	>3.5	2.8-3.5	<2.8
PT INR	<1.7	1.71-2.30	> 2.30
Ascites	None	Mild	Moderate to Severe
Hepatic encephalopathy	None	Grade I-II (or suppressed with medication)	Grade III-IV (or refractory)

Table 7

Chronic liver disease is classified into Child-Pugh class A to C, employing the added score from above. The survival rate of subjects depending in their Child-Pugh class is recalled in table 8.

Points	Class	One year survival	Two year survival
--------	-------	-------------------	-------------------

5-6	A	100%	85%
7-9	B	81%	57%
10-15	C	45%	35%

Table 8

The MELD score uses the patient's values for serum bilirubin, serum creatinine, and the international normalized ratio for prothrombin time (INR) to predict survival. It is calculated according to the following formula:

$$5 \quad MELD = 3.78[\ln \text{ serum bilirubin (mg/dL)}] + 11.2[\ln INR] + 9.57[\ln \text{ serum creatinine (mg/dL)}] + 6.43$$

When calculating the MELD score, the following applies:

- if the patient has been dialyzed twice within the last 7 days, then the value for serum creatinine used should be 4.0; and
- 10 – any value less than one is given a value of 1 (i.e. if bilirubin is 0.8, a value of 1.0 is used) to prevent the occurrence of scores below 0 (the natural logarithm of 1 is 0, and any value below 1 would yield a negative result).

The rate of mortality (within the 3 month following scoring) of a subject depending on its MELD score is as follows.

- 15 – 40 or more : 71.3% mortality
- 30-39 : 52.6% mortality
- 20-29 : 19.6% mortality

- 10-19 : 6.0% mortality
- <9 : 1.9% mortality

In addition, the Inventors have surprisingly determined a set of specific bacteria strains associated with bacterial gene clusters, which they found was significantly associated with the severity of the disease. Those specific bacterial strains may therefore be used as biomarkers to estimate the severity, and thus the outcome of the disease. The bacterial gene clusters associated with those bacterial strains have been named SEV_2, SEV_4, SEV_13, SEV_14, SEV_15, SEV_22, SEV_24, SEV_25.

Table 9 indicates taxonomic information relative to these bacterial gene clusters. BlastN against 6006 reference genomes (best hit threshold >95% alignment, >95% identity); If BlastN below threshold, BlastP against 6006 reference genomes (>= 80% of the tracer genes with the same best hit taxonomic annotation).

MGS	Assignment method*	Average identity	Assigned genes, %	Taxonomic level	Taxonomic assignment
SEV_13	blastP	99,2	98	genus	Ruminococcus
SEV_2	blastP	69,1	84	family	Acidaminococcaceae
SEV_22	blastP	64,6	100	order	Clostridiales
SEV_4	blastN	99,3	100	strain	Parabacteroides merdae ATCC 43184
SEV_14	blastP	88,4	90	genus	Leptotrichia
SEV_15	blastP	96,8	100	genus	Fusobacterium
SEV_24	blastP	93,1	90	genus	Veillonella
SEV_26	blastN	97,4	98	family	Veillonellaceae

Table 9

The inventors have further established the sequences of 50 non-redundant genes for each of the SEV, the SEQ ID are indicted below.

MGS	SEQ ID of non-redundant genes of interest.		
SEV_13	3301	to	3350
SEV_14	3351	to	3400
SEV_15	3401	to	3450
SEV_2	3451	to	3500
SEV_22	3501	to	3550
SEV_24	3551	to	3600
SEV_26	3601	to	3650
SEV_4	3651	to	3700

Table 10

The inventors have notably demonstrated that a particular set of bacterial strains, including the above SEV bacterial strains, significantly correlate with the severity of the disease, and can therefore be used as biomarkers to estimate the outcome of liver cirrhosis in patients. Those bacterial strains are the bacterial strains corresponding to the bacterial gene clusters H_16 , H_19 , H_20 , H_22 , H_33 , H_42 , L_18 , L_19 , L_39 , L_4 , L_42 , L_55 , L_59 , SEV_13 , SEV_2 , SEV_22 , SEV_4 , SEV_14 , SEV_15 , SEV_24 and SEV_26.

The invention also relates to an *in vitro* method for the diagnosis of liver cirrhosis in a subject and /or for assessing whether a subject is at risk of developing liver cirrhosis, comprising the following steps:

- a) determining from a biological sample of said subject the abundance of each of the bacterial gene clusters of a set of bacterial gene clusters, wherein each of said clusters consisting of 50 non-redundant and covariant bacterial genes belonging to the same genome, defined in tables 1 and in table 9,
- 5 wherein said set of bacterial gene clusters comprise or consist in H₁₆, H₂₂, and H₄₂ from table 1, and SEV₄, SEV₁₃ and SEV₁₅ from table 9,
- b) comparing the obtained abundances with at least one reference value,
- c) determining the diagnosis and/or risk of developing liver cirrhosis of said subject from said comparison.

10 The invention also pertains to an *in vitro* method for the prognosis of liver cirrhosis for a subject, comprising the following steps:

- a) determining from a biological sample of said subject the abundance of each of the bacterial strains H₁₆ , H₁₉ , H₂₀ , H₂₂ , H₃₃ , H₄₂ , L₁₈ , L₁₉ , L₃₉ , L₄ , L₄₂ , L₅₅ , L₅₉ , SEV₁₃ , SEV₂ , SEV₂₂ , SEV₄ , SEV₁₄ ,
- 15 SEV₁₅ , SEV₂₄ and SEV₂₆;
- b) comparing the obtained abundances with at least one reference value,
- c) determining the prognosis liver cirrhosis of said subject from said comparison.

In a further embodiment, an *in vitro* method for the prognosis of liver cirrhosis for a subject, comprising the following steps:

- 20 a) determining from a biological sample of said subject the abundance of each of the bacterial gene clusters H₁₆, H₂₂, and H₄₂ from table 1, and SEV₄, SEV₁₃ and SEV₁₅ from table 9;
- b) comparing the obtained abundances with at least one reference value,
- c) determining the prognosis liver cirrhosis of said subject from said comparison.

More precisely, the inventors have discovered that the bacterial strains corresponding to the bacterial gene clusters H_16, H_19, H_20 , H_22 , H_33 , H_42 , SEV_13 , SEV_2 , SEV_22 and SEV_4 , are less represented in patients than they are in healthy subjects, compared with the total amount of gut bacteria in the faeces sample.

On the other hand, bacterial strains corresponding to the bacterial gene clusters L_18 , L_19 , L_39 , L_4 , L_42 , L_55 , L_59 , SEV_14 , SEV_15 , SEV_24 and SEV_26 , are more represented in patients than they are in healthy subjects, compared with the total amount of gut bacteria in the faeces sample.

10 In order to implement the present method, the person skilled in the art may for instance chose to compare the abundance of the bacterial strains of interest to references values obtained in healthy subjects. When several bacterial strains of interest are used in the method of the invention, the person skilled in the art may compare the sum of abundances of those bacterial strains. Preferably, the person skilled in the art compares the bacterial strains that are over represented in patients separately from those that are underrepresented.

For instance, the abundance of each of the bacterial strains of interest among H_16, H_19, H_20, H_22 , H_33 , H_42 , SEV_13 , SEV_2 , SEV_22 and SEV_4 may be added, and the sum of the abundances of the bacterial strains obtained is compared with the sum of the abundances obtained for those same strains in healthy subjects. In addition, or alternatively, the abundance of each of the bacterial strains of interest among bacterial strains corresponding to the bacterial gene clusters L_18 , L_19 , L_39 , L_4 , L_42 , L_55 , L_59 , SEV_14 , SEV_15 , SEV_24 and SEV_26 may be added, and the sum of the abundances of the bacterial strains obtained is compared with the sum of the abundances obtained for those same strains in healthy subjects.

Generally, liver damage from cirrhosis cannot be reversed, but treatment can be administered to stop or delay further progression and reduce complications. A healthy diet is encouraged, as cirrhosis may be an energy-consuming process. Antibiotics are prescribed for infections, and various medications can help with
5 itching.

The method of prognosis of the invention enables the person skilled in the art to design a specific prophylactic treatment tailored to the needs of the patient. The prior determination of the prognosis of liver cirrhosis with the method of the invention may thus be followed by the indication and/or the administration of an
10 appropriate treatment or of therapeutic measures.

Thus the present invention also relates to a method for designing a treatment for a subject, said method comprising:

- a) determining the prognosis of the subject with a method according to the invention, and
- 15 b) designing a therapeutic treatment.

According to the invention, step b) of designing a therapeutic treatment may be followed by a step c) of administering said treatment.

Treatments for patients with cirrhosis and having a bad prognosis (i.e. being at risk of severe complications for instance) are well known from the person skilled in the
20 art, and include for instance the use of antibiotics, in particular of rifaximin. Therapeutic measures and treatments for patients with cirrhosis have been detailed in Garcia-Tsao et al.

Thus, a close follow-up is often necessary.

It will appear consistent to the skilled person that the invention further allows for monitoring the evolution of the prognostic of this disease, and the efficacy of treatments for the treatment or the prevention of liver cirrhosis.

Another object of the invention is thus a method for monitoring the efficacy of a
5 treatment of liver cirrhosis in a subject, comprising the steps of:

- a) determining from a first biological sample of said subject the abundance of the bacterial strains L_4, L_10, L_42, L_44, L_12, L_9, L_15, L_14, L_32, L_20, L_19, L_8, L_6, L_11, L_3, L_2, and L_17,
- b) determining from a second biological sample of said subject the abundance of
10 the bacterial strains L_4, L_10, L_42, L_44, L_12, L_9, L_15, L_14, L_32, L_20, L_19, L_8, L_6, L_11, L_3, L_2, and L_17,
- c) comparing the abundance obtained in step a) and the abundance obtained in step b),
- d) determining if said treatment is efficient for said subject from the comparison
15 of step c).

In another embodiment, the invention relates to a method for monitoring the efficacy of a treatment of liver cirrhosis in a subject, comprising the steps of:

- a) determining from a first biological sample of said subject the abundance of each of the bacterial gene clusters H_16, H_22, and H_42 from table 1, and
20 SEV_4, SEV_13 and SEV_15 from table 9;
- b) determining from a second biological sample of said subject the abundance of each of the bacterial gene clusters H_16, H_22, and H_42 from table 1, and SEV_4, SEV_13 and SEV_15 from table 9;

- c) comparing the abundance obtained in step a) and the abundance obtained in step b),
- d) determining if said treatment is efficient for said subject from the comparison of step c).

5 Preferably, the first sample corresponds to a sample collected before implementation of said treatment, and the second sample corresponds to a sample collected after implementation of said treatment, in the same subject

In an embodiment, the second sample corresponds to a sample collected at least one week, at least two weeks, at least three weeks, at least one month of
10 implementation of the treatment.

Preferably, when comparing the abundance obtained in step a) and the abundance obtained in step b), the person skilled in the art will compare the sum of the abundances of each of the bacterial strains L_4, L_10, L_42, L_44, L_12, L_9, L_15, L_14, L_32, L_20, L_19, L_8, L_6, L_11, L_3, L_2, and L_17 obtained for the first
15 sample, to the sum of the abundances of each of the bacterial strains L_4, L_10, L_42, L_44, L_12, L_9, L_15, L_14, L_32, L_20, L_19, L_8, L_6, L_11, L_3, L_2, and L_17 obtained for the second sample.

The treatment will be considered efficient for the subject if the abundance determined in step a) is superior to the abundance obtained in step b). In particular,
20 the treatment will be considered efficient for the subject if the sum of the abundances of each of the bacterial strains L_4, L_10, L_42, L_44, L_12, L_9, L_15, L_14, L_32, L_20, L_19, L_8, L_6, L_11, L_3, L_2, and L_17 obtained for the first sample, is superior to the sum of the abundances of each of the bacterial strains L_4,

L_10, L_42, L_44, L_12, L_9, L_15, L_14, L_32, L_20, L_19, L_8, L_6, L_11, L_3, L_2, and L_17 obtained for the second sample.

The invention further pertains to a method for monitoring the efficacy of a treatment of liver cirrhosis in a subject, comprising the steps of:

- 5 a) determining from a first biological sample of said subject the abundance of each of the bacterial strains H_16 , H_19 , H_20 , H_22 , H_33 , H_42 , L_18 , L_19 , L_39 , L_4 , L_42 , L_55 , L_59 , SEV_13 , SEV_2 , SEV_22 , SEV_4 , SEV_14 , SEV_15 , SEV_24 and SEV_26;
- b) determining from a second biological sample of said subject the abundance
10 each of the bacterial strains H_16 , H_19 , H_20 , H_22 , H_33 , H_42 , L_18 , L_19 , L_39 , L_4 , L_42 , L_55 , L_59 , SEV_13 , SEV_2 , SEV_22 , SEV_4 , SEV_14 , SEV_15 , SEV_24 and SEV_26;
- c) comparing the abundance obtained in step a) and the abundance obtained in step b),
- 15 d) determining if said treatment is efficient for said subject from the comparison of step c).

The invention further concerns a kit for the *in vitro* the diagnosis of, for assessing whether a subject is at risk of developing, and/or for the prognosis of liver cirrhosis, comprising at least one reagent for the determination of the copy number of at least
20 one gene having a sequence selected from SEQ ID Nos. 1-3700. In a preferred embodiment, the kit of the invention comprises at least one reagent for the determination of the copy number of each of the genes of sequence SEQ ID Nos. 1-3700. In another embodiment, the kit of the invention comprises reagent for the determination of at least 5, at least 10, at least 20, at least 30, at least 40 or at

least 50 bacterial genes from each of a bacterial gene clusters comprised in any of the combination of bacterial gene clusters according to the invention.

By “a reagent for the determination of the copy number of at least one gene”, it is meant a reagent which specifically allows for the determination of the copy
5 number of the said gene, i.e. a reagent specifically intended for the specific determination of the copy number of at least one gene having a sequence selected from SEQ ID Nos. 1-3700. This definition excludes generic reagents useful for the determination of the expression level of any gene, such as Taq polymerase or an amplification buffer, although such reagents may also be included in a kit according
10 to the invention.

In an embodiment, the kit of the invention comprises at least one reagent for the determination of the copy number of at least one gene having a sequence having a sequence selected from SEQ ID Nos. 1-3300. In a preferred embodiment, the kit of the invention comprises at least one reagent for the determination of the copy
15 number of each of the genes of sequence SEQ ID Nos. 1-3300.

In another embodiment, the kit of the invention comprises at least one reagent for the determination of the copy number of at least one gene having a sequence selected from SEQ ID Nos. 351 to 400, 501 to 550, 601 to 650, 701 to 750, 1201 to 1250, 2851 to 2900, 2301 to 2350, 2351 to 2400, 2701 to 2750, 2751 to 2800, 2851 to
20 2900, 3001 to 3050, 3051 to 3100 and 3301-3700. In a preferred embodiment, the kit of the invention comprises at least one reagent for the determination of the copy number of each of the genes of sequence SEQ ID Nos. 351 to 400, 501 to 550, 601 to 650, 701 to 750, 1201 to 1250, 2851 to 2900, 2301 to 2350, 2351 to 2400, 2701 to 2750, 2751 to 2800, 2851 to 2900, 3001 to 3050, 3051 to 3100 and 3301-3700.

Such a reagent for the determination of the copy number of at least one gene can be for example a dedicated microarray as described above or amplification primers specific for at least one gene having a sequence selected from SEQ ID Nos. 1-3700.

The present invention thus also relates to a kit for the *in vitro* the diagnosis of, for
5 assessing whether a subject is at risk of developing, and/or for the prognosis of liver cirrhosis, said kit comprising a dedicated microarray as described above or amplification primers specific for at least one gene having a sequence selected from SEQ ID Nos. 1-3700, preferably a sequence selected from SEQ ID Nos. 1-3300 or a sequence selected from SEQ ID Nos. 351 to 400, 501 to 550, 601 to 650, 701 to 750,
10 1201 to 1250, 2851 to 2900, 2301 to 2350, 2351 to 2400, 2701 to 2750, 2751 to 2800, 2851 to 2900, 3001 to 3050, 3051 to 3100 and 3301-3700. Here also, when the kit comprises amplification primers, while said kit may comprise amplification primers specific for other genes, said kit preferably comprises at most 100, at most 75, 50, at most 40, at most 30, preferably at most 25, at most 20, at most 15, more preferably
15 at most 10, at most 8, at most 6, even more preferably at most 5, at most 4, at most 3 or even 2 or one or even zero couples of amplification primers specific for other genes than the genes of sequences SEQ ID Nos 1-3700, preferably a sequence selected from SEQ ID Nos. 1-3300 or a sequence selected from SEQ ID Nos. 351 to 400, 501 to 550, 601 to 650, 701 to 750, 1201 to 1250, 2851 to 2900, 2301 to 2350,
20 2351 to 2400, 2701 to 2750, 2751 to 2800, 2851 to 2900, 3001 to 3050, 3051 to 3100 and 3301-3700. For example, said kit may comprise at least a couple of amplification primers for at least one gene in addition to the primers for at least one gene having a sequence selected from SEQ ID Nos. 1-3700, preferably a sequence selected from SEQ ID Nos. 1-3300 or a sequence selected from SEQ ID Nos. 351 to 400, 501 to 550, 601 to
25 650, 701 to 750, 1201 to 1250, 2851 to 2900, 2301 to 2350, 2351 to 2400, 2701 to 2750, 2751 to 2800, 2851 to 2900, 3001 to 3050, 3051 to 3100 and 3301-3700.

EXAMPLES

Methods:

Patient Information

Liver cirrhosis was diagnosed according to the international guidelines by
5 comprehensive consideration of liver biopsy, imaging examination, clinical
symptoms, physical signs, laboratory tests, medical history, progress notes and
cirrhosis associated complications. Biopsy as the golden standard for cirrhosis
diagnosis was used for 46 out of the 123 (37.4%) patients. As biopsy was counter-
indicated for patients with conditions such as refractory as cites and obvious bleeding
10 tendency, the remaining 77 (62.6%) were diagnosed using all other approaches
combined. To confirm diagnoses, we solicited outside expert opinions for each case.
Borderline or otherwise inconclusive cases were excluded from the study. After
patient discharge from the hospital, his/her case history was further reviewed for
medication history. Cases that progressed to hepatic carcinoma or those found to
15 suffer from other diseases such as hypertension and diabetes were excluded.

The control group included 114 healthy volunteers who visited the First Affiliated
Hospital of Zhejiang University in China for their annual physical examination. The
liver imaging and liver biochemistry results of all healthy controls were in the normal
range. Physical examination, routine examination of blood, urine and stools,
20 preoperative serological tests (including the detection of hepatitis B surface antigen,
hepatitis C virus antibody, treponema pallidum antibody, human immunodeficiency
virus antibody), liver function, renal function, electrolyte, liver ultrasound,
electrocardiogram, chest X-ray results were checked in the healthy controls to
exclude any abnormal samples. Comprehensive clinical information for each enrolled
25 individuals was recorded . Control group exclusion criteria included hypertension,

diabetes, obesity, metabolic syndrome, IBD, non-alcoholic fatty liver disease, coeliac disease and cancer. Individuals who received antibiotics and/or probiotics within eight weeks before enrolment were also excluded. All participants, or their legally authorized representatives, provided a written informed consent upon enrolment.

- 5 The study conformed to the ethical guidelines of the 1975 Declaration of Helsinki and was approved by the Institutional Review Board of the First Affiliated Hospital of Zhejiang University.

Human faecal sample collection and DNA extraction

Each cirrhotic patient and healthy control subject provided a fresh stool sample that
10 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 method^{16,47} as previously described. DNA concentration was measured by nanodrop (Thermo Scientific) and its
15 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
20 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, l6n, 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 BWA⁴⁸ 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.

10 *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 SOAPdenovo⁴⁹ (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 2.21⁵⁰ 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.

Construction of non-redundant human gut gene set

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 25 2000 (Illumina, San Diego, CA). This produced an average of 4.74 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 SOAPdenovo⁴⁹. Unassembled reads from 166 samples were pooled and the *de novo* assembly process was performed again for these reads (see Supplementary Methods and. 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 study²⁹. 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 set²⁹ and T2D study²².

MetaGeneMark⁵¹ (prokaryotic GeneMark.hmm version 2.8) was used to predict ORFs in scaffolds without ambiguous bases. The 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.

Then Genes from the LC, T2D and MetaHIT gene catalogs were merged to create a non-redundant gene set for subsequent analyses. 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
 5 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
 10 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
 15 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
 20 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
 5 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. Species abundance was added to obtain the genus level profile table. For some species that do not have a genus, they are denoted as unclassified genera for each species.

10 Gene abundance profiling

Reads were aligned against the gene set by using SOAPalign⁵⁰ 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
 15 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 multiple reads, it goes as follows.

$$\begin{aligned} Ab(G) &= Ab(U) + Ab(M) \\ 20 \quad Ab(U) &= U / l \\ Ab(M) &= (\sum_{i=1}^M Co * \{M\}) / l \end{aligned}$$

$Ab(U)$ 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.

$$5 \quad Co = \frac{U}{\sum_{i=1}^N Ab(U)}$$

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
 10 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.

15 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 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
 20 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.

Gene functional classification and NOG/KO abundance profiling

Protein sequences of the predicted genes were searched using NCBI blastP against eggNOG 3.0 database⁵² and KEGG gene database (KEGG FTP release 2013-01-21) with parameters *-num_descriptions 100000, -evalue 1e-5*. Genes that had alignments with bits score higher than 60 were assigned into one or more NOG or KO. We used the methods introduced in Qin J et al. Nature 2010²⁹ to calculate abundance of proteins archived in eggNOG and KEGG database. To calculate abundance of NOG or KO, we added abundances of proteins assigned into same NOG or KO, as abundance of NOG or KO, then profiles of NOG/KO were generated.

10 Gene biomarker identification

Genes from the gene-profile matrix were used in an association study aiming to identify those that are differentially abundant between the patient and the healthy groups. Wilcoxon tests were employed to compute the probabilities that frequency profiles do not differ between the patient and the healthy groups by chance alone.

15 Benjamini Hochberg multiple test correction was applied to the p-values. By performing a selection only based on a p-value threshold of $p < 0.01$ we found 541,582 genes. For specificity and computational reasons we used a very stringent significance threshold of $fdr < 0.0001$. This process identified 75,245 genes that are differentially abundant between the groups (49,830 were more abundant in the liver cirrhosis patients and 25,415 in the healthy control group). A similar p-value and group enrichment method was calculated for the NOG/KO as well.

MetaGenomic Species (MGS)

We followed the approach described by Le Chatelier et al., Nature 2013 and Nielse, Almeida et al., Nat Biotech (in press) to cluster genes from the current study into
25 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 (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 ($\rho > 0.7$).

- 5 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 ρ greater than 0.85 were fused. This procedure was applied separately to the 49,830 genes enriched in liver cirrhosis patients and the 25,415 genes enriched in healthy
- 10 controls. 21,423 out of the 25,415 “healthy” genes fall into 43 clusters composed of 51 to 2,702 genes after the first clustering step and 38 clusters of 51 to 2,970 genes after the second step. 31,386 out of the 49,830 “liver cirrhosis” genes fall into 60 clusters of 51 to 3,000 genes after the first clustering step and 28 clusters of 51 to 5,755 genes after the second step.
- 15 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
- 20 “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). 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
- 25 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),
 5 *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);
- 10 (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
 15 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 metagenomes³¹ and failed to reveal identity beyond that found with sequenced genomes.

Only a small minority of the 38 MGS enriched in healthy individuals (15.8%) could be
 20 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, reflecting a paucity of isolated and sequenced gut strains. Furthermore, it is striking that all 38 MGS enriched in healthy Chinese were
 25 found to be present in the Danish cohort. 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 and indicates a highly modified gut microbial composition.

5 Co-occurrence network of MGSs

The 66 marker profiles of the differentially abundant MGS between patient and healthy individuals were correlated separately for patients and for healthy, essentially as described by Faust et al.⁵³. For each of the 2,112 possible edges ((66*66/2)-66) we computed 1,000 permutations by renormalizing the data after each
10 step and computed Spearman correlation coefficients in order to obtain the null distributions due to the compositionality effect⁵³. For each of the edges we also computed the bootstrap distribution of the Spearman correlation coefficient in order to have the confidence interval and the corresponding variance. We next applied for each edge a z-test with the pooled variance from both distributions and computed a
15 significance p-value. Multiple testing corrections were applied to the p-values using the Benjamini-Hochberg method and only those having an $fdr < 1e-9$ were used to construct the network. This fdr threshold corresponds approximately to a $\rho > 0.4$. The network reflects strong correlations which are not spurious and which are not due to the compositionality effect. The resulting network is displayed as Figures 1
20 and 2.

Disease diagnostic MGS models

The difference in microbial signal is relatively strong between patients and healthy individuals and this can be used to predict accurately the disease status from the gut microbial data.

To build the models, the signal from the different species were combined, computing the sum of median abundance of MGS enriched in patients minus the sum of those enriched in healthy controls. For that purpose the mopred R package (developed at Metagenopolis as part of MetaOMineR suite) was used, which explores the
5 combinatory space of the models using genetic-algorithm based heuristics. The accuracy of a model was evaluated using the area under the curve (AUC).

The best models were selected on the discovery cohort (181 samples: 98 patients and 83 controls) and a confidence interval was computed using 1000 bootstraps (by randomly drawing 90% of the cohort). Next the accuracy of the model in the
10 validation cohort (56 samples: 26 patients and 31 controls) was computed along with the respective bootstraps. A model combining 6 features (MGS) gives an AUC of 0.947 in the discovery cohort and an AUC of 0.933 in the validation cohort with a narrow confidence interval.

Disease severity prediction models

15 MELD and CTP: severity scores

The severity of the disease was estimated using two different scores MELD and Child-Pugh scores largely used by clinicians computed as follows:

MELD Score = $(0.957 * \ln(\text{Serum Cr}) + 0.378 * \ln(\text{Serum Bilirubin}) + 1.120 * \ln(\text{INR}) + 0.643) * 10$ (if hemodialysis, value for Creatinine is automatically set to 4.0). Note: If
20 any score is <1, the MELD assumes the score is equal to 1.

References relative to the method.

- 1 Li, M. et al. Symbiotic gut microbes modulate human metabolic phenotypes.
Proceedings of the National Academy of Sciences of the United States of America105,
2117-2122, doi:10.1073/pnas.0712038105 (2008).
- 2 Turnbaugh, P. J. et al. An obesity-associated gut microbiome with increased
5 capacity for energy harvest. Nature444, 1027-1031, doi:10.1038/nature05414 (2006).
- 3 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-
Wheeler transform. Bioinformatics25, 1754-1760, doi:10.1093/bioinformatics/btp324
(2009).
- 4 Li, R. et al. De novo assembly of human genomes with massively parallel short
10 read sequencing. Genome research20, 265-272, doi:10.1101/gr.097261.109 (2010).
- 5 Li, R. et al. SOAP2: an improved ultrafast tool for short read alignment.
Bioinformatics25, 1966-1967, doi:10.1093/bioinformatics/btp336 (2009).
- 6 Noguchi, H., Park, J. & Takagi, T. MetaGene: prokaryotic gene finding from
environmental genome shotgun sequences. Nucleic acids research34, 5623-5630,
15 doi:10.1093/nar/gkl723 (2006).
- 7 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
20 Transplantation Society13, 1582-1588, doi:10.1002/lt.21277 (2007).
- 8 Qin, J. et al. A metagenome-wide association study of gut microbiota in type
2 diabetes. Nature490, 55-60, doi:10.1038/nature11450 (2012).

- 9 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).
- 10 Arumugam, M. et al. Enterotypes of the human gut microbiome. *Nature*473,
5 174-180, doi:10.1038/nature09944 (2011).
- 11 Powell, S. et al. eggNOG v3.0: orthologous groups covering 1133 organisms at
41 different taxonomic ranges. *Nucleic acids research*40, D284-289, doi:10.1093/nar/gkr1060 (2012).
- 12 de Hoon, M. J., Imoto, S., Nolan, J. & Miyano, S. Open source clustering
10 software. *Bioinformatics*20, 1453-1454, doi:10.1093/bioinformatics/bth078 (2004).
- 13 Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P. L. & Ideker, T. Cytoscape
2.8: new features for data integration and network visualization. *Bioinformatics*27,
431-432, doi:10.1093/bioinformatics/btq675 (2011).
- 14 Ding, C. & Peng, H. Minimum redundancy feature selection from microarray
15 gene expression data. *Journal of bioinformatics and computational biology*3, 185-205 (2005).
- 15 Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et biophysica acta*405, 442-451 (1975).
- 16 Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. & Nielsen, H. Assessing the
20 accuracy of prediction algorithms for classification: an overview. *Bioinformatics*16, 412-424 (2000).

Results

Liver cirrhosis gut microbial gene catalogue

We constructed a gene catalogue from 98 Chinese liver cirrhosis patients and 83 healthy Chinese controls using the methodology developed by MetaHIT. The catalogue, termed LC here (for Liver Cirrhosis) contained 2,688,468 non-redundant ORFs. We compared it with three other gut microbial catalogues, MetaHIT²⁹, HMP²⁵, and T2D²². To facilitate this comparison, 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 common to all catalogue. 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 catalogues were merged; the HMP was not included, as it contained Sanger, 454 or Illumina based 16S sequences, in addition to whole metagenomic data. The merged non-redundant catalogue contained 5,382,817 genes.

Gut microbial species associated with liver cirrhosis

Our investigation included two phases. The first was discovery with 98 liver cirrhosis patients and 83 healthy controls whereas the second was validation, with additional
5 25 patients and 31 controls. In the discovery phase, Wilcoxon rank-sum test corrected for multiple testing by Benjamini Hochberg method was used to identify differentially abundant genes in patients and controls. At a stringent threshold ($\text{fdr} < 0.0001$), 75,245 genes were found, 49,830 were more abundant in the patients and 25,415 in the controls. Patients and controls could be clearly separated by a PCA
10 analysis based on the 75,245 genes; this was confirmed with the validation.

To explore further the microbial genes associated with liver cirrhosis we grouped them into clusters, denoted metagenomic species (MGS) here, based on their abundance profiles^{27,30}. Of the 66 MGS, 38 and 28 were enriched in healthy individuals and patients, respectively. Their significantly different abundance distribution between healthy and liver cirrhosis subjects is shown in Fig 1. A majority (82%) were also differentially abundant in the validation cohort ($q < 0.05$), in spite of the reduced statistical power due to the smaller cohort size.

Composition of bacterial communities varies considerably as a function of the overall gene richness^{27,28} and the loss of richness is associated with obesity and IBD^{27,28,31}. A large majority of the 38 MGS enriched in the healthy individuals (33: 86.8%) was correlated with the richness at a $q < 10^{-3}$ in the Chinese cohort; 26 of these (78.8%) were similarly correlated in the Danish cohort. These observations indicate that gut communities of healthy individuals across the continents may be largely similar. Furthermore, gene richness was much lower in liver cirrhosis patients than in healthy individuals (on average 389 000 and 497 000 genes, respectively). Interestingly, among the species enriched in healthy Chinese was *Faecalibacterium prausnitzii*, which has anti-inflammatory properties and was found in a “healthy” gene-rich microbiome^{27,28}, and *Coprococcus comes*, which might contribute to gut health through butyrate production. A similar butyrate production role may be played by 3 *Lachnospiraceae* and 5 *Ruminococcaceae* enriched in healthy individuals. Lower abundance of these species in liver cirrhosis patients indicates that they have a less healthy gut microbiome.

Most interestingly, a high proportion of MGS enriched in patients belong to taxa such as *Veillonella* ($n=8$) or *Streptococcus* ($n=6$), known to include species of oral origin. However, the small intestine also harbours such species³² and the small intestinal

bacterial overgrowth (SIBO) is frequently found in liver cirrhosis patients³³. To explore the origin of the patient-enriched species, we used the HOMD³⁴ and GOLD³⁵ databases information about the origin of the closely related sequenced isolates. We also constructed a catalogue of 114 publicly available genomes for *Streptococcus*,
5 *Fusobacterium*, *Lactobacillus*, *Veillonella* and *Megasphaera* strains, originating mostly from mouth or gut (57 or 28, respectively) and used it for blastN and blastP analysis (Supplementary Methods). 13 of the species were closest to an oral isolate whereas only 6 were closest to the gut isolates, a single species being from the ileum. Comparison with the three ileum metagenomes failed to identify identity above that
10 detected by comparison with the sequenced genomes. We conclude that oral commensals invade the gut in liver cirrhosis patients. Possibly, an altered bile production in cirrhosis renders gut more permissible and/or accessible to “foreign” bacteria, as bile resistance may be required for survival in the human gut^{36,37}. As patient-enriched MGS include pathogens such as *Campylobacter* and *H.*
15 *parainfluenzae*, these also might use the oral route to invade the gut, possibly via contaminated food. The invasion species foreign to the niche may occur not only in colon but also in ileum, and contribute to the liver cirrhosis-associated SIBO. Among the patient-enriched species were *Streptococcus anginosus*, *Veillonella atypica*, *Veillonella dispar*, *Veillonella* sp. oral taxon, and *Clostridium perfringens*, which have
20 been reported to cause opportunistic infections³⁸⁻⁴⁰.

To analyse the relations between the liver cirrhosis-associated MGS we generated co-abundance based networks, for healthy individuals and liver cirrhosis patients, (Fig 2 right). A striking feature is that taxonomically related species tend to cluster, as reported previously²⁹. These observations indicate that the gut environment becomes
25 permissive for the development and maintenance of the related taxa in many individuals. Obviously, taxonomically unrelated species can also thrive in such

environments, as observed with *Campylobacter concisus*, *Haemophilus parainfluenzae* or *Fusobacterium*, which tend to be associated with *Veillonella* in patients. The overall abundance of species enriched in patients reached high levels, exceeding 5% in over a quarter and approaching the extreme of 40%, whereas it was
5 very low in healthy individuals. Interestingly, the severity of the disease was positively correlated with the abundance of a number of MGS enriched in patients and negatively correlated with those of the MGS enriched in controls (and therefore under-represented in patients; Fig 1). The disease status of the patients with the highest load of these bacteria was significantly worse than that of the patients with
10 the lowest load (Fig 2). Such a “dose response” is consistent with an active role of the enriched species in liver cirrhosis.

Microbial functions enriched in liver cirrhosis

To investigate the functional role of the gut microbiota in liver cirrhosis, we identified 13,970 eggNOG orthologues and 4,801 KEGG orthologues and associated
15 with the disease. The most abundant KEGG orthologs in patients and controls were enzyme families. The most enriched orthologs in patients was membrane transport, similar to findings for inflammatory bowel diseases^{19,20}, obesity⁴¹ and T2D²². In contrast, the most prevalent markers among the controls included those involved in carbohydrate metabolism, amino acid metabolism, energy metabolism, signal
20 transduction and the metabolism of cofactors and vitamins. At the module or pathway level, the liver cirrhosis-associated markers included assimilation or dissimilation of nitrate to or from ammonia, denitrification, GABA biosynthesis, GABA (gamma-Aminobutyrate) shunt, heme biosynthesis, phosphotransferase systems (PTS) and some types of membrane transport, such as amino acid transport. The control-
25 enriched modules included histidine metabolism, ornithine biosynthesis, creatine

pathway, carbohydrate metabolism, repair system and glycosaminoglycan metabolism.

The enrichment of the modules for ammonia production in patients suggests a potential role of gut microbiota in hepatic encephalopathy, a liver cirrhosis-related complication characterised by hyperammonemia. Overproduction of ammonia by gut bacteria might contribute to increased levels of ammonia in blood. Manganese related transport system modules enriched in patients possibly contribute to the changes of concentrations of manganese. The accumulation of manganese within the basal ganglia in patients with end-stage liver disease may have a role in the pathogenesis of chronic hepatic encephalopathy⁴², a main complication of liver cirrhosis. The hydrodynamic venous shunt and liver failure could promote this accumulation, which, in turn, causes metabolic disorders of the nerve cell proteins, affects transmission function of neural synaptic, and eventually leads to hepatic encephalopathy⁴⁰. Finally, the modules for GABA biosynthesis were enriched in the patients. GABA neurotransmitter system is involved in the pathogenesis of hepatic encephalopathy in humans⁴³. Because of the hydrodynamic venous shunt and liver failure, GABA levels in the blood are increased⁴⁴, and could go through the blood brain barrier to activate GABA receptor and cause hepatic encephalopathy. Microbiome modulation, aiming for manganese elimination and lowering of GABA levels in the gut, might provide a new therapeutic option for the treatment of hepatic encephalopathy.

Identification of disease severity associated MGS

To identify disease severity associated MGS, the liver cirrhosis patients of the discovery cohort (n=98) were used and a correlation study for all genes with either CTP or MELD parameter of the patients was performed. Based on profiles of these 98

training samples, a Spearman correlation test was performed to identify genes correlated with either CTP or MELD. Using threshold of $p < 0.001$, 18,830 genes correlated to CTP and 12,177 genes correlated to MELD were found, leading to a unique set of 25,214 genes correlated to severity.

5 Of the 25,214 correlated genes, a majority (63%) clustered into 33 MGS. These MGS were filtered according to the following criteria:

- MGS with sparse signal (very low occurrence in the cohort) were removed.
- MGS, which appeared to be non contrasted for disease vs. healthy using the whole discovery cohort were removed.
- 10 - MGS which did not clearly contrast high and low CTP patients in the discovery and validation cohorts were also removed.

Finally, after this filtering procedure only 21 MGS were obtained, 13 of which belonging to the 66 LC disease MGS and only 8 being new.

Severity prediction models

- 15 Given the high performance of the disease diagnostic MGS models, we investigated our ability to construct models that could predict the severity of the disease based on the gut MGS. To that purpose, we used the 21 severity MGS (13 previously identified as disease MGS as well as the 8 new disease severity associated MGS).

Predictions based on MELD severity score

- 20 MELD as a continuous severity score

The MELD is a continuous variable following a normal distribution. The first idea was to correlate MELD with the model score and select the best one based on the

correlation coefficient. In theory this works well for linear relations but less well for other more complex relations. The distribution of the MGS signals or the resulting models is not normal and contains outlier values having a leverage effect on the correlation. When using Spearman correlation the best N4 model found in the
5 discovery cohort has a rho of 0.56 and a rho = 0.32 in the validation cohort (see Figure 7).

The correlation operator seems not to be adapted to this situation and for this a second strategy is proposed, which is introduced hereafter.

MELD as a discrete risk (low / high severity) score

10 Using a threshold of 15, MELD score was used to split the population in two groups with mild (≤ 15) or severe (>15) risk score, the latter leading to a recommendation for liver transplantation. The cirrhotic patients selected for the cohort is biased for low severity risk score based on the >15 threshold. Only 9 out of 98 patients of the discovery cohort and 7 out of 25 patients of the validation cohort had a MELD score
15 above 15.

The models were built as for disease diagnostics models. The signal from the different species was combined, computing the sum of median abundance of MGS enriched in high severity cirrhosis minus the sum of those enriched in low severity. The accuracy of a model was evaluated using the area under the curve (AUC). The
20 best models were selected on the patients of the discovery cohort (98 patients: 89 "low severity" and 9 "high severity") and a confidence interval was computed using 1000 bootstraps (by randomly drawing 90% of the cohort). Next the accuracy of the model in the validation cohort (25 patients: 18 "low severity" and 7 "high severity") was computed along with the respective bootstraps. A model combining 6 features

(MGS) gives an AUC of 0.880 in the discovery cohort and an AUC of 0.762 in the validation cohort with a quite narrow confidence interval (see Figures 8 and 9).

Predictions based on CTP severity score

CTP as a continuous severity score

- 5 The CTP is a discrete variable that can be considered as continuous that does not follow a normal distribution. First, we correlated CTP with the model score and select the best one based on the correlation coefficient. As for MELD, the distribution of the MGS signals or the resulting models is not normal and contains outlier values having a leverage effect on the correlation. When using Spearman correlation the
- 10 best model N7 found in the discovery cohort has a rho of 0.61 and a rho = 0.14 in the validation cohort (see Figure 10).

CTP as a discrete risk (low / high severity) score

- Using a threshold of 7 as proposed a CTP score was used to split the population in two groups with mild (<7) or severe (≥ 7) risk score, the latter leading to a one year
- 15 survival probability $< 100\%$. The population of selected cirrhotic patients of the cohort is biased for high severity risk score based on this ≥ 7 threshold. 66 out of 98 patients of the discovery cohort and 15 out of 25 patients of the validation one had a CTP score at or above 7. The best models were selected on the patients of the discovery cohort (98 patients: 32 "low severity" and 66 "high severity") and a
- 20 confidence interval was computed using 1000 bootstraps (by randomly drawing 90% of the cohort). Next we computed the accuracy of the model in the validation cohort (25 patients: 10 "low severity" and 15 "high severity") along with the respective bootstraps. A model combining 5 features (MGS) gives an AUC of 0.808 in the

discovery cohort and an AUC of 0.767 in the validation cohort with a quite narrow confidence interval.

Conclusion

- 5 The liver cirrhosis associated MGS have an unprecedented statistical power to predict the disease but also its severity. They can be used as a non-invasive and highly accurate diagnostic tool but also they could be applied to stratify the diseased population in mild and severe forms. Furthermore we have also shown in the accompanying study that some of the disease-associated MGS are in high abundance
- 10 in some individuals considered to be healthy. It may be that these individuals are in a pre-cirrhotic state. This indicates that the above introduced MGS have a potential of predicting the future disease

Discussion

- To study gut microbiota in liver cirrhosis we first established a novel gut gene
- 15 catalogue (LC catalogue), including 98 liver cirrhosis patients and 83 healthy control individuals. Comparison with the previously established metaHIT and T2D gene catalogues indicated a common core of 800,000 genes and a considerable proportion of catalogue-specific genes (37.01% of metaHIT, 36.59% of T2D and 18.02% of LC) indicating that the current gene sets are still limited and should be completed by
- 20 inclusion of more individuals. Interestingly, although the T2D and LC gene sets are both derived from Chinese populations, the number of unique genes in each gene set was large. This might be due to the difference in diseases and also to the different genotypes, BMI, age⁴⁴ and dietary habits⁴⁵. Furthermore, there was a significant difference in BMI of the two studies. Nevertheless, there was no significant

difference of the abundance of main phyla between two studies ($P>0.01$); of the top 30 highest abundance genera and species, 28 and 26, respectively, were consistent in two studies, and there were no significant differences in the abundance of most of them. Furthermore, the top 4 species were exactly the same. These results point
5 towards overall similarity of the microbiota in healthy individuals.

Use of the LC gene catalogue, in conjunction with the quantitative metagenomics approach, revealed a major change of the gut microbiota in the liver cirrhosis patients, corresponding to a massive invasion of the gut by oral bacterial species, most of which were characterized previously by genome sequencing. Positive
10 correlation of the severity of the disease with the overall abundance of the oral species suggests that they may play an active role. This was not noted in a previous study, where the 16 S-based approach likely lacked the required species-level resolution, even if similar taxonomy change trends between the liver cirrhosis group and the healthy controls in the phylum, class and order levels were observed¹³. At
15 the family level, in both studies Veillonellaceae and Streptococcaceae showed enrichment in liver cirrhosis group whereas Coprococcus and Roseburia, genera affiliated to the Lachnospiraceae family were significantly decreased. In contrast, a small but significant difference in Enterobacteriaceae (median, 1.47% versus 4.58%, $p=0.039$), which was observed previously, was not detected in the current study;
20 future work might allow resolving this discrepancy. Detection of species depleted in patients as compared with healthy individuals, many of which were negatively associated with the severity of the disease opens avenues to the development of novel probiotics, which might help combat aggravation of liver cirrhosis.

- 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).
- 5 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.
10 *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).
- 15 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).
- 20 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
5 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).
- 10 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 Human Microbiome Jumpstart Reference Strains, C. et al. A catalog of
15 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).
- 20 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
5 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).
- 10 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).
- 15 23 Arumugam, M. et al. Enterotypes of the human gut microbiome. *Nature* 473, 174-180, doi:10.1038/nature09944 (2011).
- 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 Human Microbiome Project, C. A framework for human microbiome research.
20 *Nature* 486, 215-221, doi:10.1038/nature11209 (2012).
- 26 Human Microbiome Project, C. 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).
- 5 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*
- 10 *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 Koren, O. et al. A guide to enterotypes across the human body: meta-analysis
- 15 of microbial community structures in human microbiome datasets. *PLoS computational biology* 9, e1002863, doi:10.1371/journal.pcbi.1002863 (2013).
- 33 Karlsson, F. H. et al. Symptomatic atherosclerosis is associated with an altered gut metagenome. *Nature communications* 3, 1245, doi:10.1038/ncomms2266 (2012).
- 20 34 Albertsen, M. et al. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature biotechnology* 31, 533-538, doi:10.1038/nbt.2579 (2013).

- 35 Saarela, M., Mogensen, G., Fonden, R., Matto, J. & Mattila-Sandholm, T.
Probiotic bacteria: safety, functional and technological properties. *Journal of*
biotechnology 84, 197-215 (2000).
- 36 Merritt, M. E. & Donaldson, J. R. Effect of bile salts on the DNA and
5 membrane integrity of enteric bacteria. *Journal of medical microbiology* 58, 1533-
1541, doi:10.1099/jmm.0.014092-0 (2009).
- 37 Marchandin, H. et al. Prosthetic joint infection due to *Veillonella dispar*.
European journal of clinical microbiology & infectious diseases : official publication
of the European Society of Clinical Microbiology 20, 340-342 (2001).
- 10 38 Hwang, J. J., Lau, Y. J., Hu, B. S., Shi, Z. Y. & Lin, Y. H. *Haemophilus*
parainfluenzae and *Fusobacterium necrophorum* liver abscess: a case report. *Journal*
of microbiology, immunology, and infection = Wei mian yu gan ran za zhi 35, 65-67
(2002).
- 39 Xu, M. et al. Changes of fecal *Bifidobacterium* species in adult patients with
15 hepatitis B virus-induced chronic liver disease. *Microbial ecology* 63, 304-313,
doi:10.1007/s00248-011-9925-5 (2012).
- 40 Greenblum, S., Turnbaugh, P. J. & Borenstein, E. Metagenomic systems
biology of the human gut microbiome reveals topological shifts associated with
obesity and inflammatory bowel disease. *Proceedings of the National Academy of*
20 *Sciences of the United States of America* 109, 594-599, doi:10.1073/pnas.1116053109
(2012).
- 41 Krieger, D. et al. Manganese and chronic hepatic encephalopathy. *Lancet* 346,
270-274 (1995).

- 42 Ferenci, P., Schafer, D. F., Kleinberger, G., Hoofnagle, J. H. & Jones, E. A.
Serum levels of gamma-aminobutyric-acid-like activity in acute and chronic
hepatocellular disease. *Lancet* 2, 811-814 (1983).
- 43 Minuk, G. Y., Winder, A., Burgess, E. D. & Sarjeant, E. J. Serum gamma-
5 aminobutyric acid (GABA) levels in patients with hepatic encephalopathy. *Hepato-
gastroenterology* 32, 171-174 (1985).
- 44 Yatsunenko, T. et al. Human gut microbiome viewed across age and
geography. *Nature* 486, 222-227, doi:10.1038/nature11053 (2012).
- 45 Wu, G. D. et al. Linking long-term dietary patterns with gut microbial
10 enterotypes. *Science* 334, 105-108, doi:10.1126/science.1208344 (2011).

CLAIMS

1. An *in vitro* method for the diagnosis of liver cirrhosis in a subject and /or
for assessing whether a subject is at risk of developing liver cirrhosis,
5 comprising the following steps:
 - a) determining from a biological sample of said subject the abundance of
each of the bacterial gene clusters of a set of bacterial gene clusters,
wherein each of said clusters consist in 50 non-redundant and covariant
bacterial genes belonging to the same genome, defined in Tables 1,
10 wherein said set of bacterial gene clusters comprise at least 2 bacterial
gene clusters chosen in the group consisting in H_42, H_13, H_32, H_4,
H_19, H_20, H_30, H_33, H_36, H_1, H_43, H_5, H_22, H_6, H_18, H_16,
H_10, H_8, H_34, H_29, H_14, H_23, H_26, H_37, H_17, H_7, H_40, H_3,
H_11, H_15, H_12, H_38, H_21, H_24, H_2, H_9, H_28, H_25, L_18, L_4,
15 L_7, L_10, L_42, L_44, L_12, L_9, L_15, L_24, L_14, L_32, L_20, L_19, L_8,
L_1, L_6, L_11, L_3, L_2, L_17, L_25, L_13, L_5, L_59, L_55, L_40, and
L_39 from table 1,
 - b) comparing the obtained abundances with at least one reference value,
 - c) determining the diagnosis and/or risk of developing liver cirrhosis of said
20 subject from said comparison.
2. The method of claim 1, wherein said set of bacterial gene clusters of step
 - a) comprises or consists in L_7 and H_12 ; L_7 and L_25 ; L_7 and H_43 ;
L_7 and H_3 ; L_19 and L_7 ; L_4 and L_25 ; L_7 and H_9 ; L_7 and H_34 ;
25 L_4 and L_7 ; L_7 and L_9 ; L_7 and H_22 ; L_7 and H_5 ; L_7 and H_21 ;

L_7 and H_17 ; L_7 and H_11 ; L_42 and L_7 ; L_7 and H_40 ; L_7 and H_24
 ; L_4 and H_12 ; L_7 and H_33 ; L_7 and H_42 ; L_7 and H_38 ; L_7 and
 H_14 ; L_7 and L_15 ; L_9 and L_25 ; L_7 and H_2 ; L_55 and L_7 ; L_7 and
 H_25 ; L_4 and H_33 ; L_7 and H_26 ; L_7 and H_37 ; L_7 and H_10 ; L_4
 5 and H_43 ; L_7 and H_29 ; L_7 and L_13 ; L_7 and L_3 ; L_7 and H_8 ; L_7
 and L_17 ; L_7 and L_12 ; L_7 and H_32 ; L_7 and L_39 ; L_32 and L_7 ; L_7
 and H_13 ; L_7 and L_40 ; L_7 and L_1 ; L_7 and H_36 ; L_7 and H_1 ; L_59
 and L_7 ; L_7 and L_14 ; L_4 and H_26 ; L_7 and L_11 ; L_7 and H_23 ; L_7
 and L_20 ; L_7 and L_2 ; L_4 and H_37 ; L_9 and L_3 or L_4 and H_9 from
 10 table 1.

3. The method of any of the preceding claims, wherein said set of bacterial
 gene clusters of step a) comprises or consists in L_19 and L_7 and H_12;
 L_4 and L_7 and L_25; L_7 and L_25 and H_12; L_4 and L_25 and H_12;
 15 L_19 and L_7 and L_25; L_7 and L_25 and H_43; L_7 and H_43 and H_12;
 L_19 and L_7 and H_43; L_4 and L_7 and H_12; L_7 and L_25 and H_3; L_42
 and L_7 and L_25; L_42 and L_7 and H_12; L_19 and L_7 and H_3; L_4 and
 L_7 and H_43; L_7 and L_25 and H_37; L_7 and L_15 and H_12; L_7 and
 L_25 and H_9; L_55 and L_7 and H_12; L_7 and L_25 and H_24; L_4 and
 20 L_25 and H_43; L_7 and L_25 and H_34; L_4 and L_7 and H_17; L_7 and
 L_25 and L_3; L_19 and L_7 and H_34; L_7 and H_38 and H_12; L_4 and L_7
 and H_1; L_4 and L_17 and L_25; L_7 and H_21 and H_12; L_42 and L_7 and
 H_43; L_4 and L_25 and H_37; L_19 and L_7 and H_9; L_4 and H_33 and
 H_12; L_4 and L_7 and H_11 or L_7 and H_43 and H_9 from table 1.

4. The method of any of the preceding claims, wherein said set of bacterial gene clusters of step a) comprises or consists in L_19 and L_7 and L_25 and H_12; L_4 and L_7 and L_25 and H_12; L_19 and L_7 and L_25 and H_43; L_4 and L_7 and L_25 and H_43; L_4 and L_17 and L_25 and H_12; L_42 and L_7 and L_25 and H_12; L_19 and L_7 and H_43 and H_12; L_42 and L_7 and L_25 and H_43; L_19 and L_7 and L_25 and H_3; L_19 and L_7 and L_25 and H_37; L_4 and L_7 and L_25 and H_17; L_4 and L_7 and L_25 and H_3; L_7 and L_25 and H_43 and H_12 or L_4 and L_42 and L_7 and L_25 from table 1.

10

5. The method of any of the preceding claims, wherein said set of bacterial gene clusters of step a) comprises or consists in H_12 and H_43 and L_19 and L_25 and L_7; H_12 and H_37 and L_19 and L_25 and L_7; H_12 and L_17 and L_19 and L_25 and L_42; H_12 and L_19 and L_25 and L_42 and L_7; H_12 and H_43 and L_25 and L_4 and L_7; H_12 and H_43 and L_25 and L_42 and L_7; H_12 and L_17 and L_25 and L_4 and L_40; H_37 and H_43 and L_19 and L_25 and L_7; H_12 and H_37 and L_25 and L_42 and L_7; H_12 and L_25 and L_4 and L_42 and L_7; H_3 and H_43 and L_19 and L_25 and L_7; H_43 and L_19 and L_25 and L_42 and L_7; H_12 and L_17 and L_25 and L_4 and L_42; H_43 and L_25 and L_4 and L_42 and L_7 or H_12 and H_38 and L_19 and L_25 and L_7 from table 1.

15

6. The method of any of the preceding claims, wherein said set of bacterial gene clusters of step a) comprises or consists in H_12 and H_37 and L_17 and L_19 and L_2 and L_25 or H_12 and H_43 and L_17 and L_19 and L_25 and L_42 from table 1.

20

25

7. The method of any of the preceding claims, wherein said set of bacterial gene clusters of step a) comprises or consists in L_4, L_12, L_9, L_19 and L_17 from table 1; or comprises or consists in L_4, L_42, L_12, L_9 and L_17 from table 1; or comprises or consists in L_4, L_42, L_12, L_9, L_19 and L_17 from table 1.
8. The method of claim 1, wherein said set of bacterial gene clusters of step a) comprises or consists in L_4, L_10, L_42, L_44, L_12, L_9, L_15, L_14, L_32, L_20, L_19, L_8, L_6, L_11, L_3, L_2, and L_17 from table 1.
9. The method of claim 1, wherein said set of bacterial gene clusters of step a) comprises or consists in L_18, L_4, L_7, L_10, L_42, L_44, L_12, L_9, L_15, L_24, L_14, L_32, L_20, L_19, L_8, L_1, L_6, L_11, L_3, L_2, L_17, L_25, L_13, L_5, L_59, L_55, L_40, and L_39 from table 1.
10. The method of claim 1, wherein said set of bacterial gene clusters of step a) comprises or consists in H_16, H_22, and H_42 from table 1, and SEV_4, SEV_13 and SEV_15 from table 9.
11. The method according to anyone of claims 1 to 10, wherein determining the abundance of a bacterial gene cluster is performed by determining the number of copies of at least 5, at least 10, at least 20, at least 30, at least 40 or at least 50 bacterial genes indicated in table 1 from said bacterial gene cluster.

12. An *in vitro* method for the prognosis of liver cirrhosis for a subject, comprising the following steps:

- a) determining from a biological sample of said subject the abundance of each of the bacterial gene clusters L_4, L_10, L_42, L_44, L_12, L_9, L_15, L_14, L_32, L_20, L_19, L_8, L_6, L_11, L_3, L_2, and L_17 from table 1,
- b) comparing the obtained abundances with at least one reference value,
- c) determining the prognosis liver cirrhosis of said subject from said comparison.

13. An *in vitro* method for the prognosis of liver cirrhosis for a subject, comprising the following steps:

- a) determining from a biological sample of said subject the abundance of each of the bacterial gene clusters H_16 , H_19 , H_20 , H_22 , H_33 , H_42 , L_18 , L_19 , L_39 , L_4 , L_42 , L_55 , and L_59 from table 1, SEV_13 , SEV_2 , SEV_22 , SEV_4 , SEV_14 , SEV_15 , SEV_24 and SEV_26 from table 9;
- b) comparing the obtained abundances with at least one reference value,
- c) determining the prognosis liver cirrhosis of said subject from said comparison.

14. The method according to claim 13, wherein the said bacterial gene clusters of step a) are H_16, H_22, and H_42 from table 1, and SEV_4, SEV_13 and SEV_15 from table 9.

15. A method for monitoring the efficacy of a treatment of liver cirrhosis in a subject, comprising the steps of:

- a) determining from a first biological sample of said subject the abundance of each of the bacterial gene clusters L_4, L_10, L_42, L_44, L_12, L_9, L_15, L_14, L_32, L_20, L_19, L_8, L_6, L_11, L_3, L_2, and L_17 from table 1,

- b) determining from a second biological sample of said subject the abundance of each of the bacterial gene clusters L_4, L_10, L_42, L_44, L_12, L_9, L_15, L_14, L_32, L_20, L_19, L_8, L_6, L_11, L_3, L_2, and L_17 from table 1,
- c) comparing the abundance obtained in step a) and the abundance obtained in step b),
- d) determining if said treatment is efficient for said subject from the comparison of step c).

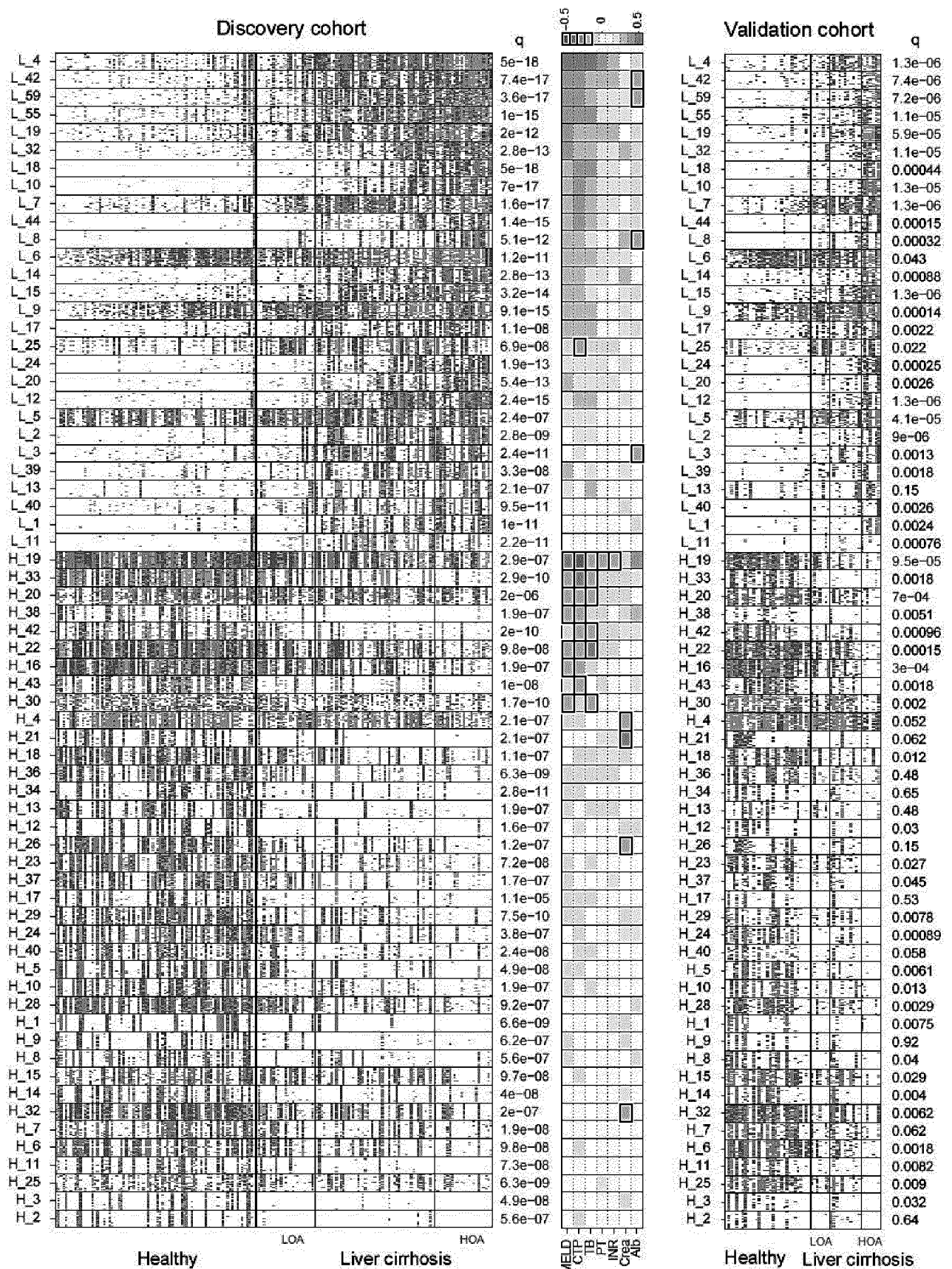
16. A method for monitoring the efficacy of a treatment of liver cirrhosis in a subject, comprising the steps of:

- a) determining from a first biological sample of said subject the abundance of each of the bacterial gene clusters H_16 , H_19 , H_20 , H_22 , H_33 , H_42 , L_18 , L_19 , L_39 , L_4 , L_42 , L_55 , L_59 from table 1 and SEV_13, SEV_2 , SEV_22 , SEV_4 , SEV_14 , SEV_15 , SEV_24 and SEV_26 from table 9;
- b) determining from a second biological sample of said subject the abundance each of the bacterial gene clusters H_16 , H_19 , H_20 , H_22 , H_33 , H_42 , L_18 , L_19 , L_39 , L_4 , L_42 , L_55 and L_59 from table 1 , SEV_13 , SEV_2 , SEV_22 , SEV_4 , SEV_14 , SEV_15 , SEV_24 and SEV_26 from table 9;
- c) comparing the abundance obtained in step a) and the abundance obtained in step b),
- d) determining if said treatment is efficient for said subject from the comparison of step c).

17. The method according to claim 16, wherein the said bacterial gene clusters of steps a) and b) are H_16, H_22, and H_42 from table 1, and SEV_4, SEV_13 and SEV_15 from table 9.

18. A kit for the in vitro the diagnosis of, for assessing whether a subject is at risk of developing, and/or for the prognosis of liver cirrhosis, comprising at least one reagent for the determination of the copy number of at least one gene having a sequence selected from SEQ ID Nos. 1-3700.
- 5

FIGURE 1



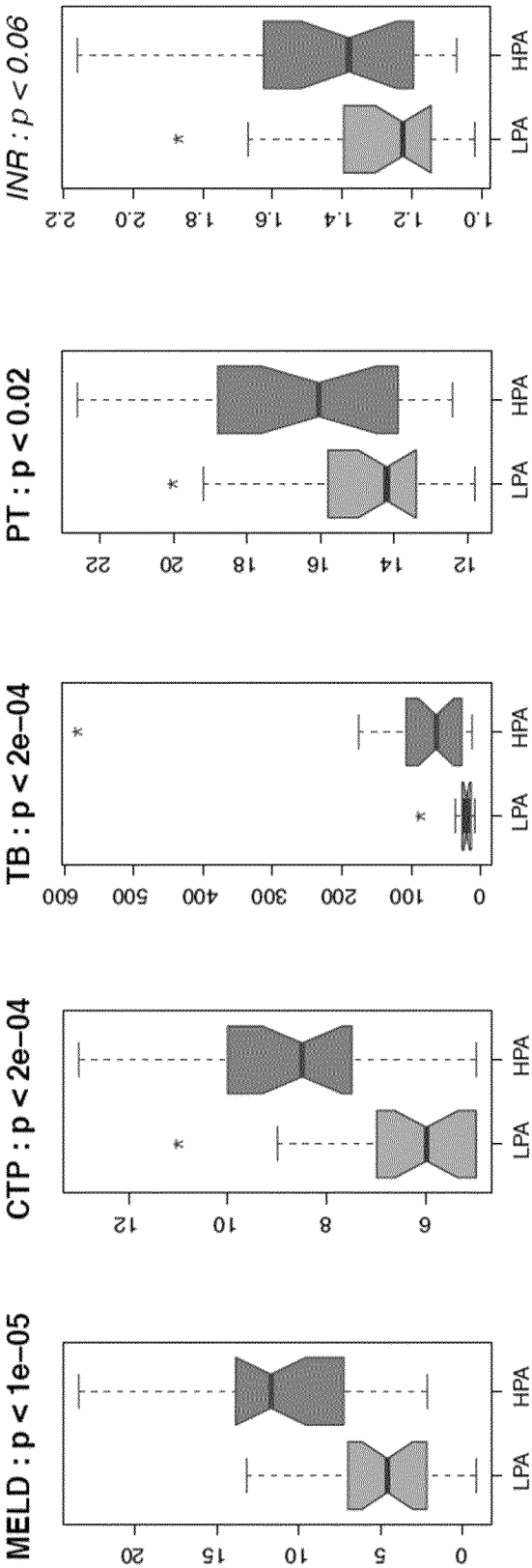


FIGURE 2

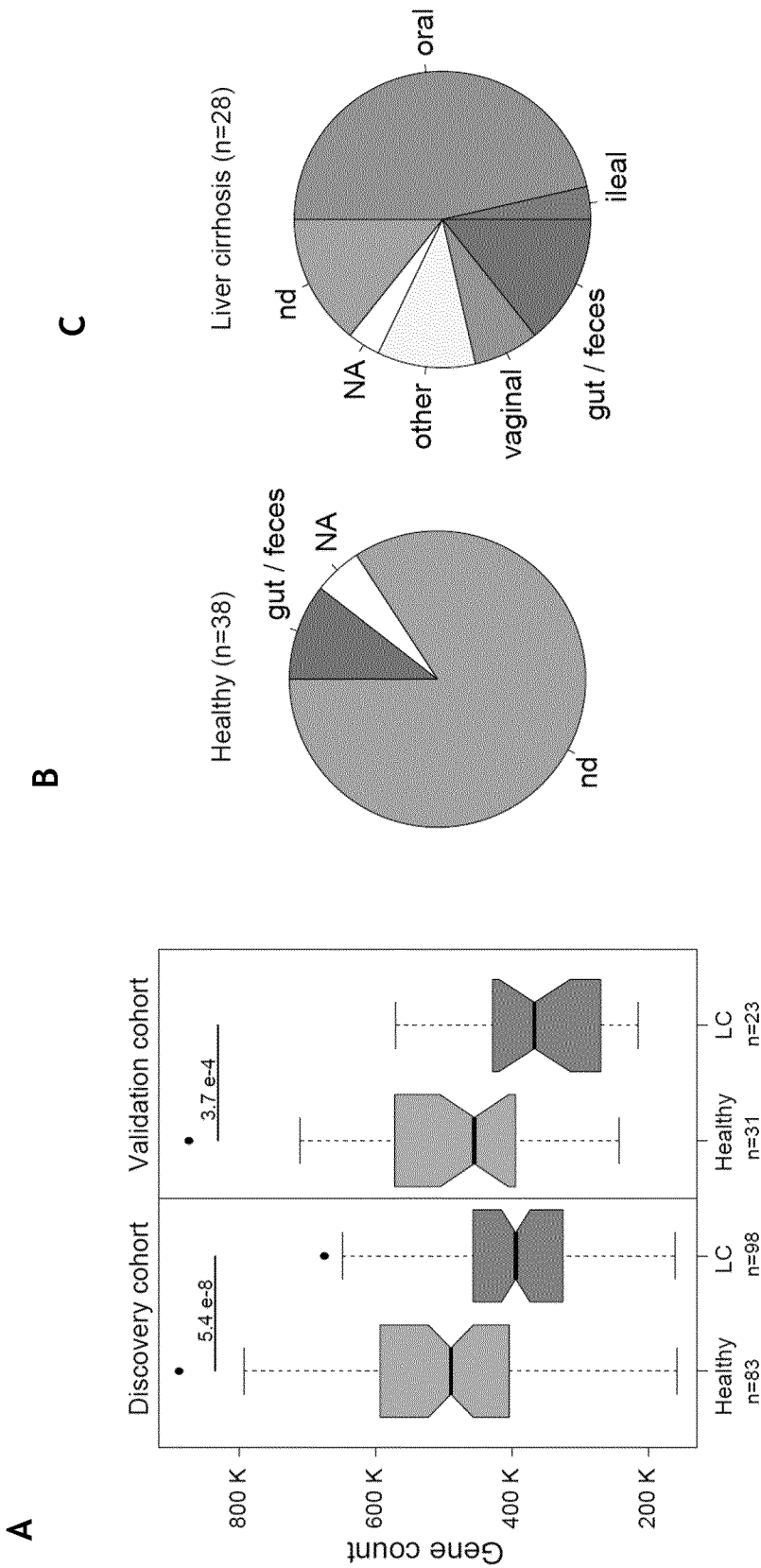


FIGURE 3

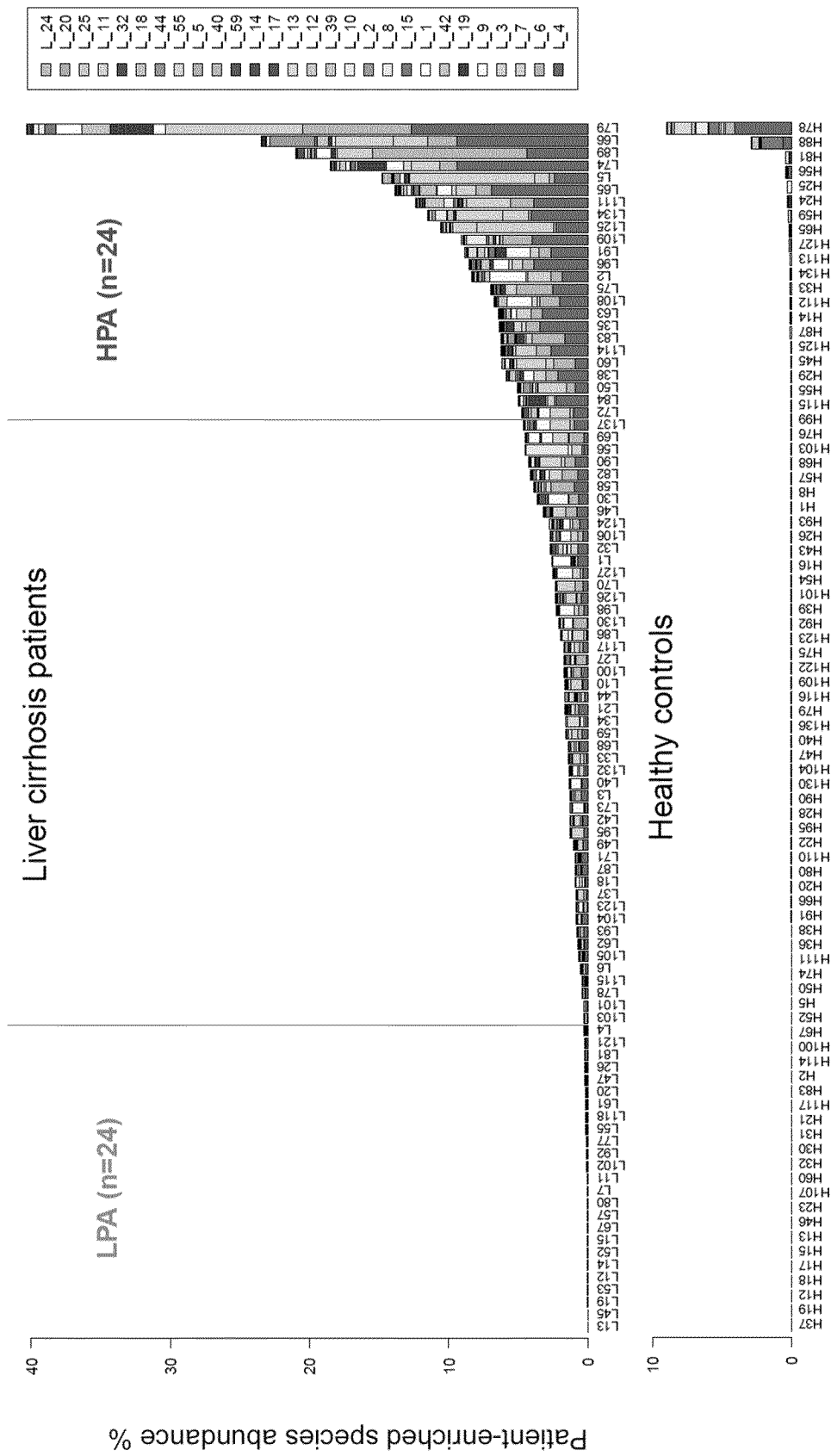


FIGURE 4

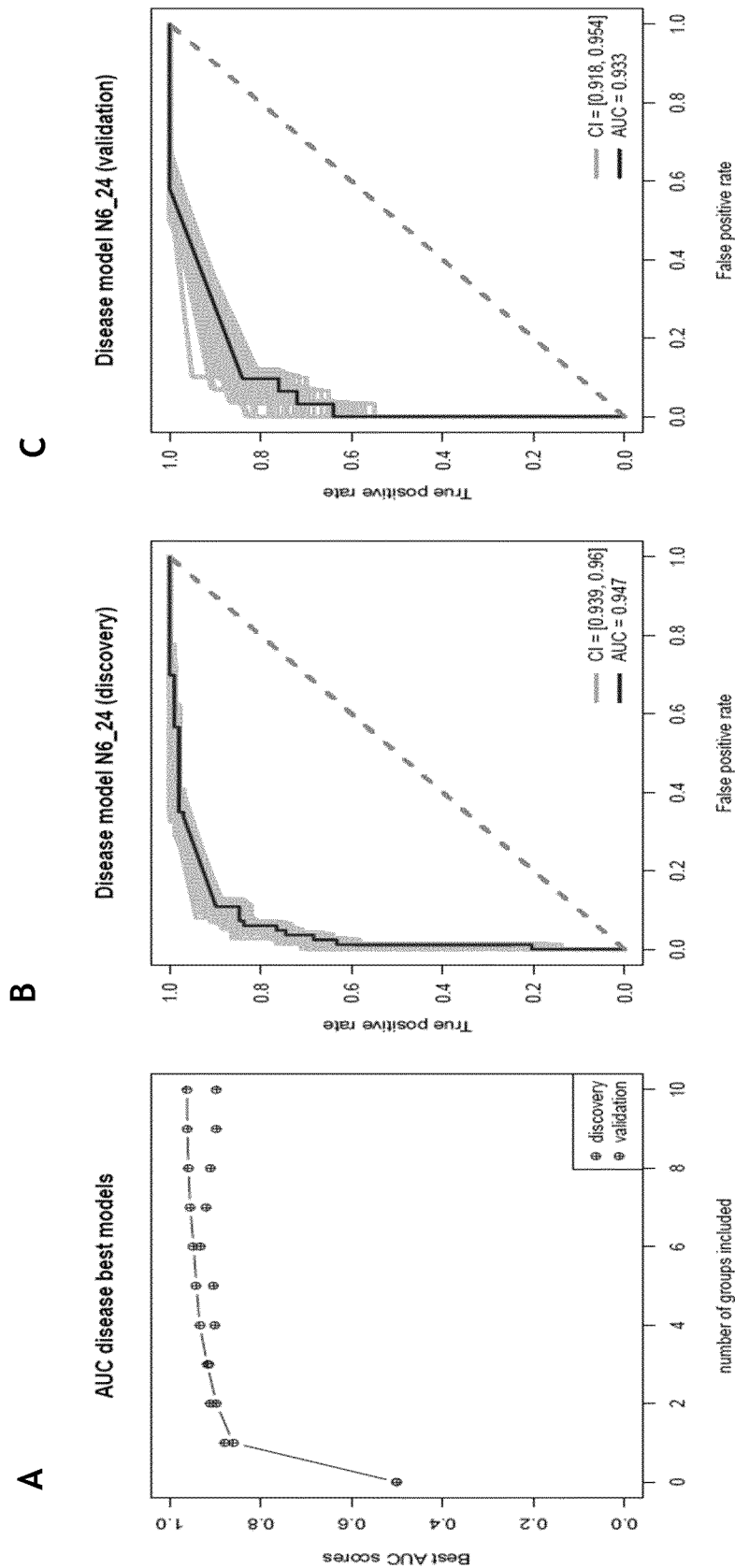


FIGURE 5

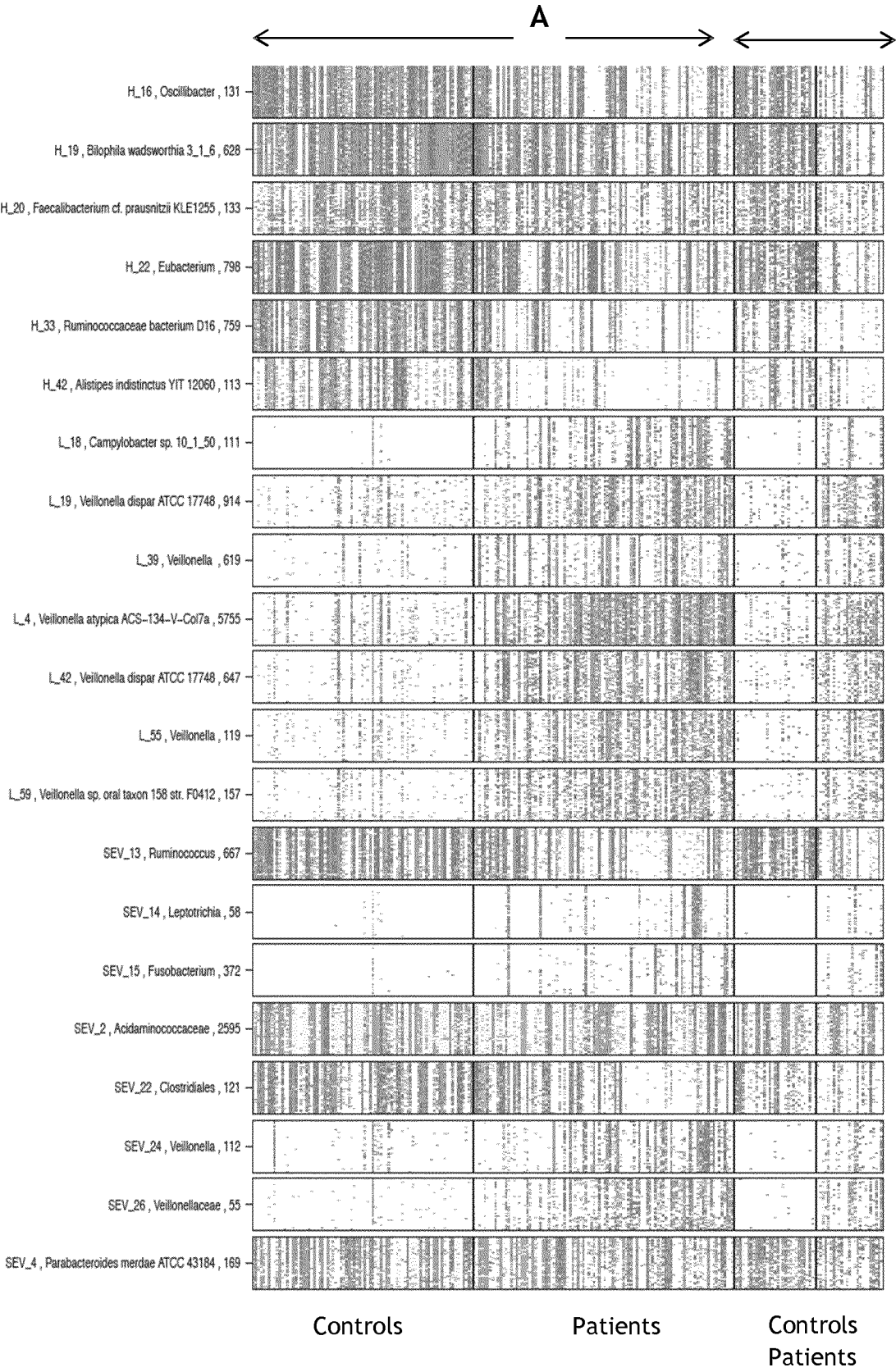


FIGURE 6

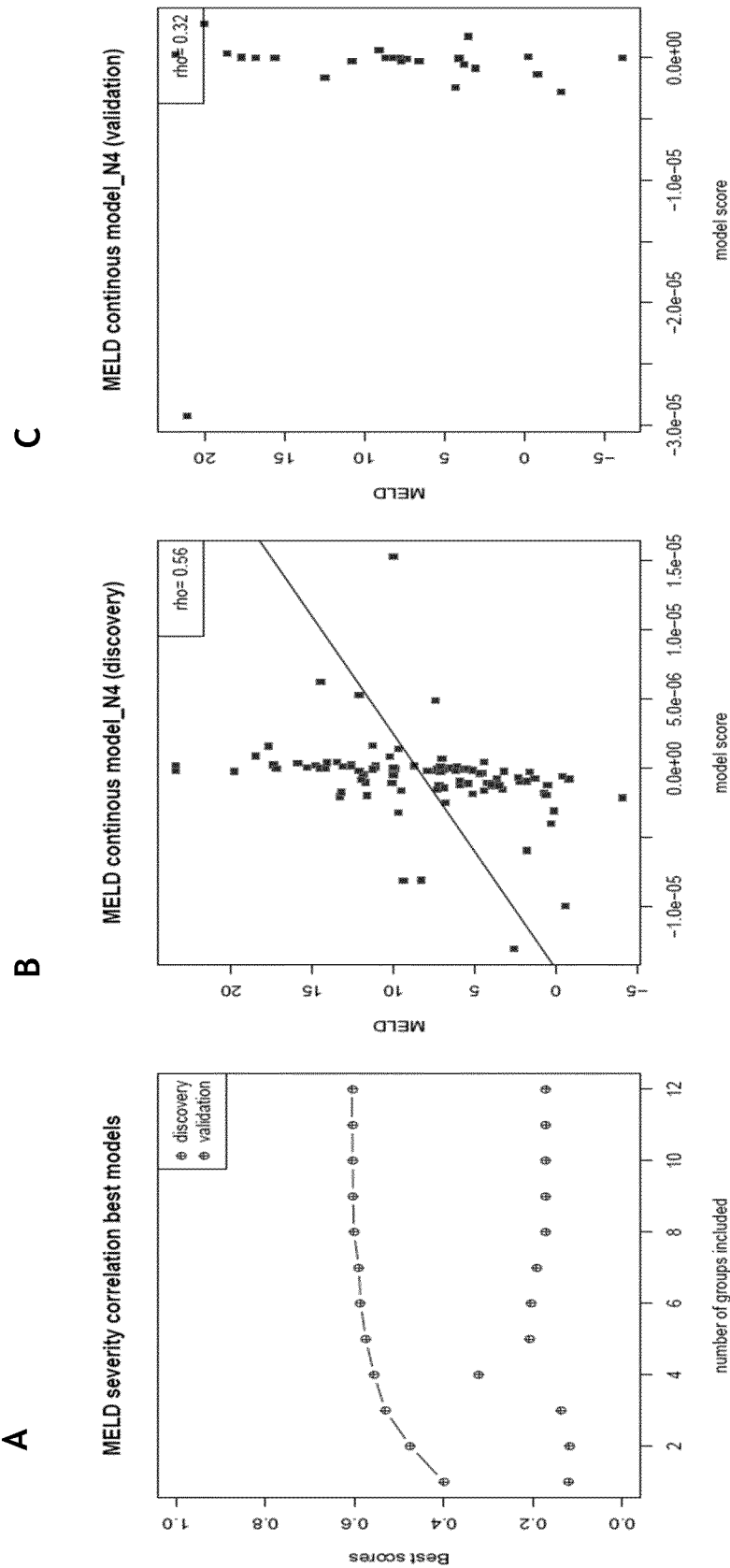


FIGURE 7

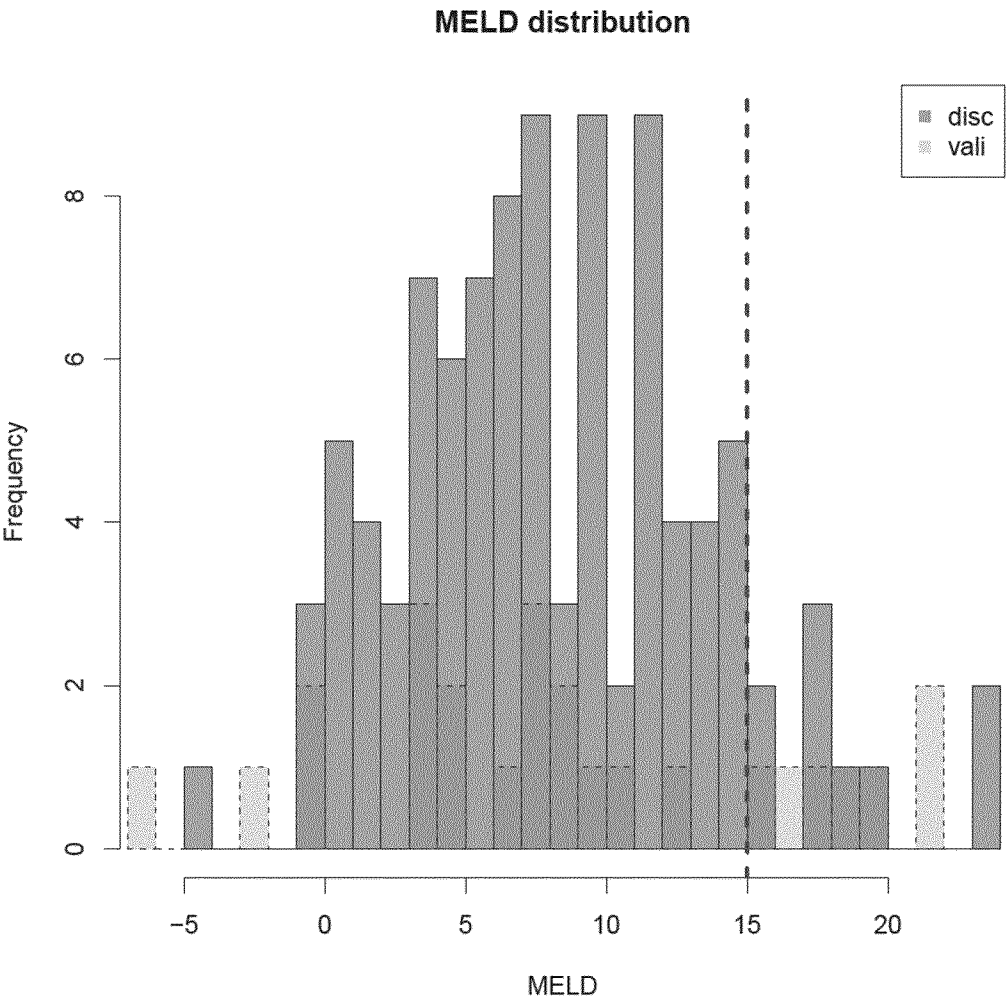


FIGURE 8

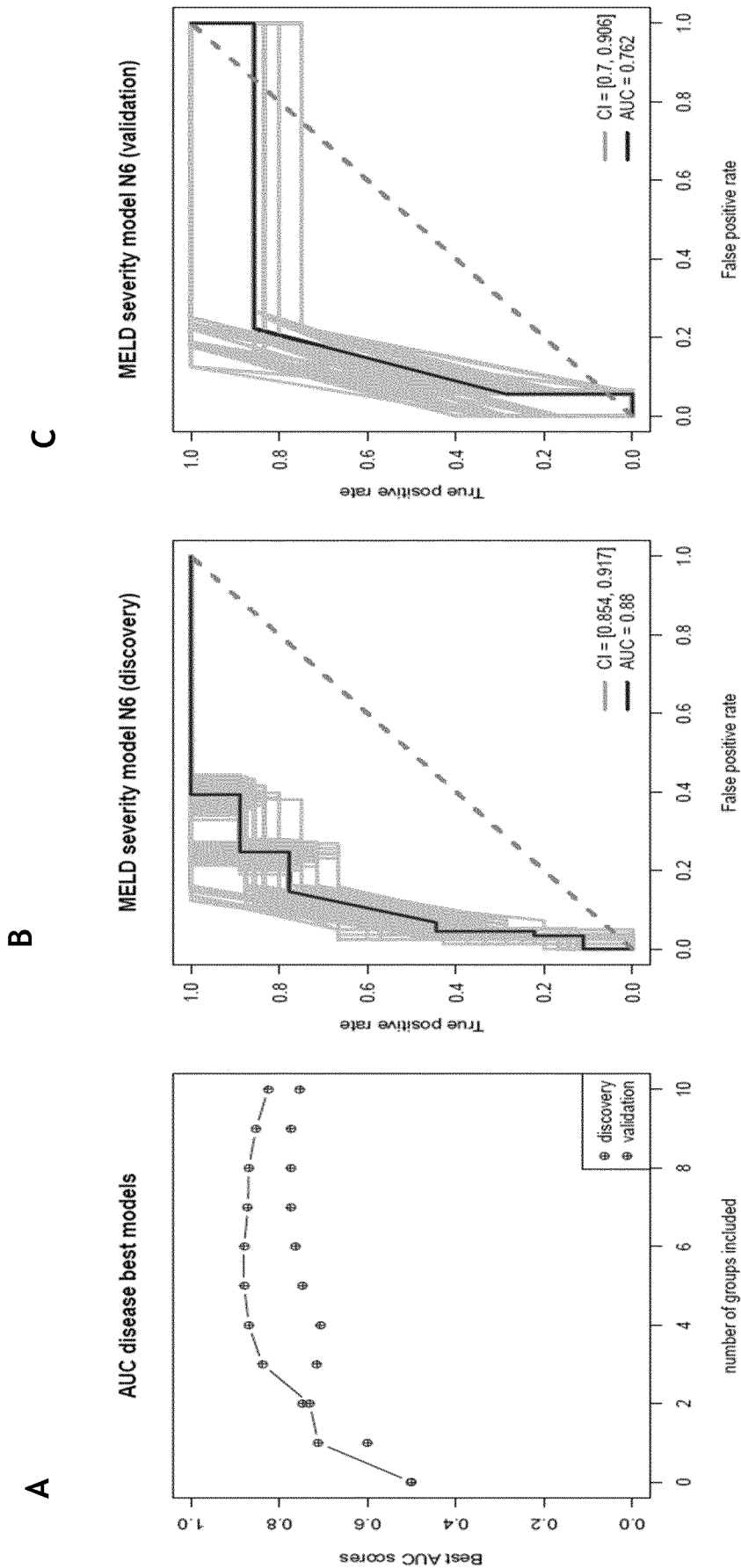


FIGURE 9

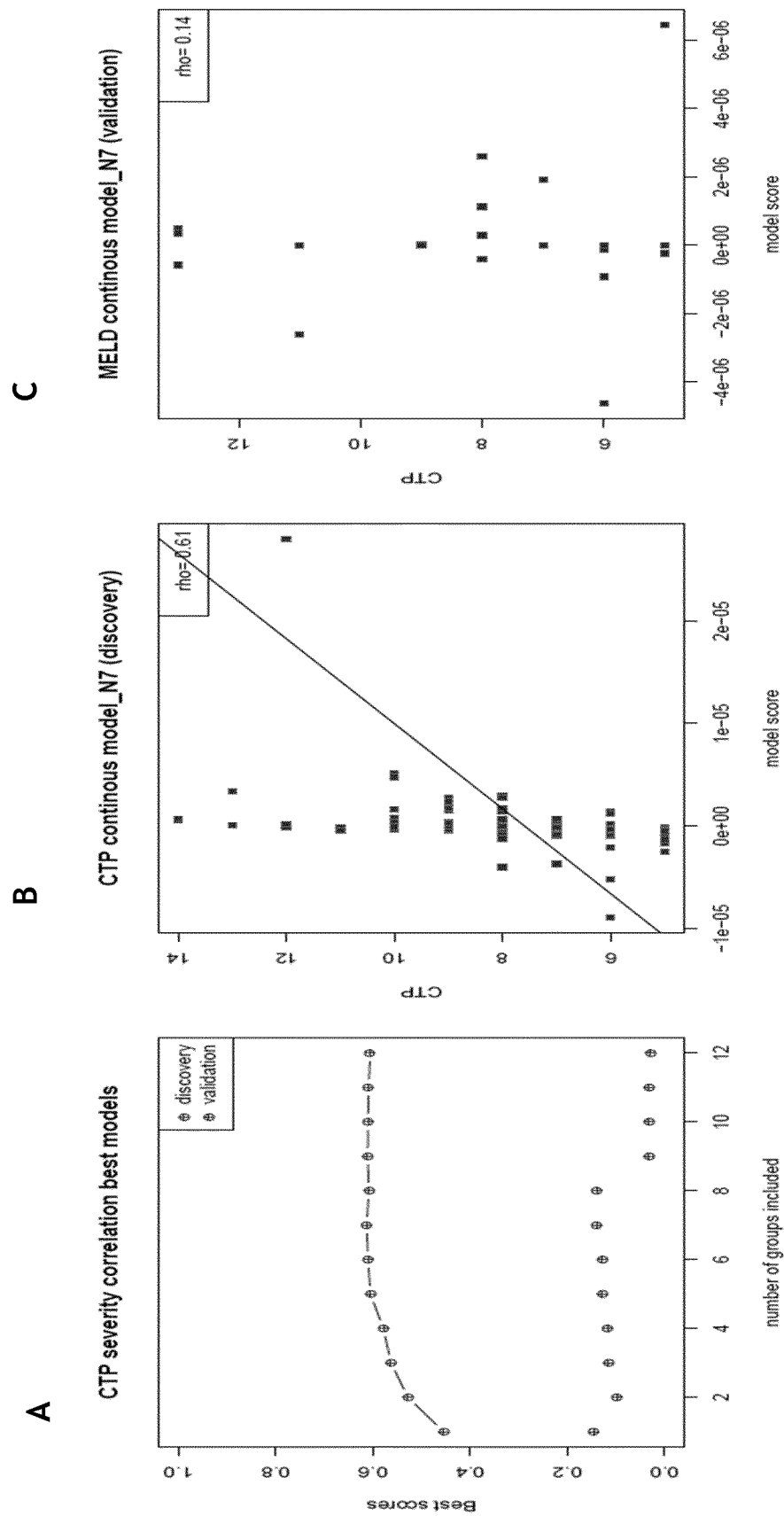
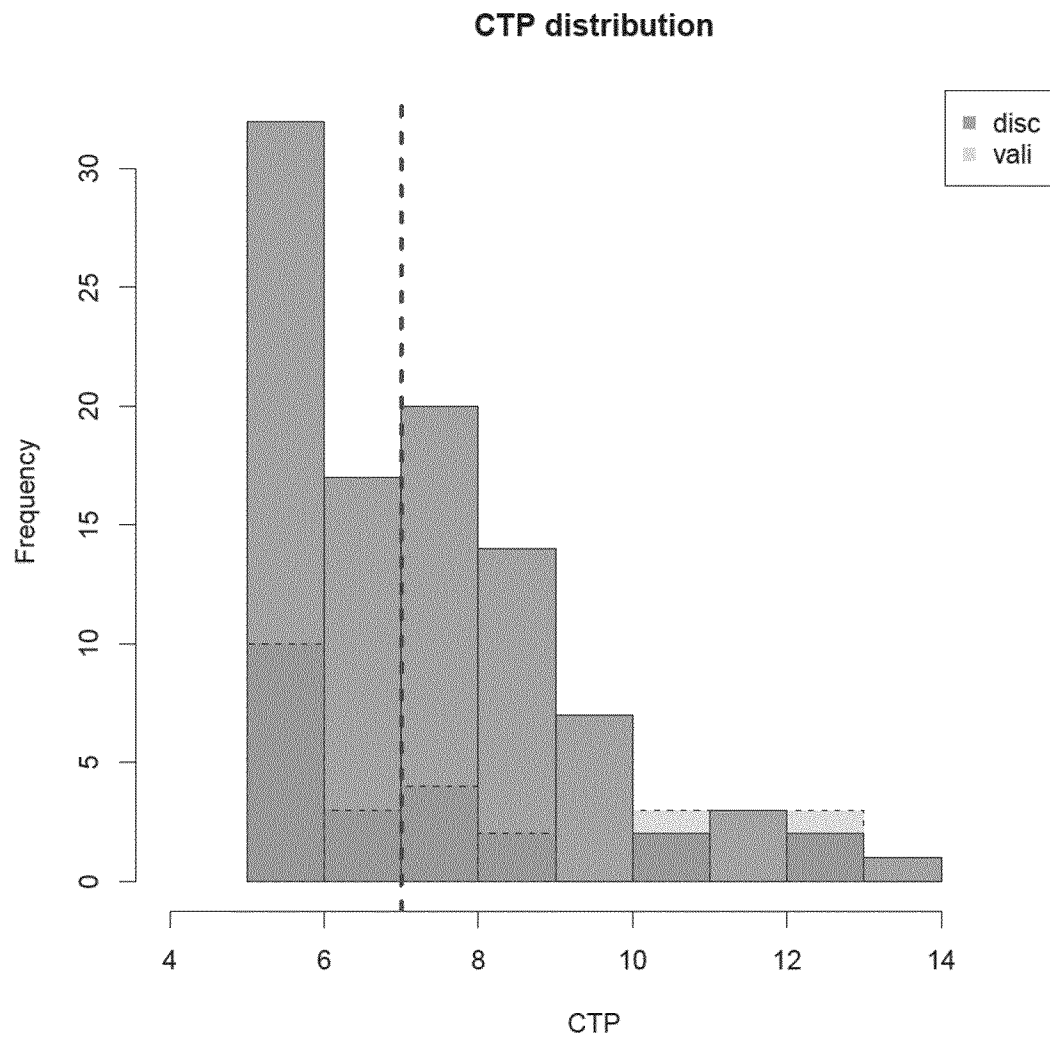


FIGURE 10

**FIGURE 11**

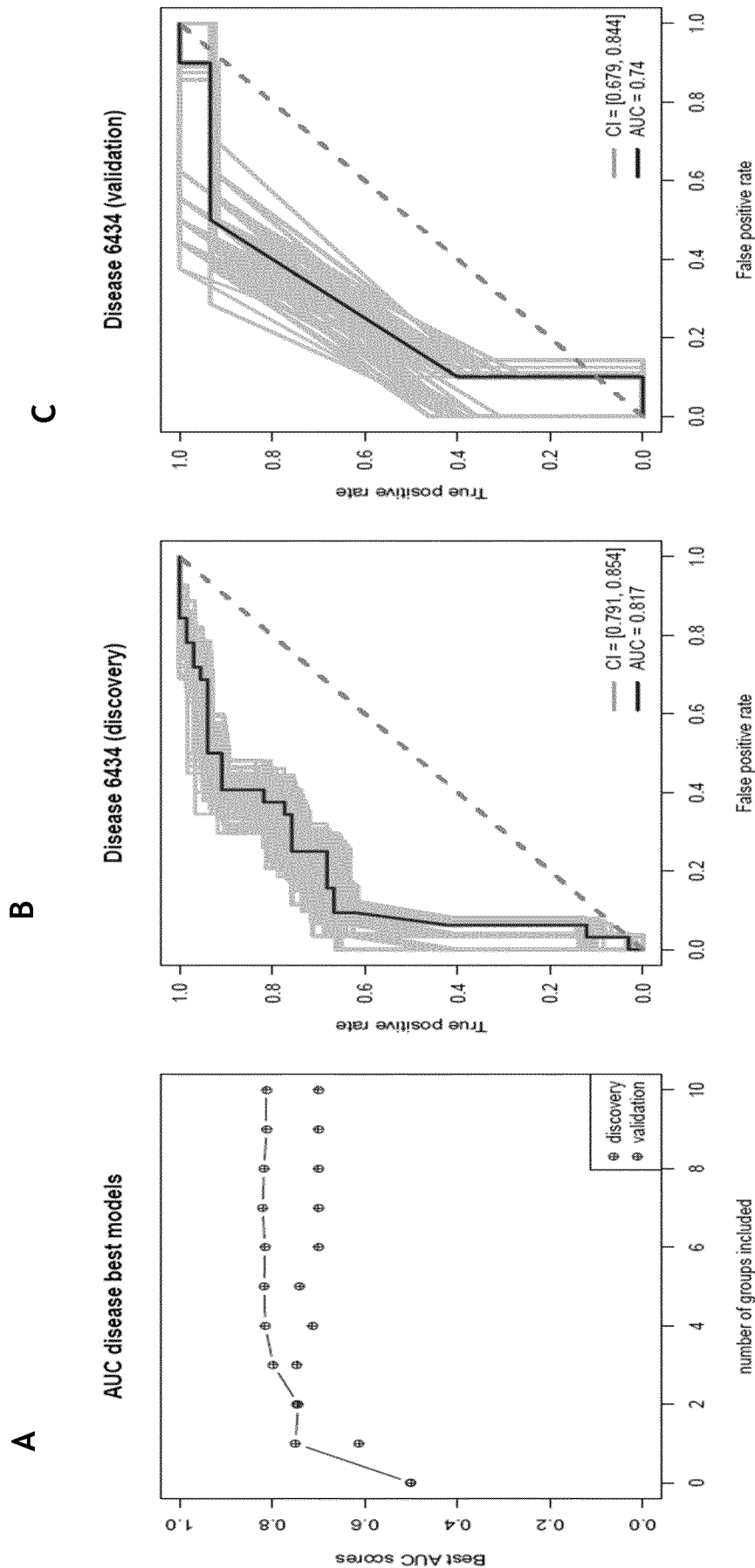


FIGURE 12

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2015/066218A. CLASSIFICATION OF SUBJECT MATTER
INV. C12Q1/68
ADD.

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)
C12Q

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)

EPO-Internal, WPI Data, BIOSIS, EMBASE

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WEI XIAO ET AL: "Abnormal fecal microbiota community and functions in patients with hepatitis B liver cirrhosis as revealed by a metagenomic approach", BMC GASTROENTEROLOGY, BIOMED CENTRAL LTD., LONDON, GB, vol. 13, no. 1, 26 December 2013 (2013-12-26), page 175, XP021172118, ISSN: 1471-230X, DOI: 10.1186/1471-230X-13-175 abstract page 4, column 1 - column 2; figure 1 ----- -/--	1-5,11, 18



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

16 September 2015

Date of mailing of the international search report

18/12/2015

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040,
Fax: (+31-70) 340-3016

Authorized officer

Werner, Andreas

INTERNATIONAL SEARCH REPORT

International application No

PCT/EP2015/066218

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CHEN YANFEI ET AL: "Characterization of fecal microbial communities in patients with liver cirrhosis", HEPATOLOGY, vol. 54, no. 2, 26 June 2011 (2011-06-26), pages 562-572, XP055159513, ISSN: 0270-9139, DOI: 10.1002/hep.24423 abstract paragraph bridging col.1-2; page 567 page 568 page 570, column 1, paragraph 1 -----	1-5,11, 18
A	US 2014/179726 A1 (BAJAJ JASMOHAN [US] ET AL) 26 June 2014 (2014-06-26) paragraph [0010]; claims 1-8 paragraph [0069]; claim 9; figure 1; table 3 -----	1-5,11, 18
A	BAJAJ J S ET AL: "Linkage of gut microbiome with cognition in hepatic encephalopathy", AMERICAN JOURNAL OF PHYSIOLOGY: GASTROINTESTINAL AND LIVER PHYSIOLOGY, AMERICAN PHYSIOLOGICAL SOCIETY, US, vol. 302, no. 1, 1 January 2012 (2012-01-01), pages G168-G175, XP002692034, ISSN: 0193-1857, DOI: 10.1152/AJPGI.00190.2011 [retrieved on 2011-09-22] abstract table 2 Materials and methods -----	1-5,11, 18
A	WO 2014/060538 A1 (AGRONOMIQUE INST NAT RECH [FR]; NOVO NORDISK AS [DK]) 24 April 2014 (2014-04-24) the whole document -----	1-5,11, 18
A	EMMANUELLE LE CHATELIER ET AL: "Richness of human gut microbiome correlates with metabolic markers", NATURE, vol. 500, no. 7464, 28 August 2013 (2013-08-28), pages 541-546, XP055087499, ISSN: 0028-0836, DOI: 10.1038/nature12506 cited in the application the whole document -----	1-5,11, 18

INTERNATIONAL SEARCH REPORT

International application No.
PCT/EP2015/066218

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.:
because they relate to subject matter not required to be searched by this Authority, namely:
2. ☒ Claims Nos.: 1-17(partially)
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:
see FURTHER INFORMATION sheet PCT/ISA/210
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).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

see additional sheet

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.
2. ☐ As all searchable claims could be searched without effort justifying an additional fees, this Authority did not invite payment of additional fees.
3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:
4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:
1-5, 11, 18(all partially)

Remark on Protest

- ☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.
- ☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.
- ☐ No protest accompanied the payment of additional search fees.

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

This International Searching Authority found multiple (groups of) inventions in this international application, as follows:

1. claims: 1-5, 11, 18(all partially)

Methods for the diagnosis of liver cirrhosis in a subject and/or for assessing whether a subject is at risk of developing liver cirrhosis, comprising determining from a biological sample of said subject the abundance of each of the bacterial gene clusters of a set of bacterial gene clusters, wherein each of said clusters consist in 50 non-redundant and covariant bacterial genes belonging to the same genome, defined in table 1, wherein said set of bacterial gene clusters comprise at least 2 bacterial gene clusters chosen in the group of claim 1, preferably wherein said set of bacterial gene clusters comprises or consists in L 7 and H12 listed in claim 2, or preferred embodiments thereof listed in claims 3-5;

Kits comprising at least one reagent for the determination of the copy number of at least one gene having a sequence selected from the corresponding SEQ ID NOs. as listed in table 6 on p.29-32, in this instance SEQ ID NO:3151-3200 and SEQ ID NO:151-200.

2-57. claims: 1-6, 11, 18(all partially)

Methods for the diagnosis of liver cirrhosis in a subject and/or for assessing whether a subject is at risk of developing liver cirrhosis, comprising determining from a biological sample of said subject the abundance of each of the bacterial gene clusters of a set of bacterial gene clusters, wherein each of said clusters consist in 50 non-redundant and covariant bacterial genes belonging to the same genome, defined in table 1, wherein said set of bacterial gene clusters comprise at least 2 bacterial gene clusters chosen in the group of claim 1, preferably wherein said set of bacterial gene clusters comprises or consists in one of the further 56 sets listed in claim 2 (in the order of appearance), or preferred embodiments thereof listed in claims 3-6;

Kits comprising at least one reagent for the determination of the copy number of at least one gene having a sequence selected from the corresponding SEQ ID NOs. as listed in table 6 on p.31-33 for each of said sets of bacterial gene clusters.

58. claims: 7-9, 12, 15(completely); 1, 11, 18(partially)

Methods for the diagnosis of liver cirrhosis in a subject and/or for assessing whether a subject is at risk of developing liver cirrhosis, comprising determining from a biological sample of said subject the abundance of each of the bacterial gene clusters of a set of bacterial gene

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

clusters, wherein each of said clusters consist in 50 non-redundant and covariant bacterial genes belonging to the same genome, defined in table 1, wherein said set of bacterial gene clusters comprise at least 2 bacterial gene clusters chosen in the group of claim 1, wherein said set of bacterial gene clusters comprises or consists in one of the sets listed in claims 7-9;

In vitro methods for the prognosis of liver cirrhosis for a subject, comprising determining from a biological sample of said subject the abundance of each of the bacterial gene clusters L_4, L_10, L_42, L_44, L_12, L_9, L_15, L_14, L_32, L_20, L_19, L_8, L_6, L_11, L_3, L_2, and L_17 from table 1; Methods thereof for monitoring the efficacy of a treatment of liver cirrhosis in a subject;

Kits comprising at least one reagent for the determination of the copy number of at least one gene having a sequence selected from the corresponding SEQ ID NOs. as listed in table 6 on p.31-33 for said sets of bacterial gene clusters.

59. claims: 10, 13, 14, 16, 17(completely); 1, 18(partially)

Methods for the diagnosis of liver cirrhosis in a subject and/or for assessing whether a subject is at risk of developing liver cirrhosis, comprising determining from a biological sample of said subject the abundance of each of the bacterial gene clusters of a set of bacterial gene clusters, wherein each of said clusters consist in 50 non-redundant and covariant bacterial genes belonging to the same genome, defined in table 1, wherein said set of bacterial gene clusters comprise at least 2 bacterial gene clusters chosen in the group of claim 1, preferably wherein said bacterial gene clusters are H_16, H_22 and H_42 from table 1 and SEV_4, SEV_13 and SEV_15 from table 9;

In vitro methods for the prognosis of liver cirrhosis for a subject, comprising determining from a biological sample of said subject the abundance of each of the bacterial gene clusters H_16, H_19, H_20, H_22, H_33, H_42, L_18, L_19, L_39, L_4, L_42, L_55, L_59 from table 1, [and] SEV_13, SEV_2, SEV_22, SEV_4, SEV_14, SEV_15, SEV_24 and SEV_26 from table 9;

Methods thereof for monitoring the efficacy of a treatment of liver cirrhosis in a subject;

Kits comprising at least one reagent for the determination of the copy number of at least one gene having a sequence selected from the corresponding SEQ ID NOs. as listed in table 6 on p.31-33 and table 10 on p.49.

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

Continuation of Box II.2

Claims Nos.: 1-17(partially)

The set of bacterial gene clusters of claim 1 refers to any combination of at least 2 gene clusters chosen in the group of 66 different bacterial gene clusters listed in claim 1. This gives rise to an extremely large number of possible combinations, whereas support and disclosure only seems to be present for the specific combinations of 2, 3, 4, 5, 6, 12, 17 or 28 bacterial gene clusters from tables 1 and 9 listed in dependent claims 2-10, as reflected in tables 2, 3, 4, 5 and corresponding passages of the description.

Moreover, claims 1-17 refer to the bacterial gene clusters of table 1 and/or table 9, which are identified only by internal designations H_..., L_... or SEV_... that have no generally recognized meaning in the field. The skilled person receives no teaching from said tables which sequences are determinative of each of said gene clusters. While claim 1 requires that "each of said clusters consist[s] in 50 non-redundant and covariant bacterial genes belonging to the same genome, defined in Table[s] 1", this definition merely amounts to an unclear desired result to be achieved. Table 1 does not even define any genomes, other than listing said unclear internal designations H_..., and L_... . The definition on p.11, last paragraph, defines "bacterial gene cluster" as any set of bacterial genes, preferably covariant bacterial genes, wherein abundance levels in samples from distinct subjects have been determined to be statistically linked. It constitutes undue burden for a skilled person to establish whether any given bacterial gene is part of one of said unclearly defined bacterial gene clusters. It also remains open to speculation which sequences would have to be quantified in order to determine the "abundance" of any of said gene clusters, or else how said bacterial gene clusters could be quantified otherwise. While the skilled person could in principle look up each of said bacterial gene clusters in tables 1 or 9, and then look up corresponding sequences in respective tables 6 or 10, the claims are not so limited, and their definition is contrary to the requirement that essential features have to be clear from the wording of the claim alone. Thus the bacterial gene clusters of the present claims and methods of their quantification are insufficiently defined and unclear.

The large number of possible combinations and unclear definition results in a lack of disclosure, support and clarity (Art. 5, 6 PCT) of said claims 1-17, to such an extent that this non-compliance with the substantive provisions was taken into consideration when determining the extent of the search of claims (PCT Guidelines 9.19 and 9.23).

Consequently, the scope has been restricted to the specific combinations of 2, 3, 4, 5, 6, 12, 17 or 28 bacterial gene clusters chosen from tables 1 and 9 as listed in dependent claims 2-10, wherein each of clusters H_..., L_... or SEV_... is defined by the 50 non-redundant and covariant bacterial genes belonging to the same genome which are respectively identified in table 6 (p.31-33) and table 10 (p.49), and which represent positively disclosed groups of sequences selected from SEQ ID NO:1-3300 and SEQ ID NO:3301-3700, respectively, as reflected in product claim 18.

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

The applicant's attention is drawn to the fact that claims relating to inventions in respect of which no international search report has been established need not be the subject of an international preliminary examination (Rule 66.1(e) PCT). The applicant is advised that the EPO policy when acting as an International Preliminary Examining Authority is normally not to carry out a preliminary examination on matter which has not been searched. This is the case irrespective of whether or not the claims are amended following receipt of the search report or during any Chapter II procedure. If the application proceeds into the regional phase before the EPO, the applicant is reminded that a search may be carried out during examination before the EPO (see EPO Guidelines C-IV, 7.2), should the problems which led to the Article 17(2) declaration be overcome.

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/EP2015/066218

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2014179726 A1	26-06-2014	US 2014179726 A1	26-06-2014
		WO 2012159023 A2	22-11-2012

WO 2014060538 A1	24-04-2014	EP 2909336 A1	26-08-2015
		US 2015267249 A1	24-09-2015
		WO 2014060538 A1	24-04-2014
