ARTICLE

doi:10.1038/nature13568

Alterations of the human gut microbiome in liver cirrhosis

Nan Qin[1,2]*, Fengling Yang[1]*, Ang Li[1]*, Edi Prifti[3]*, Yanfei Chen[1]*, Li Shao[1,2]*, Jing Guo[1], Emmanuelle Le Chatelier[3], Jian Yao[1,2], Lingjiao Wu[1], Jiawei Zhou[1], Shujun Ni[1], Lin Liu[1], Nicolas Pons[3], Jean Michel Batto[3], Sean P. Kennedy[3], Pierre Leonard[3], Chunhui Yuan[1], Wenchao Ding[1], Yuanting Chen[1], Xinjun Hu[1], Beiwen Zheng[1,2], Guirong Qian[1], Wei Xu[1], S. Dusko Ehrlich[3,4], Shusen Zheng[2,5] & Lanjuan Li[1,2]

Liver cirrhosis occurs as a consequence of many chronic liver diseases that are prevalent worldwide. Here we characterize the gut microbiome in liver cirrhosis by comparing 98 patients and 83 healthy control individuals. We build a reference gene set for the cohort containing 2.69 million genes, 36.1% of which are novel. Quantitative metagenomics reveals 75,245 genes that differ in abundance between the patients and healthy individuals (false discovery rate < 0.0001) and can be grouped into 66 clusters representing cognate bacterial species; 28 are enriched in patients and 38 in control individuals. Most (54%) of the patient-enriched, taxonomically assigned species are of buccal origin, suggesting an invasion of the gut from the mouth in liver cirrhosis. Biomarkers specific to liver cirrhosis at gene and function levels are revealed by a comparison with those for type 2 diabetes and inflammatory bowel disease. On the basis of only 15 biomarkers, a highly accurate patient discrimination index is created and validated on an independent cohort. Thus microbiota-targeted biomarkers may be a powerful tool for diagnosis of different diseases.

Cirrhosis is an advanced liver disease resulting from acute or chronic liver injury, including alcohol abuse, obesity and hepatitis virus infection. The prognosis for patients with decompensated liver cirrhosis is poor, and they frequently require liver transplantation[1]. The liver interacts directly with the gut through the hepatic portal and bile secretion[2] systems. Enteric dysbiosis, especially the translocation of bacteria[3] and their products[4,5] across the gut epithelial barrier, is involved in the progression of liver cirrhosis. However, the phylogenetic and functional composition changes in the human gut microbiota that are related to this progression remain obscure[5]. Some studies have revealed that alterations in the gut microbiota are important in complications of end-stage liver cirrhosis[6] (such as spontaneous bacterial peritonitis[7] and hepatic encephalopathy[8]) and the induction and promotion of liver damage in early-stage liver disease[9] (such as alcoholic liver disease[10] and non-alcoholic fatty liver disease[11]), but definitive associations of gut microbiota and liver pathology in humans are still lacking[12]. Studies of patients with liver cirrhosis[13] and of mouse models for alcoholic liver disease[10] have revealed a similar and substantial alteration in the gut microbiota, as measured by sequencing of 16S ribosomal RNA genes. How these phylogenetic alterations relate to changes in the functioning of this ecosystem is, however, unclear.

The role of gut microbiota in human health and disease[14] has recently received considerable attention. Chronic diseases, such as obesity[15–18], inflammatory bowel disease (IBD)[19,20], diabetes mellitus[21,22], metabolic syndrome[23], symptomatic atherosclerosis[24] and non-alcoholic fatty liver disease[10], have been associated with gut microbiota. The US National Institutes of Health Human Microbiome Project (HMP) generated a large data set from different anatomical sites among 242 healthy individuals and created a large human microbiome gene resource[25,26]. Quantitative metagenomics analysis[27,28] developed by the MetaHIT consortium revealed a significant loss of gut microbial richness associated with the risk of metabolic syndrome related co-morbidities. Here we apply a similar analysis to contrast microbiota from 123 patients with liver cirrhosis and 114 healthy counterparts of Han Chinese origin.

Gene catalogue of gut microbes

We constructed a gene catalogue from 98 Chinese patients with liver cirrhosis and 83 healthy Chinese control individuals (Supplementary Table 1) using the methodology developed by MetaHIT. The liver cirrhosis catalogue contained 2,688,468 non-redundant open reading frames (ORFs). We compared it with three other gut microbial catalogues: MetaHIT[29], HMP[25] and T2D[22]. 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 catalogues (Extended Data Fig. 1a). The liver cirrhosis catalogue, MetaHIT, HMP and T2D gene sets contained 794,647, 1,419,517, 2,620,096 and 623,570 unique genes, respectively. Genes from the liver cirrhosis, 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 (Extended Data Fig. 1b).

Phylogenetic profiles of gut microbes

The sequencing reads (36.67%) were aligned against 4,398 reference genomes from the National Center for Biotechnology Information and the HMP (Supplementary Table 2). After correction for population stratification that could be related to non-liver cirrhosis-related factors (see Methods), the relative abundances of phylum, class, order, family, genus and species between liver cirrhosis and control groups were compared (Extended Data Fig. 2). Phylotypes with a median relative abundance larger than 0.01% of the total abundance in either the healthy control

[1]State Key Laboratory for Diagnosis and Treatment of Infectious Disease, The First Affiliated Hospital, College of Medicine, Zhejiang University, 310003 Hangzhou, China. [2]Collaborative Innovation Center for Diagnosis and Treatment of Infectious Diseases, Zhejiang University, 310003 Hangzhou, China. [3]Metagenopolis, Institut National de la Recherche Agronomique, 78350 Jouy en Josas, France. [4]King's College London, Centre for Host-Microbiome Interactions, Dental Institute Central Office, Guy's Hospital, London Bridge, London SE1 9RT, UK. [5]Key Laboratory of Combined Multi-organ Transplantation, Ministry of Public Health, the First Affiliated Hospital, Zhejiang University, 310003 Hangzhou, China.
*These authors contributed equally to this work.

©2014 Macmillan Publishers Limited. All rights reserved

group or the liver cirrhosis group were included for comparison. At the phylum level, Bacteroidetes and Firmicutes dominated the faecal microbial communities of both groups (Fig. 1a, b). Compared with healthy controls, patients with liver cirrhosis had fewer Bacteroidetes (Fig. 1a), but higher levels of Proteobacteria and Fusobacteria (Fig. 1b).

At the genus level, *Bacteroides* was the dominant phylotype in both groups, but was significantly decreased in the liver cirrhosis group. Of the remaining genera, *Veillonella*, *Streptococcus*, *Clostridium* and *Prevotella* were enriched in the liver cirrhosis group, while *Eubacterium* and *Alistipes* were dominant in the healthy controls (Fig. 1a, b). The most abundant species in both liver cirrhosis and the healthy control groups were primarily from the *Bacteroides* genus. Of the 20 species that increased the most in abundance in the liver cirrhosis group, four were *Streptococcus* spp. and six were *Veillonella* spp., suggesting that the two genera might play an important role in liver cirrhosis. Of the species that decreased the most in abundance in the liver cirrhosis group, 12 were Bacteroidetes and seven were Firmicutes, specifically from the order Clostridiales.

Gut microbial species associated with cirrhosis

Our investigation included two phases. The first was discovery, where we compared 98 patients with liver cirrhosis and 83 healthy controls. The second was validation, with additional 25 patients and 31 controls. In the discovery phase, a Wilcoxon rank-sum test corrected for multiple testing by the Benjamini and Hochberg method was used to identify differentially abundant genes in patients and controls. At a stringent threshold (false discovery rate (FDR) < 0.0001), 75,245 genes were found: 49,830 were more abundant in the patients and 25,415 in the controls (Methods). Patients and controls could be clearly separated by principal component analysis based on the 75,245 genes; this was confirmed with the validation samples (Supplementary Table 3 and Extended Data Fig. 1c).

To explore further the microbial genes associated with liver cirrhosis we grouped them into clusters, denoted metagenomic species (MGS) here, on the basis of their abundance profiles[27,30]. Of the 66 MGS, 38 and 28 were enriched in healthy individuals and patients, respectively. The significantly different abundance distribution between healthy and liver cirrhosis subjects is shown in Fig. 2 and Supplementary Table 4. 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 $q < 10^{-3}$ in the Chinese cohort; 26 of these (78.8%) were similarly correlated in a Danish cohort (Extended Data Fig. 3). These observations indicate that gut communities of bacteria in healthy individuals across continents may be largely similar. Furthermore, gene richness was much lower in patients with liver cirrhosis than in healthy individuals (on average 389,000 and 497,000 genes, respectively; Supplementary Table 5 and Extended Data Fig. 4, top left). Interestingly, among the species enriched in healthy Chinese, were *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 three Lachnospiraceae and five Ruminococcaceae enriched in healthy individuals. A lower abundance of these species in patients with liver cirrhosis indicates that these individuals 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 (Supplementary Table 4). However, the small intestine also harbours such species[32] and small-intestinal bacterial

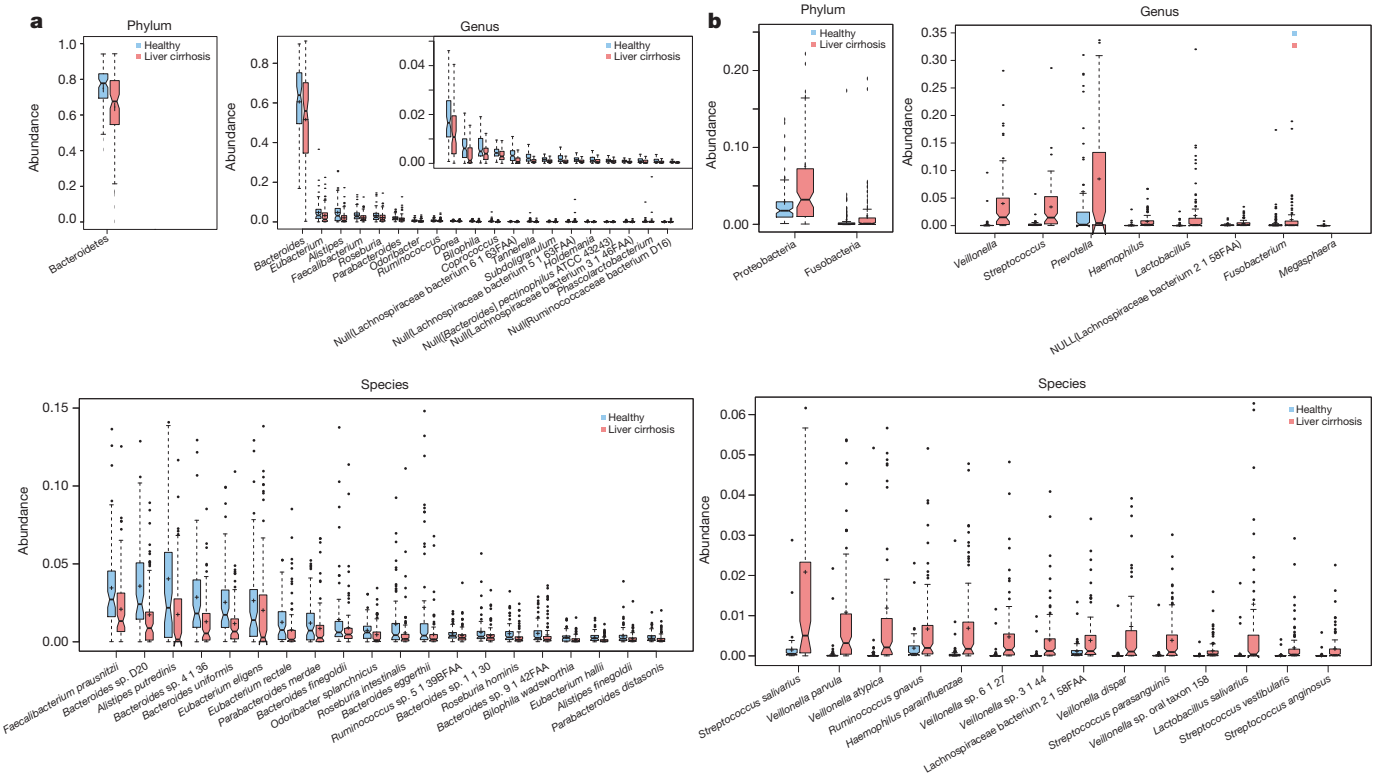


Figure 1 | Differentially abundant phyla in patients ($n = 98$) and healthy individuals ($n = 83$). The phylotypes decreased (**a**) and increased (**b**) in patients with liver cirrhosis at the phylum, genus and species levels. Blue and red represent healthy controls and patients with liver cirrhosis, respectively. Only the 20 most abundant species in each group are shown for clarity. The phylotypes with median relative abundances greater than 0.01% of total abundance in either the healthy control group or the liver cirrhosis group are included (FDR < 0.01, Wilcoxon rank-sum test corrected by the Benjamini and Hochberg method). The boxes represent the interquartile range (IQR), from the first and third quartiles, and the inside line represents the median. The whiskers denote the lowest and highest values within 1.5 IQR from the first and third quartiles. The circles represent outliers beyond the whiskers. The notches show the 95% confidence interval for the medians. If the notches of two boxes do not overlap, it gives evidence of a significant difference between the medians.

©2014 Macmillan Publishers Limited. All rights reserved

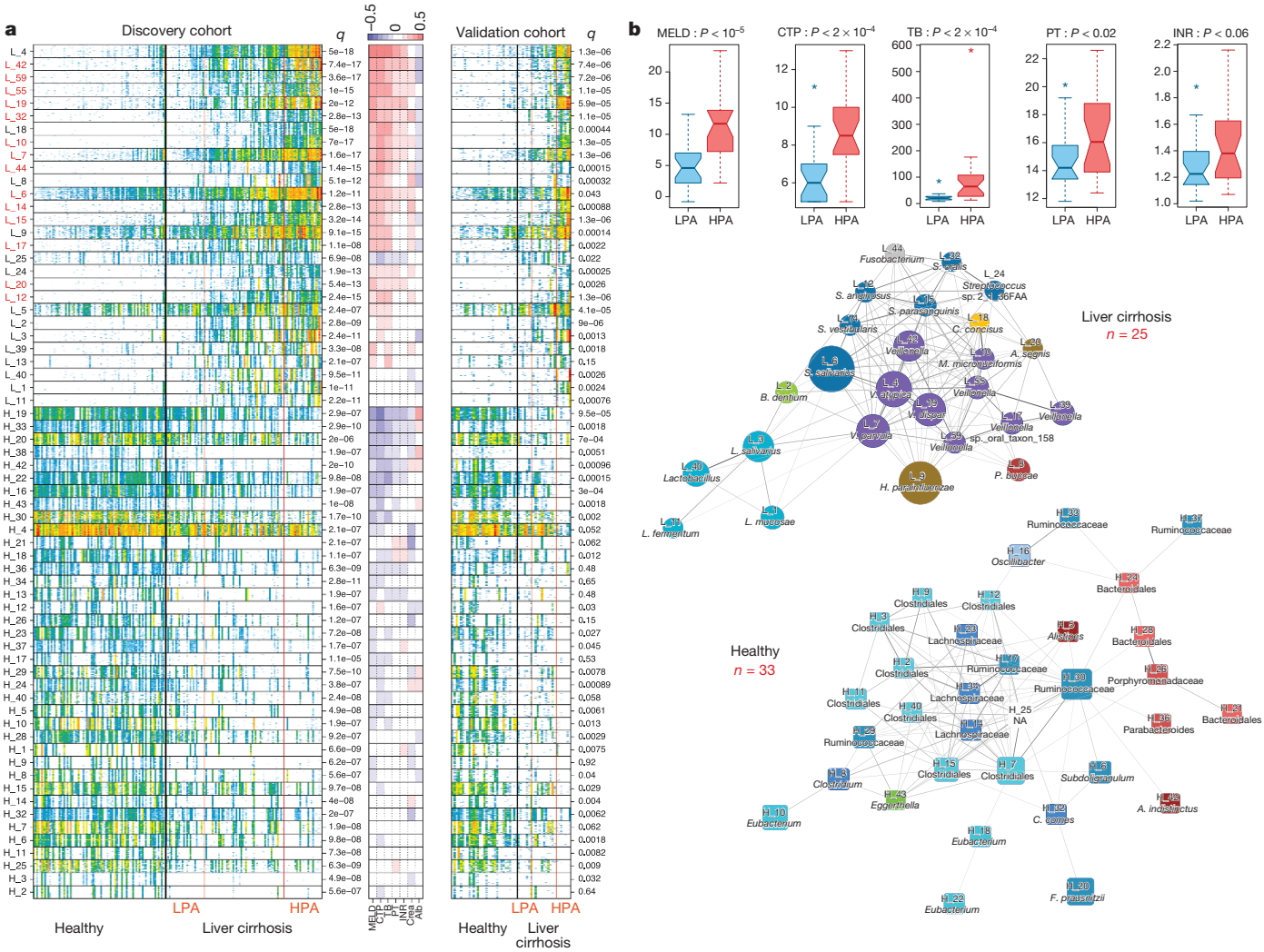


Figure 2 | Differentially abundant MGS in patients ($n = 123$) and healthy individuals ($n = 114$). **a**, Abundance of 50 'tracer' genes for each species in the discovery ($n_{patients} = 98$, $n_{healthy} = 83$) and validation cohorts ($n_{patients} = 25$, $n_{healthy} = 31$); oral species are highlighted in red. Genes are in rows, abundance is indicated by colour gradient (white, not detected; red, most abundant); the enrichment significance is shown (q indicates the Mann–Whitney P values corrected by the Benjamini and Hochberg method). Individuals are shown in columns, ordered by increasing abundance of patient-enriched species. Correlation of the species abundance and patients' clinical parameters in the discovery cohort are indicated in colour code (red and blue for positive and negative correlations; intensity reflects the level of correlation). MELD, model for end-stage liver disease; CTP, Child–Turcotte–Pugh score; TB, total bilirubin; PT, prothrombintime test; INR, international normalized ratio describing coagulation of the blood in patients with liver cirrhosis; Crea, creatinine level; Alb, albumin level. **b**, Top, clinical parameters of patients for the lowest and highest patient-enriched species abundance (LPA and HPA, respectively; $n = 24$ for each). P values indicate the significance of the difference by Mann–Whitney U-test except MELD (Student's t-test). Middle and bottom, abundance-based species correlation network enriched in patients with liver cirrhosis ($n = 25$) and healthy individuals ($n = 33$), respectively. Two nodes are linked if the pooled variance z-test shows an FDR $< 10^{-9}$ when accounting for the compositionality effect (see Methods). The edge width is proportional to the correlation strength. The node size is proportional to the mean abundance in the respective population. Nodes with the same colour are classified in the same phylogenetic order level.

overgrowth is frequently found in patients with liver cirrhosis[33]. To explore the origin of the patient-enriched species, we used information from the HOMD[34] and GOLD[35] databases about the origin of the closely related sequenced isolates. We also constructed a catalogue of 114 publicly available genomes for *Streptococcus*, *Fusobacterium*, *Lactobacillus*, *Veillonella and Megasphaera* strains, originating mostly from mouth or gut (57 or 28, respectively; Supplementary Table 6) and used it for blastN and blastP analysis (Methods). Thirteen of the species were closest to an oral isolate whereas only six were closest to the gut isolates, a single species being from the ileum (Supplementary Table 4 and Extended Data Fig. 4, top right). Comparison with the three ileum metagenomes failed to reveal identity above that detected by comparison with the sequenced genomes (Methods). We conclude that oral commensals invade the gut in patients with liver cirrhosis. Possibly, an altered bile production in cirrhosis renders the 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 *Haemophilus 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 the colon but also in the ileum, and contribute to the small-intestinal bacterial overgrowth associated with liver cirrhosis. Among the patient-enriched species were *Streptococcus anginosus*, *Veillonella atypica*, *Veillonella dispar*, *Veillonella* sp. oral taxon and *Clostridium perfringens*, which have been reported to cause opportunistic infections[38–40].

To analyse the relations between the liver-cirrhosis-associated MGS, we generated networks based on co-abundance, for healthy individuals and patients with liver cirrhosis (Fig. 2b). A striking feature is that taxonomically related species tend to cluster, as reported previously[29]. These observations indicate that the gut environment becomes permissive for

©2014 Macmillan Publishers Limited. All rights reserved

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*, *H. 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 very low in healthy individuals (Extended Data Fig. 4, bottom). 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. 2a). The disease status of the patients with the highest load of these bacteria was significantly worse than that of the patients with the lowest load (Fig. 2b, top). 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 4,801 KEGG (Kyoto Encyclopedia of Genes and Genomes database) orthologues and 13,970 eggNOG (evolutionary genealogy of genes: Non-supervised Orthologous Groups database) orthologues associated with the disease (Supplementary Tables 7 and 8). The most abundant KEGG orthologues in patients and controls were enzyme families. The most enriched orthologues in patients were membrane transport, similar to findings for IBDs[19,20], obesity[41] and T2D[22]. In contrast, the most prevalent markers among the controls included those involved in carbohydrate metabolism, amino-acid metabolism, energy metabolism, signal transduction and the metabolism of cofactors and vitamins (Extended Data Fig. 5). At the module or pathway level, the liver-cirrhosis-associated markers included assimilation or dissimilation of nitrate to or from ammonia, denitrification, GABA (γ-aminobutyric acid) biosynthesis, GABA shunt, haem biosynthesis, phosphotransferase systems and some types of membrane transport, such as amino-acid transport. The control-enriched modules included histidine metabolism, ornithine biosynthesis, creatine pathway, carbohydrate metabolism, repair systems and glycosaminoglycan metabolism (Supplementary Table 9).

The enrichment of the modules for ammonia production in patients suggests a potential role of gut microbiota in hepatic encephalopathy, a complication related to liver cirrhosis that is characterized 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 in 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 encephalopathy[42], 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 enzymes, affects transmission function of neural synapses and eventually leads to hepatic encephalopathy[40]. Finally, the modules for GABA biosynthesis were enriched in the patients. The GABA neurotransmitter system is involved in the pathogenesis of hepatic encephalopathy in humans[43]. Because of the hydrodynamic venous shunt and liver failure, GABA levels in the blood are increased[44], and could go through the blood–brain barrier to activate GABA receptor and cause hepatic encephalopathy. Microbiome modulation, aiming at manganese elimination and lowering of GABA levels in the gut, might provide a new therapeutic option for the treatment of hepatic encephalopathy.

Microbial dysbiosis in chronic diseases

It is unclear whether a gut microbial dysbiosis in type 2 diabetes (T2D)[22], IBD[41] and liver cirrhosis[13] is similar or unique for each disease. We compared the differences between the gut microbiota from patients with liver cirrhosis, T2D and IBD, and organized the disease-associated gene, KEGG orthologue group and eggNOG orthologue group markers into patient- and control-enriched groups. We then identified markers common to different disease pairs (T2D and liver cirrhosis, liver cirrhosis and IBD, and IBD and T2D) and to the three diseases (Supplementary Table 10). Different diseases displayed a relatively unique profile, even if some markers were shared (Extended Data Fig. 6a, b). Most liver-cirrhosis-enriched markers had low *P* values (Extended Data Fig. 6c), implying that patients with liver cirrhosis had more severe dysbiosis than patients with T2D. Functional differences between liver cirrhosis and T2D were also detected at the pathway level, even if there was a significant increase in membrane transport markers in both (Extended Data Figs 7 and 8). Most functional markers in both diseases were from categories of carbohydrate metabolism, metabolism of cofactors and vitamins, amino-acid metabolism and signal transduction. In contrast, most cell motility markers in the KEGG orthologue group were enriched in liver cirrhosis or T2D but not both, possibly indicating a unique role in each disease (Extended Data Fig. 8a, b). However, similar cell motility markers and pathways in the KEGG orthologue group were enriched both in liver cirrhosis and in T2D controls, suggesting a possible role in health (Extended Data Figs 8c, d and 9a, b).

Gene markers that identify patients with liver cirrhosis

We used a pattern recognition technique to identify patients by gut microbiota information in the discovery cohort ($n = 181$). For this we selected 46,000 genes, half enriched in patients and half in controls (Supplementary Table 11). From this set we selected 15 optimal gene markers by a minimum redundancy–maximum relevance (mRMR) method combined with an incremental feature search, which showed the highest value of Matthews correlation coefficient (Extended Data Fig. 9c). A support vector machine discriminator was constructed using the same samples and 15 gene markers (Supplementary Table 12), with the training and leave-one-out cross-validation AUC (area under the receiver operating characteristic curve) achieving 0.918 (confidence interval: 0.881–0.955) (Fig. 3b) and 0.838, respectively. The validation cohort of 31 healthy controls and 25 patients with liver cirrhosis showed an AUC value of 0.836 (95% confidence interval 0.730–0.943) (Fig. 3c) for these samples, confirming that the gut microbiota information could be applied to identify patients accurately.

To facilitate the clinical application of the 15 optimal gene markers, we propose a patient discrimination index (PDI). The high correlation coefficient value between the ratio of patients in our cohort and the PDI (Fig. 3a and Supplementary Table 13) indicates that the PDI could be used to identify patients with liver cirrhosis. The discriminatory power of the PDI was then validated using an independent group (Fig. 3d). The average PDI index between the control and the patient groups was significantly different ($P < 8.18 \times 10^{-5}$, Wilcoxon rank-sum test), confirming the potential use of gut microbiota information for identifying patients with liver cirrhosis.

Discussion

To study gut microbiota in liver cirrhosis we first established a novel gut gene catalogue (liver cirrhosis catalogue), including 98 patients with liver cirrhosis and 83 healthy control individuals. Comparison with the previously established MetaHIT and T2D[22] gene catalogues indicated a common core of approximately 800,000 genes and a considerable proportion of catalogue-specific genes (37.01% of MetaHIT, 36.59% of T2D and 18.02% of liver cirrhosis), indicating that the current gene sets are still limited and should be completed by inclusion of more individuals. Interestingly, although the T2D and liver cirrhosis 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 disease profiles and to the different genotypes, body mass indices, age[45] and dietary habits[46] (Supplementary Table 14 and Extended Data Fig. 10). Nevertheless, there was no significant difference in the abundance of main phyla ($P > 0.01$); of the top 30 most abundant genera and species, 28 and 26, respectively, were the same in both studies, and there were no significant differences in abundance for most of them. Furthermore, the top four species were exactly the same. These results, and the similarity of

©2014 Macmillan Publishers Limited. All rights reserved

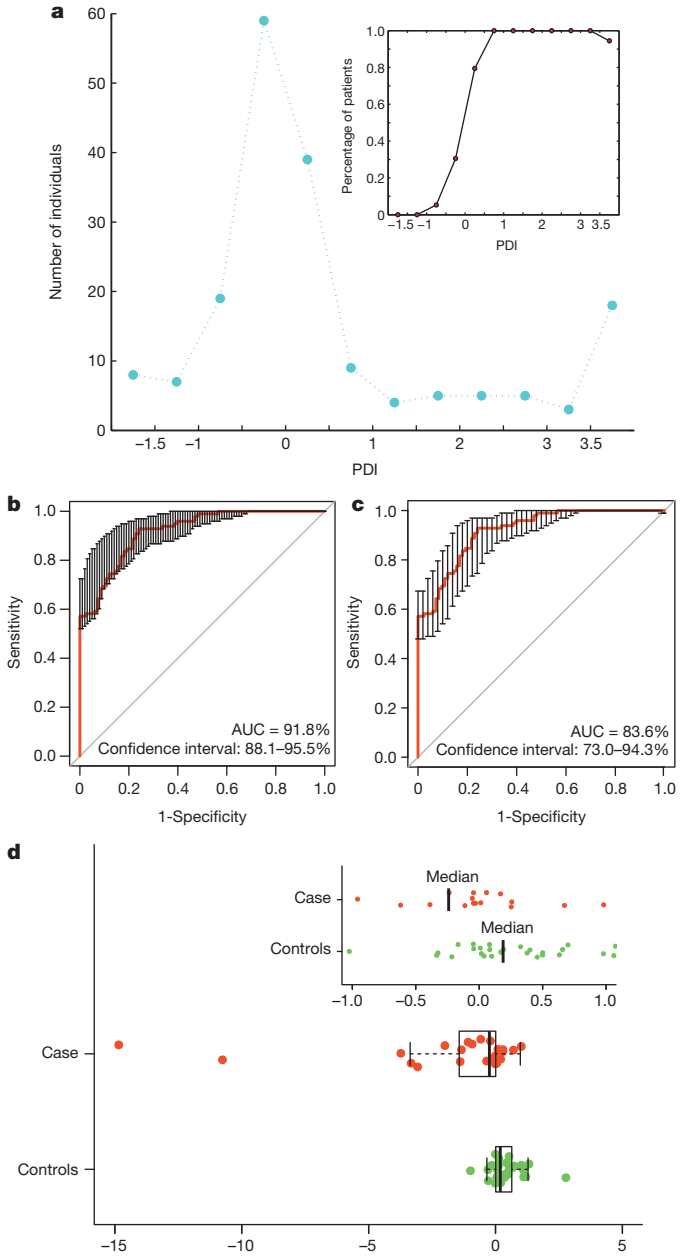


Figure 3 | PDI on the basis of gut microbial biomarkers. **a**, A PDI was calculated for each individual from 15 gene markers selected using the mRMR approach to evaluate the risk of liver cirrhosis. The filled blue circles show the distribution of liver cirrhosis indices for all individuals (bins of 0.5 PDI units were used; values less than −1.5 and greater than 3.5 were grouped). Inset, the proportion of patients with liver cirrhosis in the corresponding bins. **b**, **c**, The AUC is shown for the training (**b**) and validation (**c**) samples. **d**, The liver cirrhosis PDI was computed for an additional 25 liver cirrhosis samples and 31 healthy control samples. The box depicts the interquartile range between the first and third quartiles (25th and 75th percentiles, respectively); the line inside denotes the median. Inset, the PDI without the outliers.

controls with the healthy Danish population, point towards overall similarity of the microbiota in healthy individuals.

Use of the liver cirrhosis gene catalogue, in conjunction with the quantitative metagenomics approach, revealed a major change of the gut microbiota in the patients with liver cirrhosis, mainly because of a massive invasion of the gut by oral bacterial species. Correlation of the severity of the disease with the abundance of the invading species suggests that they may play an active role in the pathology. This was not noted in a previous study, where the 16S-based approach probably lacked the required species-level resolution, even if similar trends in taxonomy change between the liver cirrhosis group and the healthy controls at the phylum, class and order levels were observed[13]. Some of the MGS depleted in patients were negatively associated with the severity of the disease (Fig. 2). This opens avenues to the development of novel probiotics, which might help combat the aggravation of liver cirrhosis. More generally, modulation of microbiota to correct the major dysbioses we report might open new avenues to treatment of liver cirrhosis.

A combination of 15 microbial genes discriminates patients with liver cirrhosis from healthy individuals, with a high specificity. This could lead to a new way of monitoring and preventing liver cirrhosis. None of the 15 markers found in the liver cirrhosis study overlapped with the 50 markers found in the T2D study[22], indicating that diagnosis of different diseases with microbiota-targeted biomarkers may be a powerful tool for disease detection.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 7 April 2013; accepted 9 June 2014.
Published online 23 July 2014.

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. *J. Hepatol.* **56,** 1283–1292 (2012).
2. Cesaro, C. *et al.* Gut microbiota and probiotics in chronic liver diseases. *Digest. Liver Dis.* **43,** 431–438 (2011).
3. Wiest, R. & Garcia-Tsao, G. Bacterial translocation (BT) in cirrhosis. *Hepatology* **41,** 422–433 (2005).
4. Nolan, J. P. The role of intestinal endotoxin in liver injury: a long and evolving history. *Hepatology* **52,** 1829–1835 (2010).
5. Gill, S. R. *et al.* Metagenomic analysis of the human distal gut microbiome. *Science* **312,** 1355–1359 (2006).
6. Garcia-Tsao, G. & Wiest, R. Gut microflora in the pathogenesis of the complications of cirrhosis. *Best Pract. Res. Clin. Gastroenterol.* **18,** 353–372 (2004).
7. Wiest, R., Krag, A. & Gerbes, A. Spontaneous bacterial peritonitis: recent guidelines and beyond. *Gut* **61,** 297–310 (2012).
8. Bass, N. M. *et al.* Rifaximin treatment in hepatic encephalopathy. *N. Engl. J. Med.* **362,** 1071–1081 (2010).
9. Benten, D. & Wiest, R. Gut microbiome and intestinal barrier failure–the ''Achilles heel'' in hepatology? *J. Hepatol.* **56,** 1221–1223 (2012).
10. Yan, A. W. *et al.* Enteric dysbiosis associated with a mouse model of alcoholic liver disease. *Hepatology* **53,** 96–105 (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. *Proc. Natl Acad. Sci. USA* **107,** 14691–14696 (2010).
12. Cho, I. & Blaser, M. J. The human microbiome: at the interface of health and disease. *Nature Rev. Genet.* **13,** 260–270 (2012).
13. Chen, Y. *et al.* Characterization of fecal microbial communities in patients with liver cirrhosis. *Hepatology* **54,** 562–572 (2011).
14. Nelson, K. E. *et al.* A catalog of reference genomes from the human microbiome. *Science* **328,** 994–999 (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 (2006).
16. Turnbaugh, P. J. *et al.* An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444,** 1027–1031 (2006).
17. Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457,** 480–484 (2009).
18. Ley, R. E. *et al.* Obesity alters gut microbial ecology. *Proc. Natl Acad. Sci. USA* **102,** 11070–11075 (2005).
19. Lepage, P. *et al.* Twin study indicates loss of interaction between microbiota and mucosa of patients with ulcerative colitis. *Gastroenterology* **141,** 227–236 (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 (2010).
21. Wen, L. *et al.* Innate immunity and intestinal microbiota in the development of type 1 diabetes. *Nature* **455,** 1109–1113 (2008).
22. Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490,** 55–60 (2012).
23. Vijay-Kumar, M. *et al.* Metabolic syndrome and altered gut microbiota in mice lacking Toll-like receptor 5. *Science* **328,** 228–231 (2010).
24. Karlsson, F. H. *et al.* Symptomatic atherosclerosis is associated with an altered gut metagenome. *Nature Commun.* **3,** 1245 (2012).
25. The Human Microbiome Project Consortium. A framework for human microbiome research. *Nature* **486,** 215–221 (2012).
26. The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486,** 207–214 (2012).
27. Le Chatelier, E. *et al.* Richness of human gut microbiome correlates with metabolic markers. *Nature* **500,** 541–546 (2013).

©2014 Macmillan Publishers Limited. All rights reserved

28. Cotillard, A. *et al.* Dietary intervention impact on gut microbial gene richness. *Nature* **500,** 585–588 (2013).
29. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464,** 59–65 (2010).
30. Nielsen, H. B. *et al.* Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nature Biotechnol*. http://dx.doi.org/10.1038/nbt.2939 (2014).
31. Albertsen, M. *et al.* Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature Biotechnol.* **31,** 533–538 (2013).
32. Zoetendal, E. G. *et al.* The human small intestinal microbiota is driven by rapid uptake and conversion of simple carbohydrates. *ISME J.* **6,** 1415–1426 (2012).
33. Bauer, T. M. *et al.* Small intestinal bacterial overgrowth in human cirrhosis is associated with systemic endotoxemia. *Am. J. Gastroenterol.* **97,** 2364–2370 (2002).
34. Chen, T. *et al.* The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. *Database* **2010,** baq013 (2010).
35. Pagani, I. *et al.* The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.* **40,** D571–D579 (2012).
36. Saarela, M., Mogensen, G., Fonden, R., Matto, J. & Mattila-Sandholm, T. Probiotic bacteria: safety, functional and technological properties. *J. Biotechnol.* **84,** 197–215 (2000).
37. Merritt, M. E. & Donaldson, J. R. Effect of bile salts on the DNA and membrane integrity of enteric bacteria. *J. Med. Microbiol.* **58,** 1533–1541 (2009).
38. Marchandin, H. *et al.* Prosthetic joint infection due to *Veillonella dispar*. *Eur. J Clin. Microbiol. Infect. Dis.* **20,** 340–342 (2001).
39. 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. *J. Microbiol. Immunol. Infect.* **35,** 65–67 (2002).
40. Xu, M. *et al.* Changes of fecal *Bifidobacterium* species in adult patients with hepatitis B virus-induced chronic liver disease. *Microb. Ecol.* **63,** 304–313 (2012).
41. 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. *Proc. Natl Acad. Sci. USA* **109,** 594–599 (2012).
42. Krieger, D. *et al.* Manganese and chronic hepatic encephalopathy. *Lancet* **346,** 270–274 (1995).
43. 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* **ii,** 811–814 (1983).
44. Minuk, G. Y., Winder, A., Burgess, E. D. & Sarjeant, E. J. Serum gamma-aminobutyric acid (GABA) levels in patients with hepatic encephalopathy. *Hepatogastroenterology* **32,** 171–174 (1985).
45. Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* **486,** 222–227 (2012).
46. Wu, G. D. *et al.* Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334,** 105–108 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by the National Program on Key Basic Research Project (2013CB531401), the National Natural Science Foundation of China (81301475 and 81330011), the Science Fund for Creative Research Groups of the National Natural Science Foundation of China (81121002), the Technology Group Project for Infectious Disease Control of Zhejiang Province (2009R50041) and the Metagenopolis grant ANR-11-DPBS-0001. We thank Q. Cao, K. Su, J. Shao and A. Ghozlane for help with data computation, and H. Zhang, H. Lu, Q. Bao, J. Ge, J. Jiang, Z. Ren and M. Ye for assistance with sample collection. We are thankful to the MetaHIT consortium for generating the gut gene set and the Human Microbiome Project for generating the reference genomes from human gut microbes.

Author Contributions L.J.L., S.D.E., S.S.Z. and N.Q. designed the project. L.J.L., S.P.K. and N.Q. managed the project. F.L.Y., N.Q., Y.F.C., J.G., G.R.Q., X.J.H. and B.W.Z. collected samples and performed clinical study. J.G., Y.T.C. and W.X. performed DNA extraction experiments. Y.J., L.J.W., J.W.Z. and S.J.N. performed library construction and sequencing. L.J.L. and S.D.E. designed the analysis. N.Q., A.L., E.P., E.L.C., L.L., N.P., P.L., J.M.B., C.H.Y. and W.C.D. analysed the data. A.L. and N.Q. did the functional annotation analyses. L.S., E.P., E.L.C. and A.L. analysed the statistics. N.Q., F.L.Y., L.S. and E.P. wrote the paper. L.J.L. and S.D.E. revised the paper.

Author Information The raw Illumina read data for all samples have been deposited in the European Bioinformatics Institute European Nucleotide Archive under accession number ERP005860. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to L.J.L. (ljli@zju.edu.cn), S.S.Z. (zyzsss@zju.edu.cn) or S.D.E. (dusko.ehrlich@jouy.inra.fr).

©2014 Macmillan Publishers Limited. All rights reserved

METHODS

Patient information. Liver cirrhosis was diagnosed according to the international guidelines by comprehensive consideration of liver biopsy, imaging examination, clinical symptoms, physical signs, laboratory tests, medical history, progress notes and cirrhosis-associated complications. Biopsy as the 'gold 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 ascites and obvious bleeding 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 discharge of the patient from the hospital, their case history was further reviewed for medication history. Cases that progressed to hepatic carcinoma or those found to 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, 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 and chest X-ray results were checked in the healthy controls to exclude any abnormal samples. Comprehensive clinical information for each enrolled individual was recorded (Supplementary Table 1). Exclusion criteria for the control group included hypertension, diabetes, obesity, metabolic syndrome, IBD, non-alcoholic fatty liver disease, coeliac disease and cancer. Individuals who received antibiotics and/or probiotics within 8 weeks before enrolment were also excluded. All participants, or their legally authorized representatives, provided a written informed consent upon enrolment. 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 individual provided a fresh stool sample that was delivered immediately from our hospital to the laboratory in an ice bag using insulating polystyrene foam containers. In the laboratory it was divided into five aliquots of 200 mg and immediately stored at −80 °C. A frozen aliquot (200 mg) of each faecal sample was processed by phenol trichloromethane DNA extraction[16,47] as previously described. DNA concentration was measured by NanoDrop (Thermo Scientific) and its molecular size was estimated by agarose gel electrophoresis.

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

Quality control of reads. Reads that mapped to human genome together with their mated/paired reads were removed from each sample using BWA[48] with parameter '-n 0.2'. Then quality control used the following criteria: (1) reads containing more than 3 *N* bases were removed; (2) reads containing more than 50 bases with low quality (Q2) were removed; (3) 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 the next step of the analysis.

***De novo* assembly of the Illumina short reads.** Considering that *k*-mers with very low frequencies might arise from sequencing errors, they were not used in assembly by SOAPdenovo[49] (version 1.05), which is based on De Brujin graph construction. SOAPdenovo (version 1.05) was used in Illumina short read 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 lengths less than 500 bp. Finally we tested series of *k*-mer values (from 31 to 59), then chose one with the longest N50 value for the remaining scaffolds. For each sample, we mapped clean data against scaffolds using SOAPalign version 2.21 (ref. 50) with parameters '-u -2 -m 200'. Unused data from each sample were pooled and split into four parts (considering memory limit). Unused reads were repeatedly assembled with the same parameters but only one *k*-mer value, -K 55, was chosen.

Construction of non-redundant human gut gene set. Total DNA was extracted from the faecal samples of 98 Chinese patients with liver cirrhosis and 83 healthy Chinese controls (Supplementary Table 1) and sequenced using an Illumina HiSeq 2000 (Illumina). This produced an average of 4.74 gigabases (Gb) of high-quality sequence for each sample, providing a total of 858 Gb of sequence data (Supplementary Table 15). The reads were assembled into contigs for all samples using the assembly software SOAPdenovo[49].Unassembled reads from 166 samples were pooled and the *de novo* assembly process was performed again for these reads (Extended Data Fig. 9d). 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 48,822 bp (Supplementary Table 15). To predict microbial genes for each of the 181 samples, we applied the methodology used in the MetaHIT human gut gene catalogue study[29]. The non-redundant human gut gene set was built by pairwise 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 set[29] and T2D study[22].

MetaGeneMark[51] (prokaryotic GeneMark.hmm version 2.8) was used to predict ORFs in scaffolds without ambiguous bases. The program predicted 13,371,697 ORFs using a 100 bp cut-off for prediction (Supplementary Table 15). 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 'liver cirrhosis gene set' was established by removing redundant ORFs, defined as those sharing 95% identity over 90% of the shorter ORF length in pairwise alignments. The final non-redundant liver cirrhosis gut gene set contained 2,688,468 ORFs, with an average length of 750 bp and 42% of reads could be aligned to the gene catalogue.

Then genes from the liver cirrhosis, T2D and MetaHIT catalogues 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 alignment result of two ORFs, the shorter one would not be removed as a redundancy. We used MetaGeneMark to predict genes in assembled contigs originally from MetaHIT and T2D 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 the same reference genomes could be assigned into two types, as follows. (1) Unique reads (*U*): reads having alignments with only one genome. These reads were denoted as unique reads. (2) Multiple reads (*M*): reads having alignments with more than one genome. If these genomes came from one species, we denoted these reads as unique reads. If they were from more than one species, we denoted 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, the computation is

$$\mathrm{Ab}(S) = \mathrm{Ab}(U) + \mathrm{Ab}(M)$$

$$\mathrm{Ab}(U) = U/l$$

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

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

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

For these reads, we add a unique abundance of *N* species as the denominator. Before we calculate the abundance of species *S*, we calculate 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.

Gene abundance profiling. Reads were aligned against the gene set by using SOAPalign[50] with parameters '-r -m 200 -x 1000'. We counted a 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 non-aligned read were in the non-translated region or not. If true, both reads were validated for gene count; if not, both reads were discarded.

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

©2014 Macmillan Publishers Limited. All rights reserved

$$\mathrm{Ab}(G) = \mathrm{Ab}(U) + \mathrm{Ab}(M)$$

$$\mathrm{Ab}(U) = U/l$$

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

Ab(U) and Ab(M) are the abundances of unique and multiple reads, respectively, and l is length of gene G. For each multiple read, 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.

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

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

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

Gene count determination. Gene counts were computed essentially as described in ref. 27. Briefly, data were downsized to adjust for sequencing depth and technical variability by randomly selecting 6.2 million reads mapped to the merged gene catalogue for each sample and then computing the mean number of genes over 30 random drawings (Supplementary Table 4). This was possible for all but two patients with liver cirrhosis from the validation cohort (with insufficient number of mapped reads), who were excluded from this analysis. The results are displayed in Extended Data Fig. 4 top left.

Gene functional classification and orthologue group abundance profiling. Protein sequences of the predicted genes were searched using National Center for Biotechnology Information blastP against the eggNOG 3.0 database[52] and the KEGG gene database (KEGG FTP release 21 January 2013) with parameters '-num_descriptions 100000, -evalue 1e-5'. Genes that had alignments with a bits score higher than 60 were assigned into one or more eggNOG or KEGG orthologue groups. We used the methods introduced in ref. 29 to calculate abundance of proteins archived in the eggNOG and KEGG databases. To calculate abundances of eggNOG or KEGG orthologue groups, we added abundances of proteins assigned into the same eggNOG or KEGG orthologue groups, as abundances of eggNOG or KEGG orthologue groups, then profiles of eggNOG/KEGG orthologue groups were generated.

Gene biomarker identification. Genes from the gene-profile matrix were used in an association study aimed at identifying those that were differentially abundant between the patient and the healthy control groups. Wilcoxon tests were employed to compute the probabilities that frequency profiles did not differ between the patient and the healthy control groups by chance alone. Benjamini and Hochberg multiple test correction was applied to the P values. By performing a selection only based on a 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 were differentially abundant between the groups (49,830 were more abundant in the patients with liver cirrhosis and 25,415 in the healthy control group). A similar P value and group enrichment method was calculated for the NOG/KEGG orthologue groups as well.

MGS. We followed the approach described in refs 27 and 30 to cluster genes from the current study into MGS. Briefly, in a first step the pairwise Spearman's 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 ($\rho > 0.7$). To correct for the concomitant loss of sensitivity, we performed a second step whereby the mean abundance signal of each cluster of at least 50 genes was computed, using the 50 most connected genes of a cluster. The clusters that had $\rho > 0.85$ were fused. This procedure was applied separately to the 49,830 genes enriched in patients with liver cirrhosis and the 25,415 genes enriched in healthy controls. Of the 25,415 'healthy' genes, 21,423 fell into 43 clusters composed of 51–2,702 genes after the first clustering step, and 38 clusters of 51–2,970 genes after the second step. Of the 'liver cirrhosis' genes, 31,386 out of 49,830 fell into 60 clusters of 51–3,000 genes after the first clustering step, and 28 clusters of 51–5,755 genes after the second step.

To verify that the genes from a given cluster belonged to the same genome and to annotate the MGS taxonomically, we performed blastN and blastP analyses using a collection of 6,006 genomes (the available reference genomes from the National Center for Biotechnology Information and the set of draft gastrointestinal genomes from the Data Analysis and Coordination Center of the HMP and MetaHIT (3 August 2012 version)). MGS were assigned to a given genome when more than 80% of its 'tracer genes'[27] matched the same genome using blastN, at a threshold of 95% identity over 90% of gene length. Six 'healthy' and 24 'liver cirrhosis' MGS could thus be assigned to the strain level (see Extended Data Fig. 9e, f and Supplementary Table 4). The remaining MGS were annotated using blastP analysis and assigned to a given taxonomical level from genus to superkingdom level if more than 80% of their 50 tracer genes had the same level of assignment[27]. All but one of the 36 remaining species could thus be assigned to a given genus, family or order (see Supplementary Table 4). The quality of the clustering was thus validated by the homogenous annotation of its marker genes, which also held true for all of the MGS genes (data not shown). The abundance of the 66 MGS in each individual was computed using the 50 tracer genes.

To explore the origin of the species-level annotated MGS, we constructed a reference catalogue, grouping 114 publicly available *Streptococcus* (57), *Fusobacterium* (26), *Lactobacillus* (16), *Veillonella* (12) and *Megasphaera* (3) genomes, mostly of oral (50) or gut (28) isolates (Supplementary Table 6). The 16 liver cirrhosis MGS that were assigned to the corresponding genera were compared with the genomes, using blastN. A score (T) was computed for each MGS, taking into account (1) the proportion of genes above 95% identity and 90% coverage (Q), (2) the average identity (R), (3) the average coverage (S) and (4) $T = Q \times R \times S$.

A majority of the MGS enriched in patients with liver cirrhosis (15 out of 28) were of oral origin by this criterion whereas six were from gut or faeces, including a single species from the ileum (Supplementary Table 4 and Extended Data Fig. 4 top right). To explore further the origin of the liver-cirrhosis-enriched MGS, we compared them by blastN with the genes from three available ileum metagenomes[31] 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 assigned species phylogenetic information by comparison with sequenced gut genomes using blastN (95% identity and 90% overlap; Supplementary Table 4). Annotation to comparable taxonomic levels was observed for the 58 gut MGS analysed in the context of gene richness in a Danish cohort[27] (Extended Data Fig. 9e, f), reflecting a paucity of isolated and sequenced gut strains. Furthermore, it is striking that all 38 MGS enriched in healthy Chinese were found in the Danish cohort (Extended Data Fig. 3). In sharp contrast with the MGS enriched in healthy subjects, an overwhelming majority of the MGS enriched in patients (24 out of 28) could be assigned to a species. Such a difference has a vanishingly low probability of being caused by chance alone (1.3×10^{-21} by a χ^2 test, Extended Data Fig. 9e, f) and indicates a highly modified composition of gut microbes.

Co-occurrence network of MGS. The 66 marker profiles of the differentially abundant MGS between patient and healthy individuals were correlated separately for patients and for healthy individuals, essentially as described in ref. 53. For each of the 2,112 possible edges [$(66 \times 66/2) - 66$] we computed 1,000 permutations by renormalizing the data after each step and computed Spearman's correlation coefficients to obtain the null distributions due to the compositionality effect[53]. For each of the edges we also computed the bootstrap distribution of the Spearman's correlation coefficients 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 significance P value. Multiple testing corrections were applied to the P values using the Benjamini and Hochberg method, and only those having FDR $< 10^{-9}$ were used to construct the network. This FDR threshold corresponds approximately to $\rho > 0.4$. The network reflects strong correlations that are not spurious and that are not due to the compositionality effect. The resulting network is displayed as Fig. 2.

Marker selection by mRMR. Patient discrimination gene markers (23,000 from healthy controls and 23,000 from patients, selected as most discriminant by the Wilcoxon rank-sum test upon adjustment for age, performed as described in ref. 54; Supplementary Table 11) were selected with a two-step scheme (using the side Channel Attack R package). All markers retained were first filtered by the mRMR algorithm[55] (using the side Channel Attack R package), and the top 180 best ones were selected for further analysis. Then, we performed an incremental search to select the optimal subset of genes, named as markers. Concisely, genes were sequentially added into the subset with a step of 5, the performance of which was evaluated on the basis of linear discriminant analysis and leave-one-out cross-validation. Here, Matthews correlation coefficient is a balanced measure taking into account true and false positives and negatives; it is superior to accuracy or error rate when the classes (healthy and diseased, etc.) are of very different sizes. Matthews correlation coefficient (MCC) is defined as

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$$

where TP, TN, FP and FN are true positive, true negative, false positive and false negative, respectively. We finally selected a set of 15 gut microbial gene markers as the optimal selection for patient discrimination.

©2014 Macmillan Publishers Limited. All rights reserved

Model construction and validation. On the basis of the 15 metagenomic markers described above, a support vector machine classifier (radial basis function kernel and default parameters) was constructed for patient discrimination (realized by the e1071 package of R software), the performance of which was assessed by receiver operating characteristic analysis. The AUC and corresponding 95% confidence intervals for training and validation data sets, obtained by using the pROC package of R software (10,000 bootstrap replicates), were 0.97 (0.95–0.99) and 0.889 (0.79–0.98), respectively.

Definition of PDI. To facilitate clinical application of the selected 15 metagenomic markers, we defined a more straightforward index (PDI) for discrimination of patients. For each individual sample, the PDI of sample j that was denoted by I_j was computed as follows:

$$I_j^d = \sum_{i \in N} A_{ij}$$

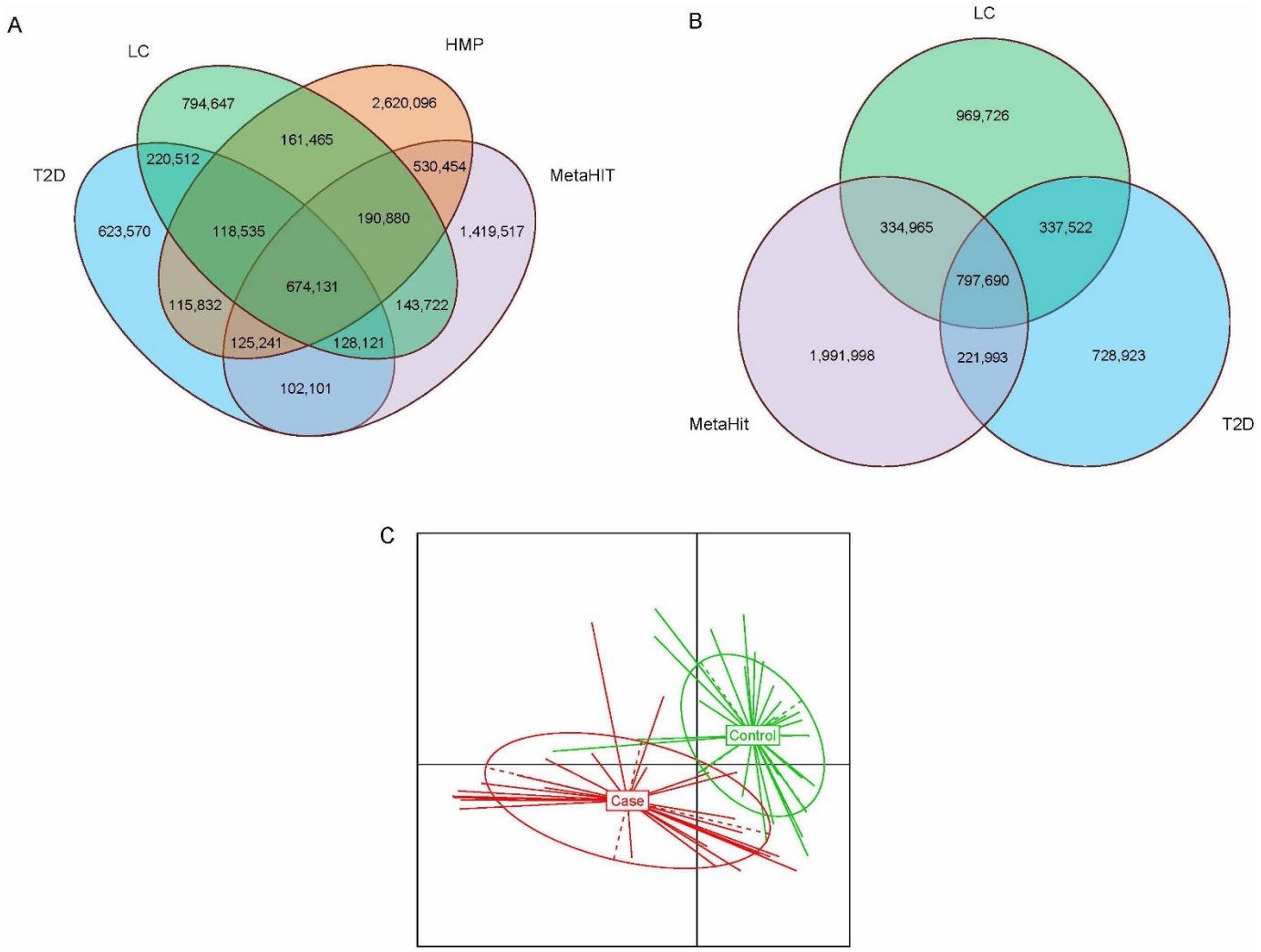
$$I_j^n = \sum_{i \in M} A_{ij}$$

$$I_j = \left(\frac{I_j^d}{|N|} - \frac{I_j^n}{|M|} \right) \times 10^6$$

where A_{ij} is the relative abundance of marker i in sample j. N and M are subsets of patient- and control-enriched markers in these 15 selected gut metagenomic markers, respectively. Moreover, $|N|$ and $|M|$ are the sizes of these two sets.

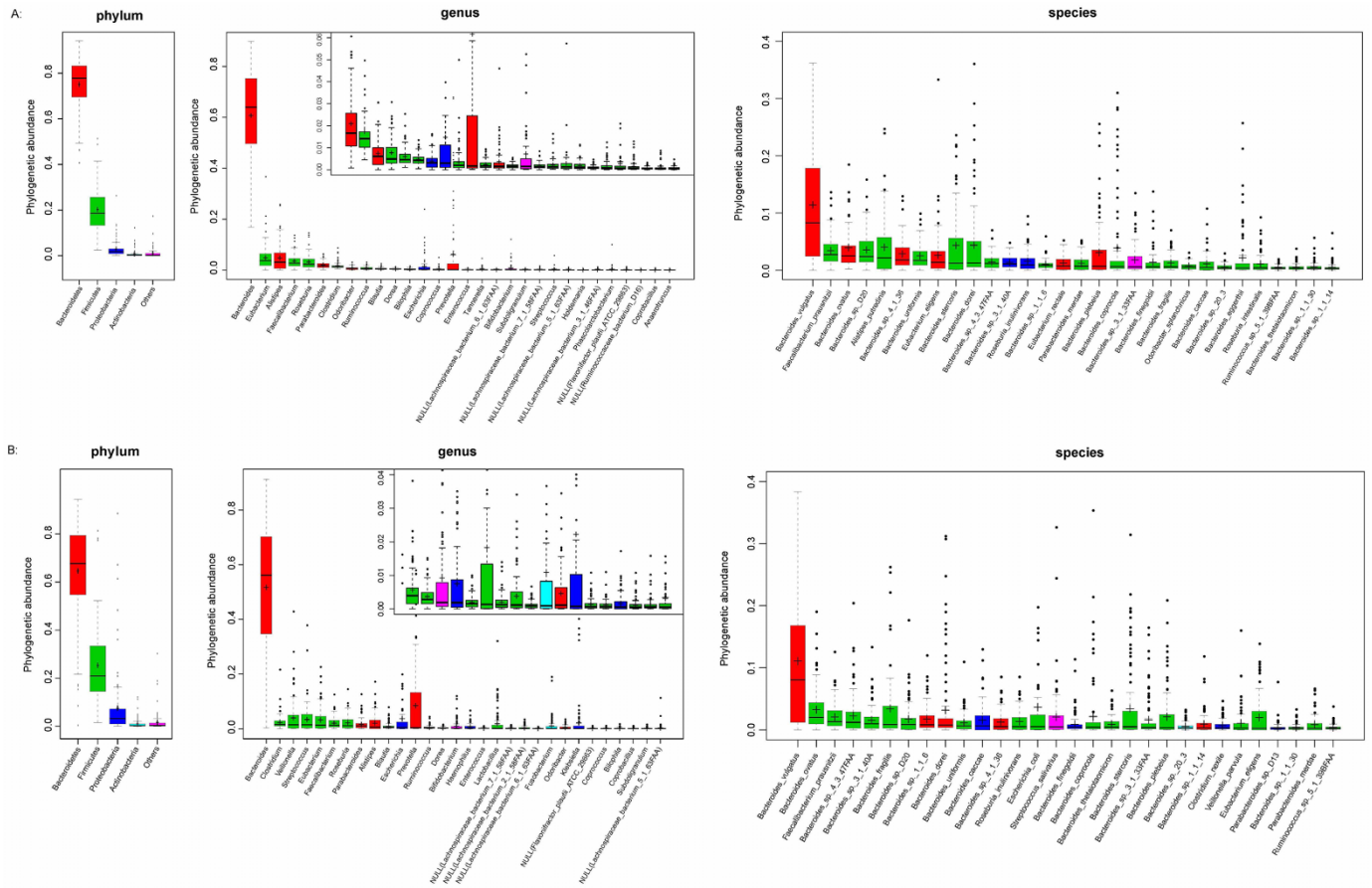
47. Li, M. *et al.* Symbiotic gut microbes modulate human metabolic phenotypes. *Proc. Natl Acad. Sci. USA* **105,** 2117–2122 (2008).
48. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
49. Li, R. *et al.* De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20,** 265–272 (2010).
50. Li, R. *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25,** 1966–1967 (2009).
51. Noguchi, H., Park, J. & Takagi, T. MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res.* **34,** 5623–5630 (2006).
52. Powell, S. *et al.* eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res.* **40,** D284–D289 (2012).
53. Faust, K. *et al.* Microbial co-occurrence relationships in the human microbiome. *PLOS Comput. Biol.* **8,** e1002606 (2012).
54. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genet.* **38,** 904–909 (2006).
55. Ding, C. & Peng, H. Minimum redundancy feature selection from microarray gene expression data. *J. Bioinform. Comput. Biol.* **3,** 185–205 (2005).

©2014 **Macmillan Publishers Limited. All rights reserved**



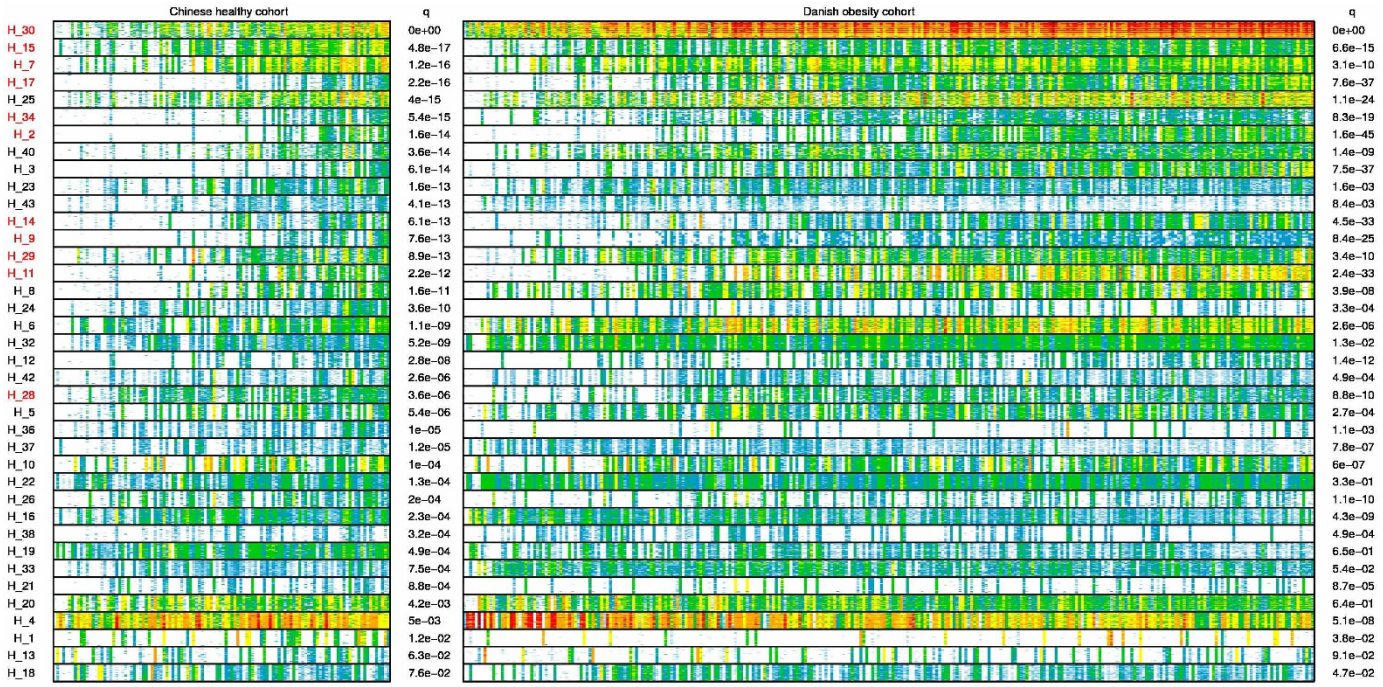
Extended Data Figure 1 | Venn diagram comparing the current major human microbiome gene set and the results of a principal component analysis of biomarkers distributed between patients with liver cirrhosis and healthy controls. **a**, Venn diagram of the four currently available major human microbiome gene sets. The total gene number in each gene set and the overlapping areas are indicated. **b**, Venn diagram of the three major human gut gene sets (LC, liver cirrhosis gene set; T2D, type 2 diabetes gene set; MetaHIT, MetaHIT gene set; HMP, HMP gene set). **c**, Visualization of the principal component analysis results for the liver-cirrhosis-associated genes that differed significantly in the discovery cohort (FDR < 0.0001, Wilcoxon rank-sum test adjusted for multiple testing). The principal component analysis is built here using these genes in the validation cohort (25 patients with liver cirrhosis in red, 31 healthy controls in green).

©2014 Macmillan Publishers Limited. All rights reserved



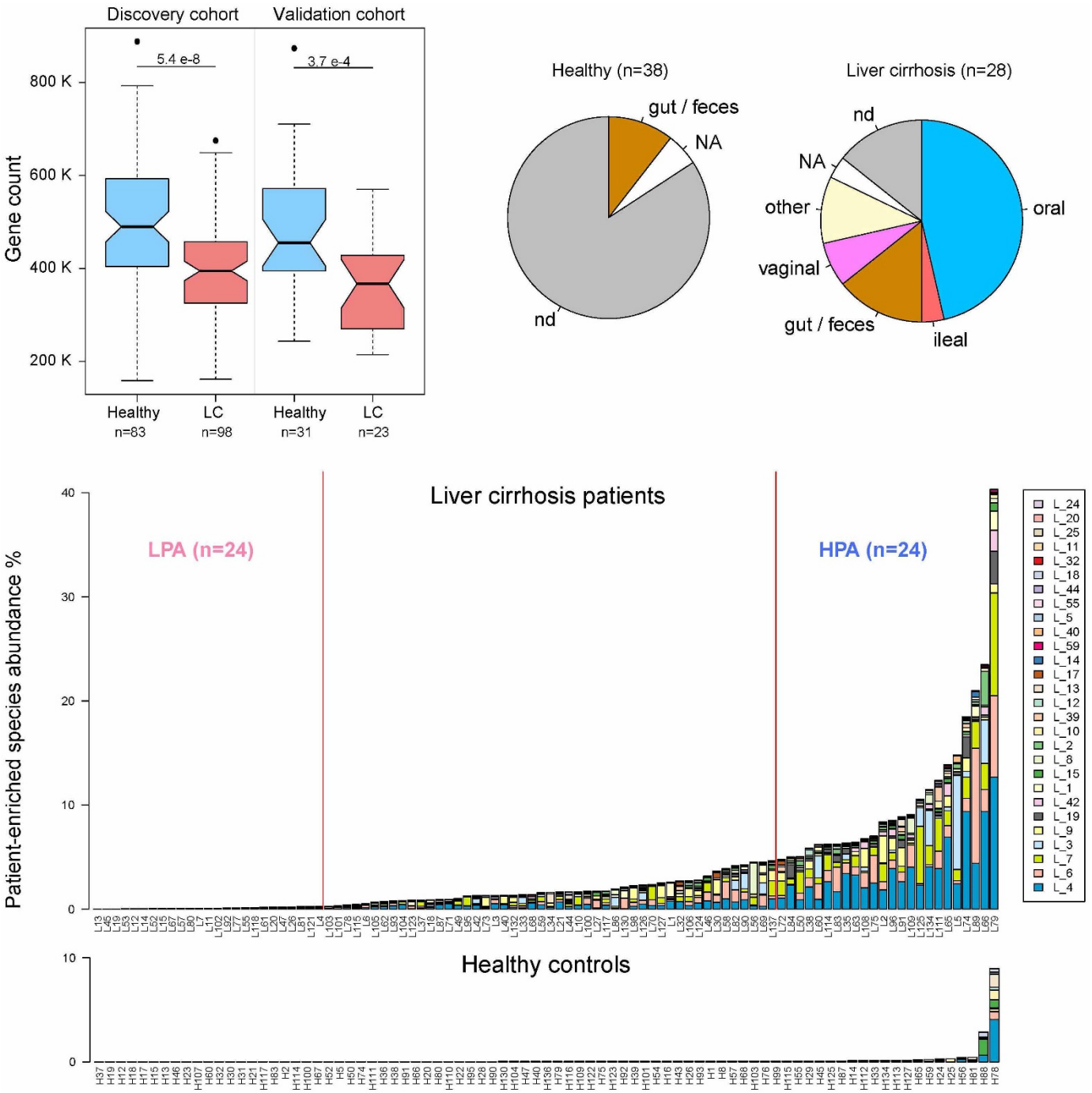
Extended Data Figure 2 | Phylogenetic abundance at the phylum, genus and species levels from liver cirrhosis and healthy control samples. **a**, Phylogenetic abundance variation box plot at the phylum level and the 30 most abundant phylotypes at the genus and species levels in the healthy controls are shown. Red, green, blue, turquoise and purple represent Bacteroidetes, Firmicutes, Proteobacteria, Actinobacteria and other phyla, respectively. The colour of each genus and species corresponds with the colour of its respective phylum. **b**, Phylogenetic abundance variation box plot at the phylum level and the 30 most abundant phylotypes at the genus and species levels in the liver cirrhosis are shown (see Methods for the calculations). The boxes represent the interquartile range, from the first and third quartiles, and the inside line represents the median. The whiskers denote the lowest and highest values within an interquartile range of 1.53 from the first and third quartiles. The circles represent outliers beyond the whiskers.

©2014 Macmillan Publishers Limited. All rights reserved



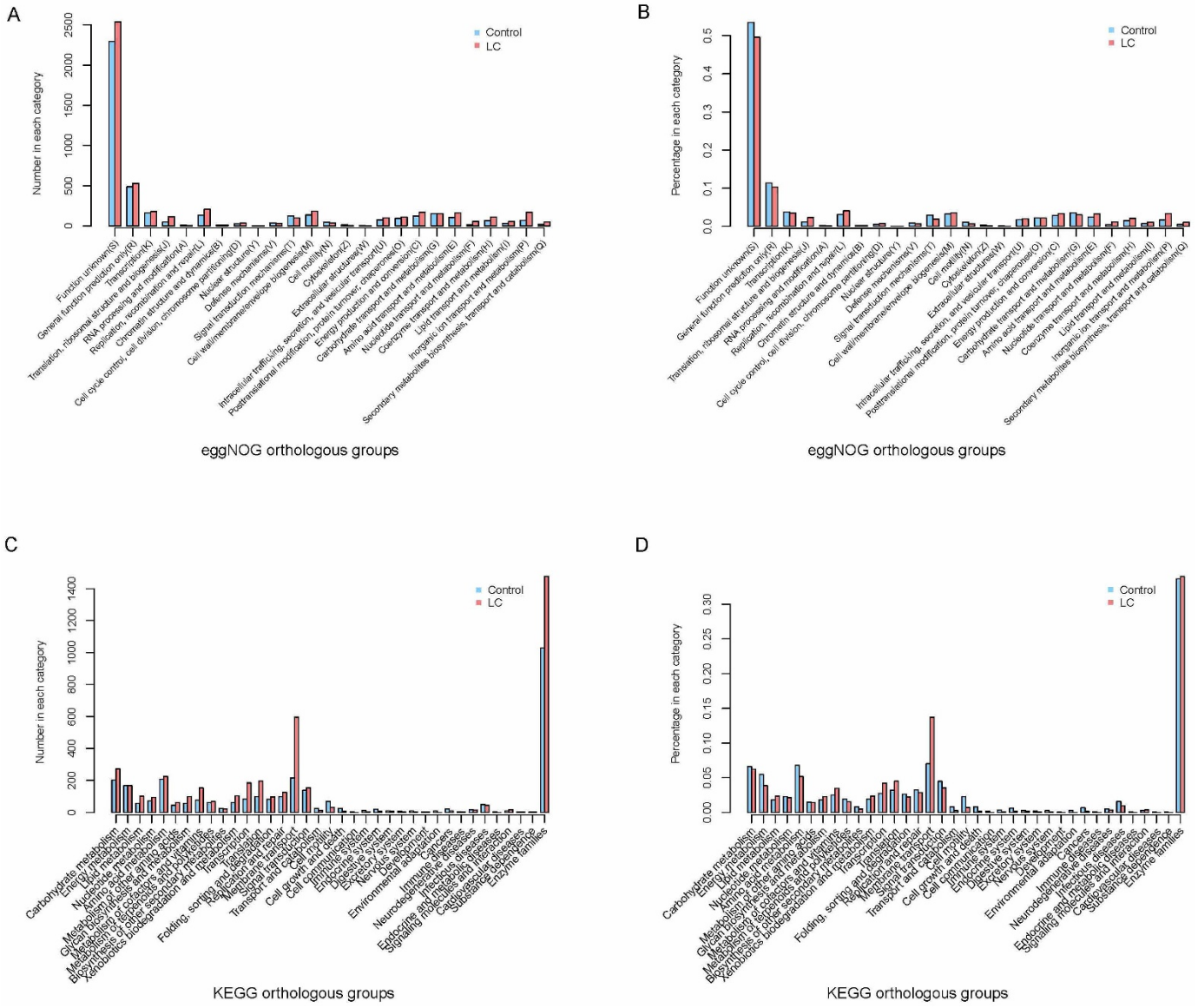
Extended Data Figure 3 | MGS enriched in healthy Chinese individuals ($n = 114$) are present in Danish individuals ($n = 292$). Presence and abundance of 50 'tracer' genes for each species; genes are in rows; abundance is indicated by colour gradient (white, not detected; red, most abundant). Individuals, ordered by increasing gene count, are in columns. Significance of correlation of species abundance (computed as mean abundance of the tracer genes) and gene count (q value, FDR adjusted) is given. Species in the Chinese cohort that were identical to those previously found, as correlated with the gene diversity in the Danish cohort[27], are highlighted in red. Left, the Chinese healthy cohort. Right, the Danish obesity cohort.

©2014 Macmillan Publishers Limited. All rights reserved



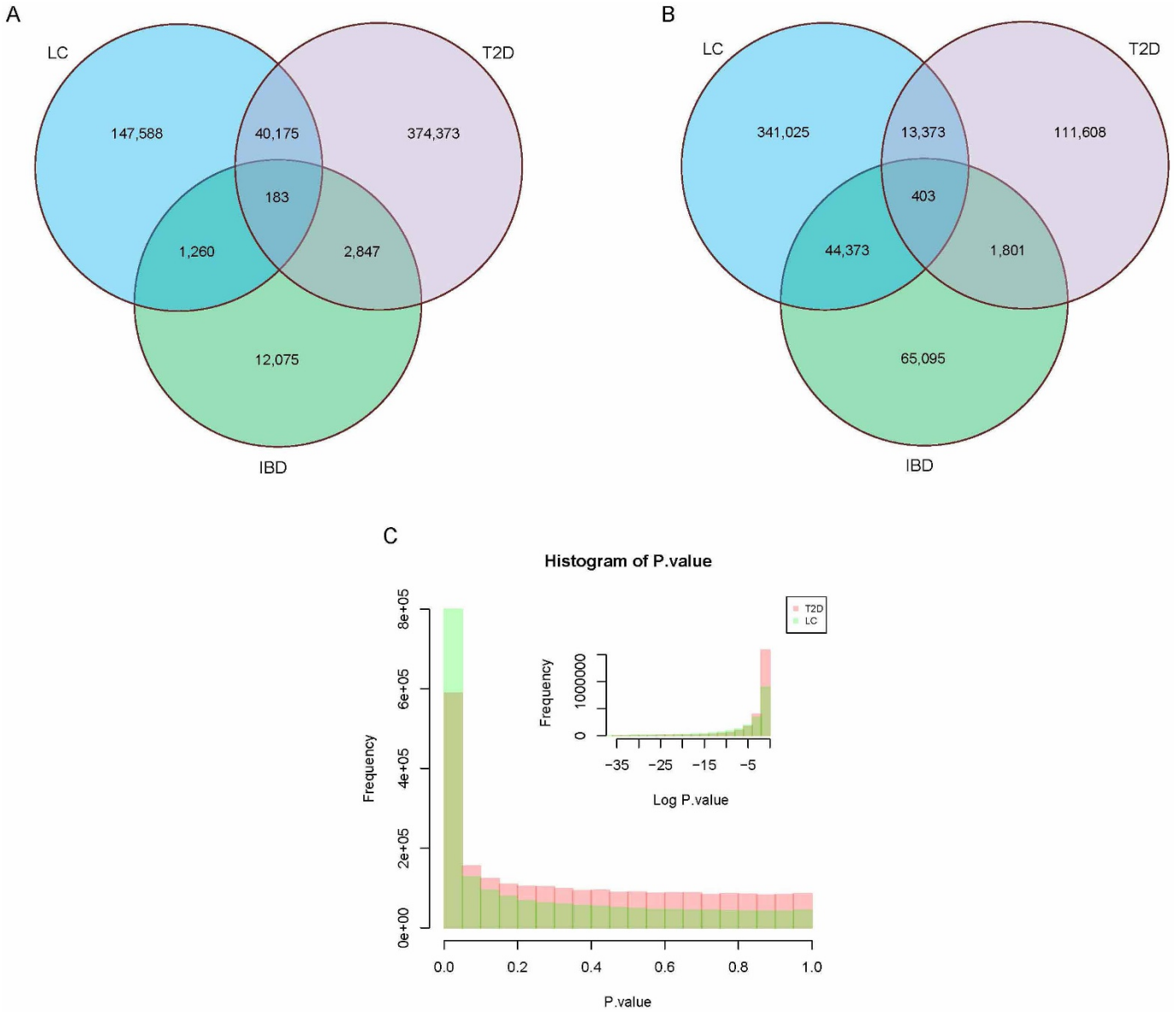
Extended Data Figure 4 | Massive changes in the gut microbiome in liver cirrhosis. Top left, healthy individuals have more gut microbial genes than patients with liver cirrhosis. Gene count was computed after downsizing the mapped reads to a level of 6.2 million (ref. 27). The significance of the difference was computed using a Student's *t*-test. Bottom, abundance of patient-enriched species ($n = 28$) in patients with liver cirrhosis ($n = 98$) and healthy controls ($n = 83$). 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 data set. Bar length indicates the relative abundance of a given species depicted by a different colour. Patients were ordered by the total patient-enriched species abundance; LPA and HPA quartiles ($n = 24$) are separated by red vertical lines. Top right, oral species are frequent in patients with liver cirrhosis. MGS enriched in healthy controls are largely not assigned to a species level, while those enriched in patients with liver cirrhosis are largely assigned to a species level and are mostly of oral origin (see Methods for species assignment).

©2014 Macmillan Publishers Limited. All rights reserved



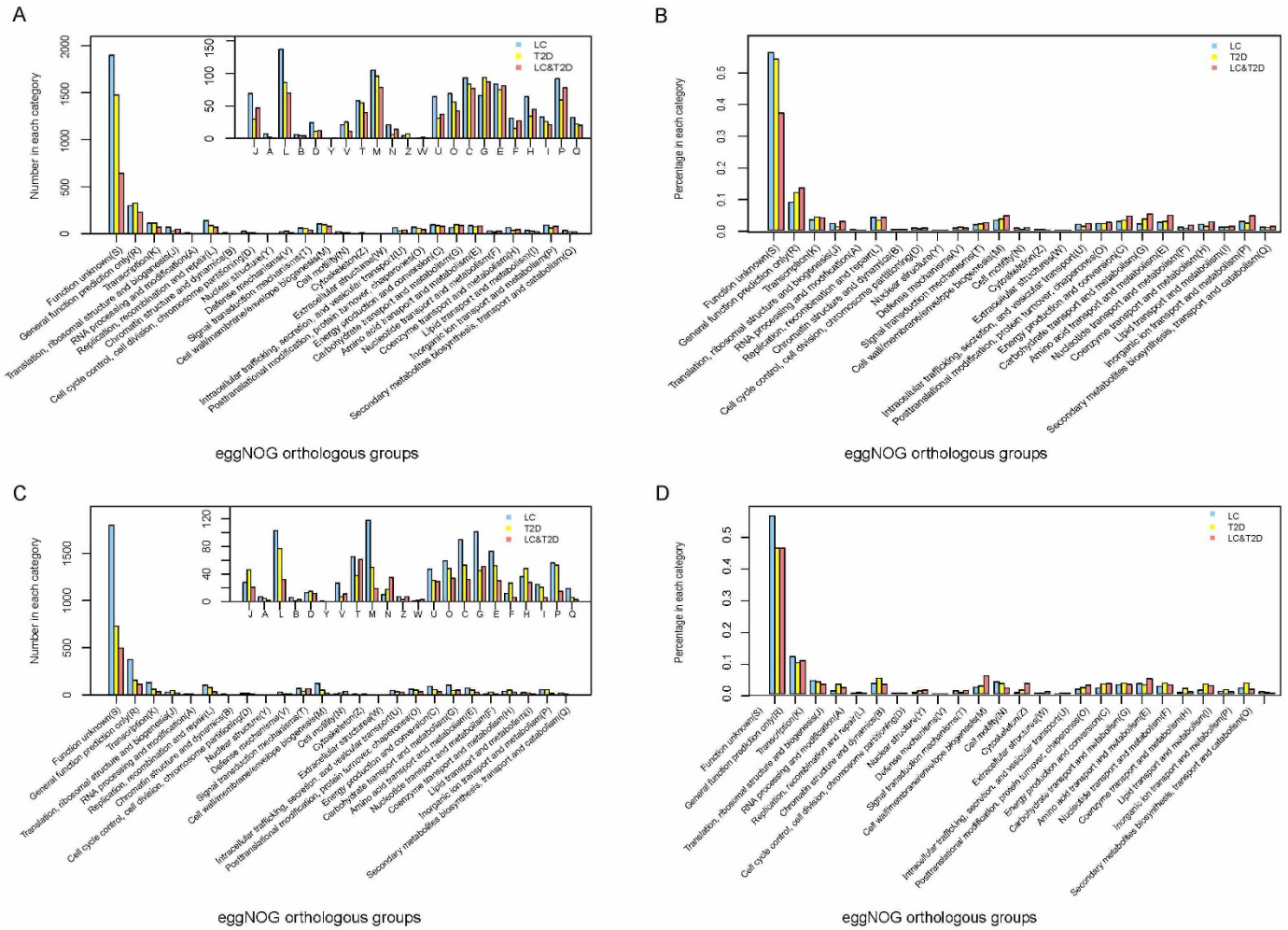
Extended Data Figure 5 | The distribution of eggNOG orthologue group and KEGG functional categories for liver-cirrhosis-related markers. **a**, Comparison between the liver-cirrhosis-enriched and control-enriched eggNOG orthologue group markers for 24 eggNOG orthologue group functional categories shown by number. **b**, Comparison between the liver-cirrhosis-enriched and control-enriched eggNOG orthologue group markers for 24 eggNOG orthologue group functional categories shown by percentage. **c**, Comparison between the liver-cirrhosis-enriched and control-enriched KEGG orthologue group markers for each KEGG functional category shown by number. **d**, Comparison between the liver-cirrhosis-enriched and control-enriched KEGG orthologue group markers for each KEGG functional category shown by percentage.

©2014 Macmillan Publishers Limited. All rights reserved



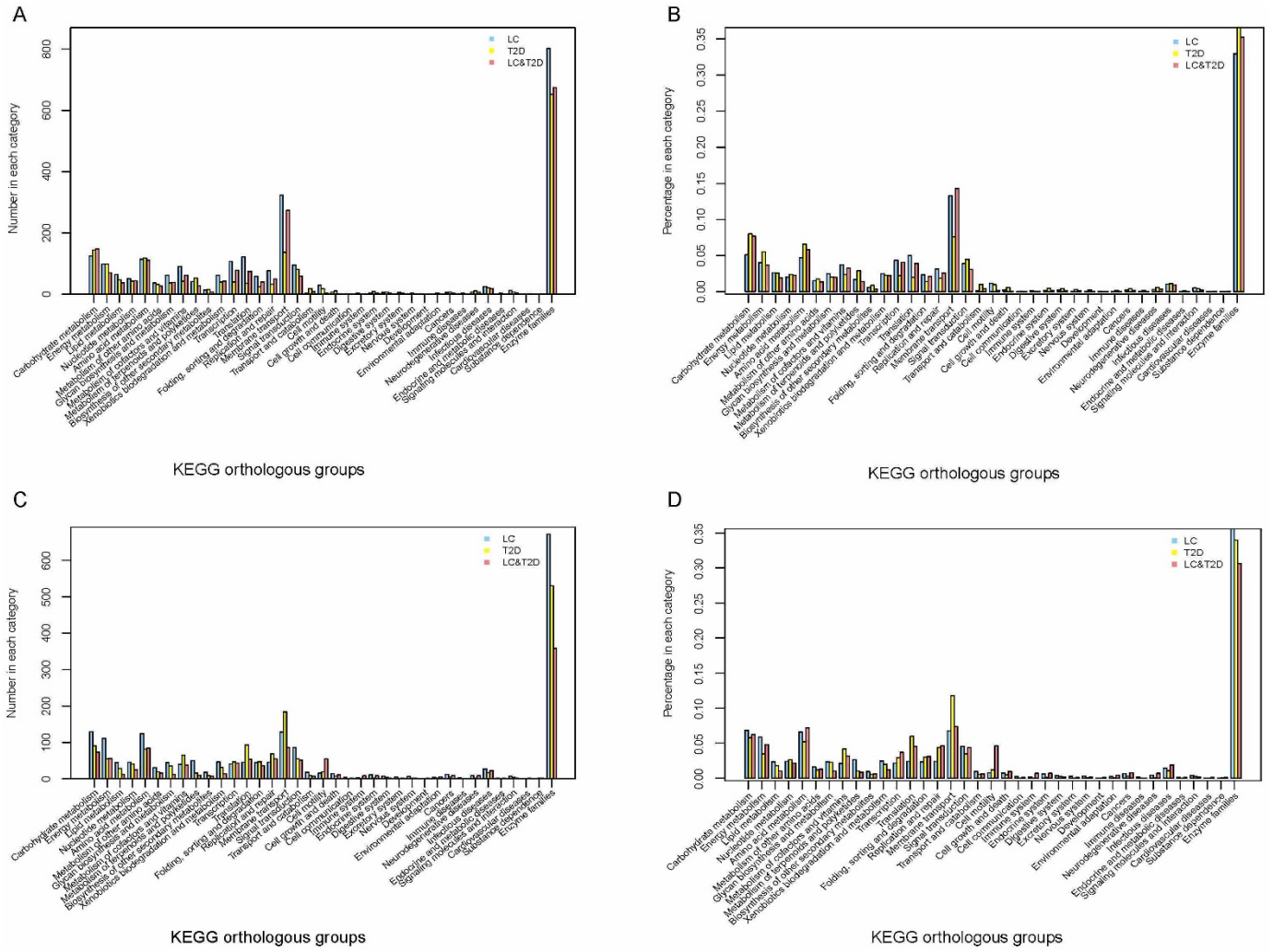
Extended Data Figure 6 | A comparison of the gene markers for the different groups. **a**, Venn diagram showing a gene marker comparison of case-enriched gene markers from the liver cirrhosis, T2D and IBD studies. **b**, Venn diagram showing a gene marker comparison of control-enriched gene markers from the liver cirrhosis, T2D and IBD studies. **c**, The length of the bar (*y* axis) represents the number of genes; the *P* value in the related range is shown on the *x* axis. The pink and light green bars show genes involved in type 2 diabetes and liver cirrhosis, respectively. Inset, the log *P* value of the gene markers between the two studies.

©2014 Macmillan Publishers Limited. All rights reserved



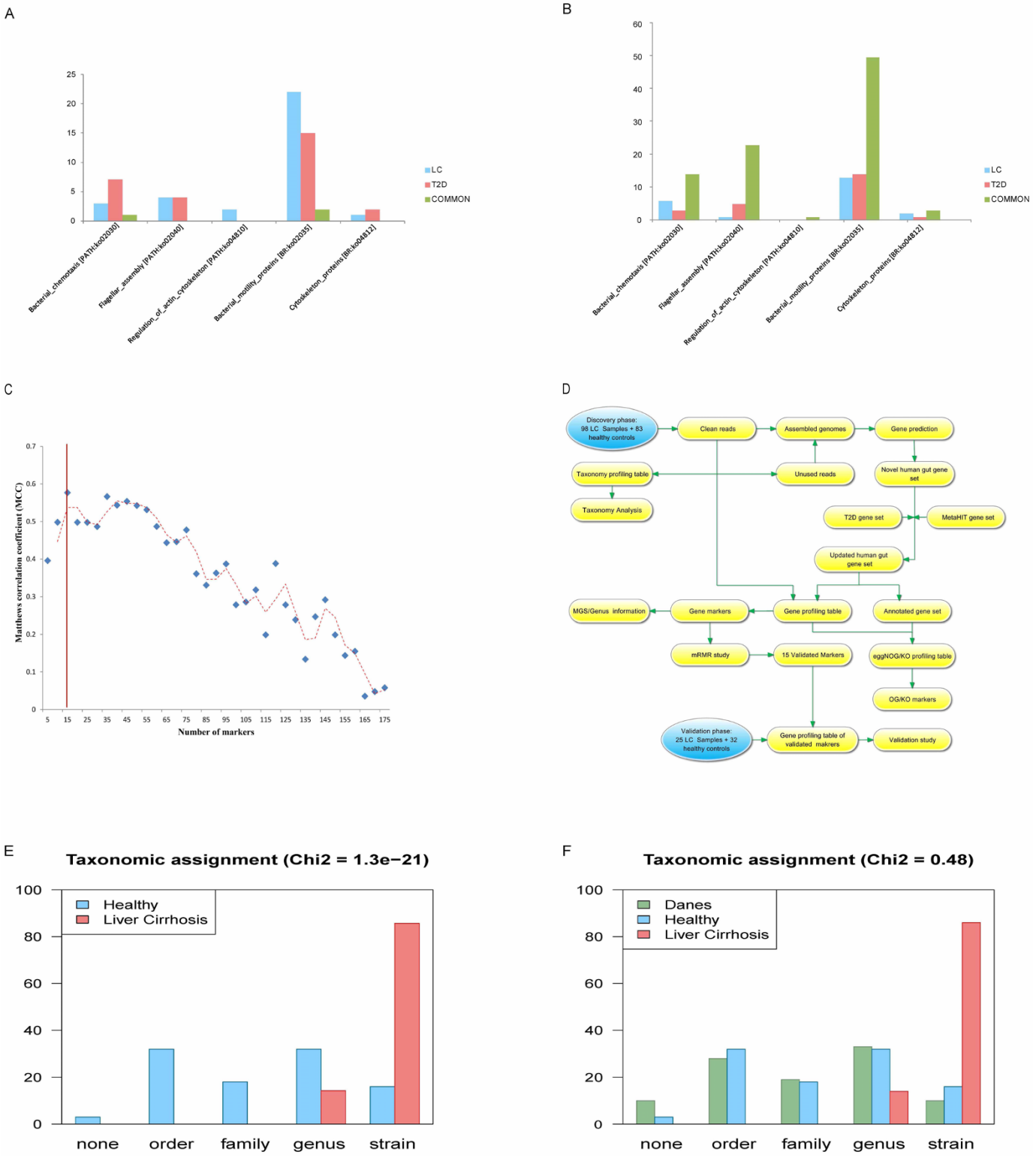
Extended Data Figure 7 | The distribution of eggNOG functional categories for case-enriched and control-enriched gene markers in liver cirrhosis only, T2D only and the liver cirrhosis/T2D groups. **a**, Comparison of the eggNOG orthologue group functional categories for case-enriched gene markers shown by number. **b**, Comparison of the eggNOG orthologue group functional categories for case-enriched gene markers shown by percentage. **c**, Comparison of the eggNOG orthologue group functional categories for the control-enriched gene markers shown by number. **d**, Comparison of the eggNOG orthologue group functional categories for the control-enriched gene markers shown by percentage.

©2014 Macmillan Publishers Limited. All rights reserved



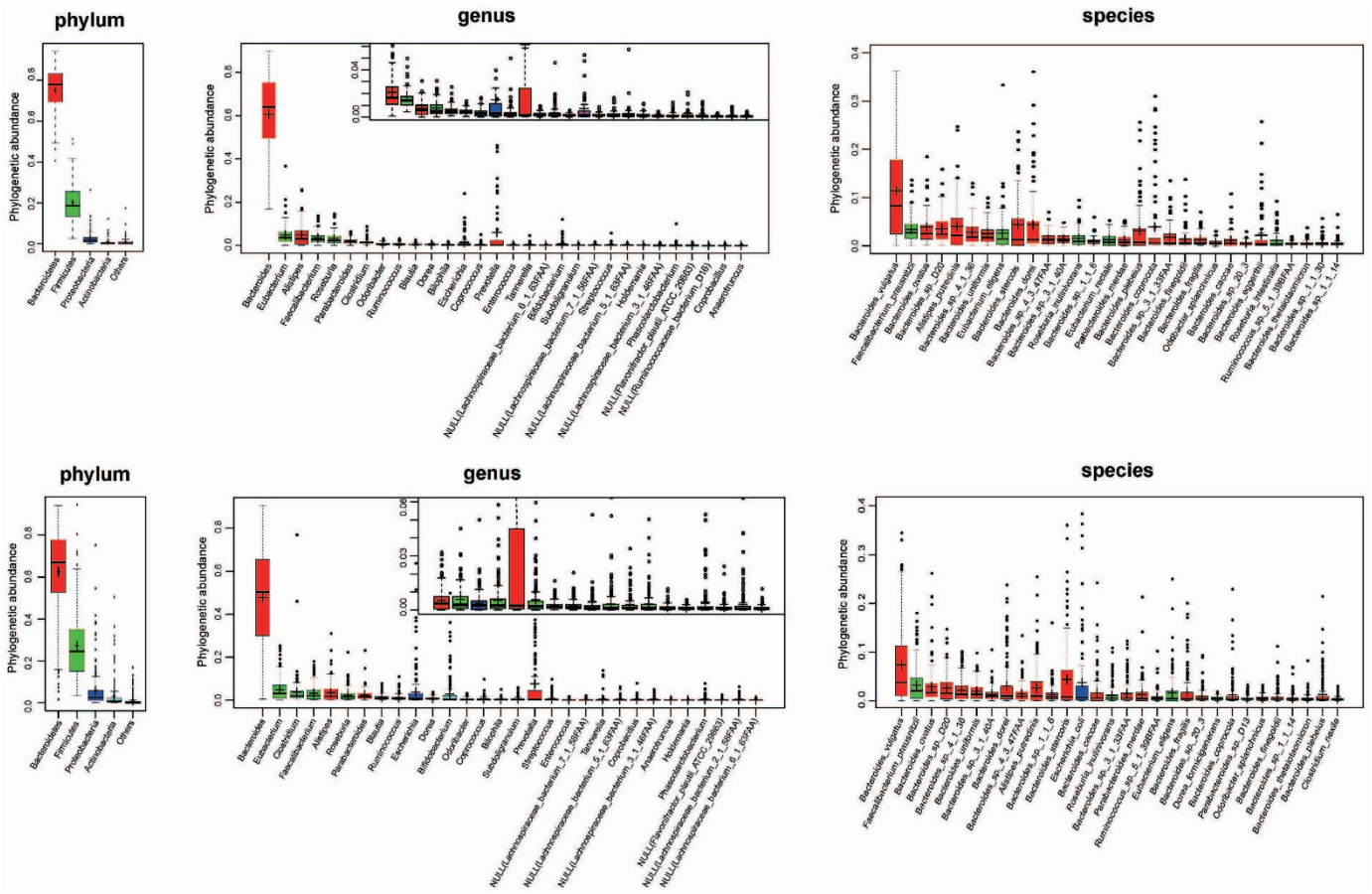
Extended Data Figure 8 | The distribution of the KEGG functional categories for case-enriched and control-enriched gene markers in liver cirrhosis only, T2D only or the liver cirrhosis/T2D group. **a**, Comparison of the KEGG pathway categories for the case-enriched gene markers shown by number. **b**, Comparison of the KEGG pathway categories for the case-enriched gene markers shown by percentage. **c**, Comparison of the KEGG pathway categories for the control-enriched gene markers shown by number. **d**, Comparison of the KEGG pathway categories for the control-enriched gene markers shown by percentage.

©2014 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 9 | Estimating the optimum number of markers and establishing the taxonomic assignment of MGS. **a**, Comparison of the case-enriched gene markers. **b**, Comparison of the control-enriched gene markers. **c**, The mRMR method was used to identify the liver-cirrhosis-associated markers. Sequential subsets were generated at five-marker intervals. For each subset, the error rate was estimated using a leave-one-out cross-validation of a linear discrimination classifier. The optimum (highest value of the Matthews correlation coefficient) subset contains 15 gene markers. **d**, The study included a discovery and a validation phase. Volunteers for both phases were recruited in the same hospital. Both direct read mapping and *de novo* assembly were performed for each sample. A taxonomy profiling table was established for taxonomy analysis. A novel gut gene set was established, and annotated. Identification of the MGS, finding markers and validating markers is also shown. **e**, MGS enriched in Chinese patients with liver cirrhosis and healthy individuals. Species-level assignment was deduced from the best BlastN hits of genes from a given MGS at thresholds of the average of more than 95% identity and more than 90% overlap with genes from a sequenced genome. For MGS where these thresholds were not reached, an assignment was attributed at the lowest taxonomy level where at least 80% of the genes had the same best hit BlastP taxonomy; in all cases these criteria held true at higher taxonomic levels. **f**, Taxonomic assignments of 58 species related to gut gene richness in a Danish cohort[27].

©2014 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 10 | Phylogenetic abundance of healthy controls in the discovery stage in the liver cirrhosis and T2D studies. The relative abundance of top bacterial phylotypes at the phylum, genus and species levels, respectively, in the liver cirrhosis study (top three panels) and in the T2D study (bottom three panels).

©2014 Macmillan Publishers Limited. All rights reserved