Vol 464 | 4 March 2010 | doi:10.1038/nature08821

A human gut microbial gene catalogue established by metagenomic sequencing

Junjie Qin[1]*, Ruiqiang Li[1]*, Jeroen Raes[2,3], Manimozhiyan Arumugam[2], Kristoffer Solvsten Burgdorf[4], Chaysavanh Manichanh[5], Trine Nielsen[4], Nicolas Pons[6], Florence Levenez[6], Takuji Yamada[2], Daniel R. Mende[2], Junhua Li[1,7], Junming Xu[1], Shaochuan Li[1], Dongfang Li[1,8], Jianjun Cao[1], Bo Wang[1], Huiqing Liang[1], Huisong Zheng[1], Yinlong Xie[1,7], Julien Tap[6], Patricia Lepage[6], Marcelo Bertalan[9], Jean-Michel Batto[6], Torben Hansen[4], Denis Le Paslier[10], Allan Linneberg[11], H. Bjørn Nielsen[9], Eric Pelletier[10], Pierre Renault[6], Thomas Sicheritz-Ponten[9], Keith Turner[12], Hongmei Zhu[1], Chang Yu[1], Shengting Li[1], Min Jian[1], Yan Zhou[1], Yingrui Li[1], Xiuqing Zhang[1], Songgang Li[1], Nan Qin[1], Huanming Yang[1], Jian Wang[1], Søren Brunak[9], Joel Doré[6], Francisco Guarner[5], Karsten Kristiansen[13], Oluf Pedersen[4,14], Julian Parkhill[12], Jean Weissenbach[10], MetaHIT Consortium†, Peer Bork[2], S. Dusko Ehrlich[6] & Jun Wang[1,13]

To understand the impact of gut microbes on human health and well-being it is crucial to assess their genetic potential. Here we describe the Illumina-based metagenomic sequencing, assembly and characterization of 3.3 million non-redundant microbial genes, derived from 576.7 gigabases of sequence, from faecal samples of 124 European individuals. The gene set, ~150 times larger than the human gene complement, contains an overwhelming majority of the prevalent (more frequent) microbial genes of the cohort and probably includes a large proportion of the prevalent human intestinal microbial genes. The genes are largely shared among individuals of the cohort. Over 99% of the genes are bacterial, indicating that the entire cohort harbours between 1,000 and 1,150 prevalent bacterial species and each individual at least 160 such species, which are also largely shared. We define and describe the minimal gut metagenome and the minimal gut bacterial genome in terms of functions present in all individuals and most bacteria, respectively.

It has been estimated that the microbes in our bodies collectively make up to 100 trillion cells, tenfold the number of human cells, and suggested that they encode 100-fold more unique genes than our own genome[1]. The majority of microbes reside in the gut, have a profound influence on human physiology and nutrition, and are crucial for human life[2,3]. Furthermore, the gut microbes contribute to energy harvest from food, and changes of gut microbiome may be associated with bowel diseases or obesity[4–8].

To understand and exploit the impact of the gut microbes on human health and well-being it is necessary to decipher the content, diversity and functioning of the microbial gut community. 16S ribosomal RNA gene (rRNA) sequence-based methods[9] revealed that two bacterial divisions, the Bacteroidetes and the Firmicutes, constitute over 90% of the known phylogenetic categories and dominate the distal gut microbiota[10]. Studies also showed substantial diversity of the gut microbiome between healthy individuals[4,8,10,11]. Although this difference is especially marked among infants[12], later in life the gut microbiome converges to more similar phyla.

Metagenomic sequencing represents a powerful alternative to rRNA sequencing for analysing complex microbial communities[13–15]. Applied to the human gut, such studies have already generated some 3 gigabases (Gb) of microbial sequence from faecal samples of 33 individuals from the United States or Japan[8,16,17]. To get a broader overview of the human gut microbial genes we used the Illumina Genome Analyser (GA) technology to carry out deep sequencing of total DNA from faecal samples of 124 European adults. We generated 576.7 Gb of sequence, almost 200 times more than in all previous studies, assembled it into contigs and predicted 3.3 million unique open reading frames (ORFs). This gene catalogue contains virtually all of the prevalent gut microbial genes in our cohort, provides a broad view of the functions important for bacterial life in the gut and indicates that many bacterial species are shared by different individuals. Our results also show that short-read metagenomic sequencing can be used for global characterization of the genetic potential of ecologically complex environments.

Metagenomic sequencing of gut microbiomes

As part of the MetaHIT (Metagenomics of the Human Intestinal Tract) project, we collected faecal specimens from 124 healthy, overweight and obese individual human adults, as well as inflammatory bowel disease (IBD) patients, from Denmark and Spain (Supplementary Table 1). Total DNA was extracted from the faecal specimens[18] and an average of 4.5 Gb (ranging between 2 and 7.3 Gb) of sequence was generated for each sample, allowing us to capture most of the

[1]BGI-Shenzhen, Shenzhen 518083, China. [2]European Molecular Biology Laboratory, 69117 Heidelberg, Germany. [3]VIB—Vrije Universiteit Brussel, 1050 Brussels, Belgium. [4]Hagedorn Research Institute, DK 2820 Copenhagen, Denmark. [5]Hospital Universitari Val d'Hebron, Ciberehd, 08035 Barcelona, Spain. [6]Institut National de la Recherche Agronomique, 78350 Jouy en Josas, France. [7]School of Software Engineering, South China University of Technology, Guangzhou 510641, China. [8]Genome Research Institute, Shenzhen University Medical School, Shenzhen 518000, China. [9]Center for Biological Sequence Analysis, Technical University of Denmark, DK-2800 Kongens Lyngby, Denmark. [10]Commissariat à l'Energie Atomique, Genoscope, 91000 Evry, France. [11]Research Center for Prevention and Health, DK-2600 Glostrup, Denmark. [12]The Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK. [13]Department of Biology, University of Copenhagen, DK-2200 Copenhagen, Denmark. [14]Institute of Biomedical Sciences, University of Copenhagen & Faculty of Health Science, University of Aarhus, 8000 Aarhus, Denmark.

*These authors contributed equally to this work.

†Lists of authors and affiliations appear at the end of the paper.

novelty (see Methods and Supplementary Table 2). In total, we obtained 576.7 Gb of sequence (Supplementary Table 3).

Wanting to generate an extensive catalogue of microbial genes from the human gut, we first assembled the short Illumina reads into longer contigs, which could then be analysed and annotated by standard methods. Using SOAPdenovo[19], a de Bruijn graph-based tool specially designed for assembling very short reads, we performed *de novo* assembly for all of the Illumina GA sequence data. Because a high diversity between individuals is expected[8,16,17], we first assembled each sample independently (Supplementary Fig. 3). As much as 42.7% of the Illumina GA reads was assembled into a total of 6.58 million contigs of a length >500 bp, giving a total contig length of 10.3 Gb, with an N50 length of 2.2 kb (Supplementary Fig. 4) and the range of 12.3 to 237.6 Mb (Supplementary Table 4). Almost 35% of reads from any one sample could be mapped to contigs from other samples, indicating the existence of a common sequence core.

To assess the quality of the Illumina GA-based assembly we mapped the contigs of samples MH0006 and MH0012 to the Sanger reads from the same samples (Supplementary Table 2). A total of 98.7% of the contigs that map to at least one Sanger read were collinear over 99.6% of the mapped regions. This is comparable to the contigs that were generated by 454 sequencing for one of the two samples (MH0006) as a control, of which 97.9% were collinear over 99.5% of the mapped regions. We estimate assembly errors to be 14.2 and 20.7 per megabase (Mb) of Illumina- and 454-based contigs, respectively (see Methods and Supplementary Fig. 5), indicating that the short- and long-read-based assemblies have comparable accuracies.

To complete the contig set we pooled the unassembled reads from all 124 samples, and repeated the *de novo* assembly process. About 0.4 million additional contigs were thus generated, having a length of 370 Mb and an N50 length of 939 bp. The total length of our final contig set was thus 10.7 Gb. Some 80% of the 576.7 Gb of Illumina GA sequence could be aligned to the contigs at a threshold of 90% identity, allowing for accommodation of sequencing errors and strain variability in the gut (Fig. 1), almost twice the 42.7% of sequence that was assembled into contigs by SOAPdenovo, because assembly uses more stringent criteria. This indicates that a vast majority of the Illumina sequence is represented by our contigs.

To compare the representation of the human gut microbiome in our contigs with that from previous work, we aligned them to the reads from the two largest published gut metagenome studies (1.83 Gb of Roche/454 sequencing reads from 18 US adults[8], and 0.79 Gb of Sanger reads from 13 Japanese adults and infants[17]), using the 90% identity threshold. A total of 70.1% and 85.9% of the reads from the Japanese and US samples, respectively, could be aligned to our contigs (Fig. 1), showing that the contigs include a high fraction of sequences from previous studies. In contrast, 85.7% and 69.5% of our contigs were not covered by the reads from the Japanese and US samples, respectively, highlighting the novelty we captured.

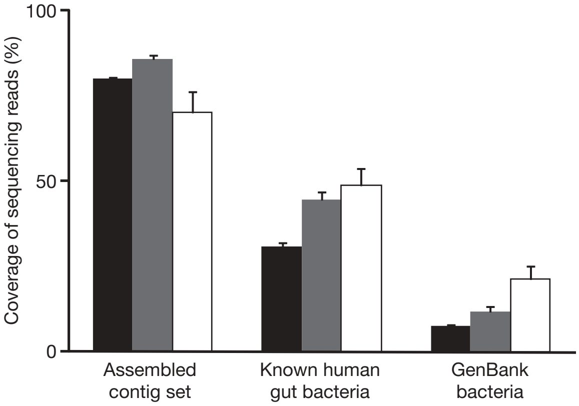

Figure 1 | Coverage of human gut microbiome. The three human microbial sequencing read sets—Illumina GA reads generated from 124 individuals in this study (black; $n = 124$), Roche/454 reads from 18 human twins and their mothers (grey; $n = 18$) and Sanger reads from 13 Japanese individuals (white; $n = 13$)—were aligned to each of the reference sequence sets. Mean values ± s.e.m. are plotted.

Only 31.0–48.8% of the reads from the two previous studies and the present study could be aligned to 194 public human gut bacterial genomes (Supplementary Table 5), and 7.6–21.2% to the bacterial genomes deposited in GenBank (Fig. 1). This indicates that the reference gene set obtained by sequencing genomes of isolated bacterial strains is still of a limited scale.

A gene catalogue of the human gut microbiome

To establish a non-redundant human gut microbiome gene set we first used the MetaGene[20] program to predict ORFs in our contigs and found 14,048,045 ORFs longer than 100 bp (Supplementary Table 6). They occupied 86.7% of the contigs, comparable to the value found for fully sequenced genomes (~86%). Two-thirds of the ORFs appeared incomplete, possibly due to the size of our contigs (N50 of 2.2 kb). We next removed the redundant ORFs, by pair-wise comparison, using a very stringent criterion of 95% identity over 90% of the shorter ORF length, which can fuse orthologues but avoids inflation of the data set due to possible sequencing errors (see Methods). Yet, the final non-redundant gene set contained as many as 3,299,822 ORFs with an average length of 704 bp (Supplementary Table 7).

We term the genes of the non-redundant set ‘prevalent genes’, as they are encoded on contigs assembled from the most abundant reads (see Methods). The minimal relative abundance of the prevalent genes was $\sim 6 \times 10^{-7}$, as estimated from the minimum sequence coverage of the unique genes (close to 3), and the total Illumina sequence length generated for each individual (on average, 4.5 Gb), assuming the average gene length of 0.85 kb (that is, $3 \times 0.85 \times 10^3 / 4.5 \times 10^9$).

We mapped the 3.3 million gut ORFs to the 319,812 genes (target genes) of the 89 frequent reference microbial genomes in the human gut. At a 90% identity threshold, 80% of the target genes had at least 80% of their length covered by a single gut ORF (Fig. 2b). This indicates that the gene set includes most of the known human gut bacterial genes.

We examined the number of prevalent genes identified across all individuals as a function of the extent of sequencing, demanding at least two supporting reads for a gene call (Fig. 2a). The incidence-based coverage richness estimator (ICE), determined at 100 individuals (the highest number the EstimateS[21] program could accommodate), indicates that our catalogue captures 85.3% of the prevalent genes. Although this is probably an underestimate, it nevertheless indicates that the catalogue contains an overwhelming majority of the prevalent genes of the cohort.

Each individual carried 536,112 ± 12,167 (mean ± s.e.m.) prevalent genes (Supplementary Fig. 6b), indicating that most of the 3.3 million gene pool must be shared. However, most of the prevalent genes were found in only a few individuals: 2,375,655 were present in less than 20%, whereas 294,110 were found in at least 50% of individuals (we term these ‘common’ genes). These values depend on the sampling depth; sequencing of MH0006 and MH0012 revealed more of the catalogue genes, present at a low abundance (Supplementary Fig. 7). Nevertheless, even at our routine sampling depth, each individual harboured 204,056 ± 3,603 (mean ± s.e.m.) common genes, indicating that about 38% of an individual’s total gene pool is shared. Interestingly, the IBD patients harboured, on average, 25% fewer genes than the individuals not suffering from IBD (Supplementary Fig. 8), consistent with the observation that the former have lower bacterial diversity than the latter[22].

Common bacterial core

Deep metagenomic sequencing provides the opportunity to explore the existence of a common set of microbial species (common core) in

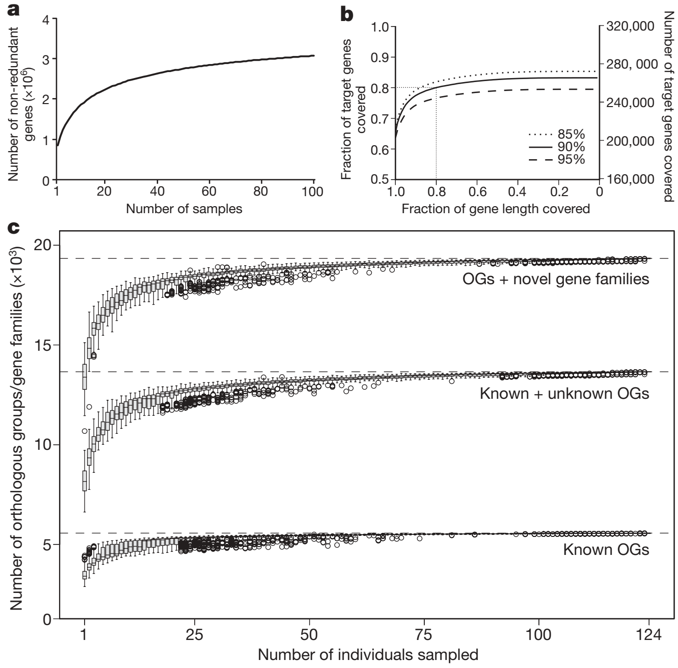

Figure 2 | Predicted ORFs in the human gut microbiome. a, Number of unique genes as a function of the extent of sequencing. The gene accumulation curve corresponds to the S_{obs} (Mao Tau) values (number of observed genes), calculated using EstimateS[21] (version 8.2.0) on randomly chosen 100 samples (due to memory limitation). **b**, Coverage of genes from 89 frequent gut microbial species (Supplementary Table 12). **c**, Number of functions captured by number of samples investigated, based on known (well characterized) orthologous groups (OGs; bottom), known plus unknown orthologous groups (including, for example, putative, predicted, conserved hypothetical functions; middle) and orthologous groups plus novel gene families (>20 proteins) recovered from the metagenome (top). Boxes denote the interquartile range (IQR) between the first and third quartiles (25th and 75th percentiles, respectively) and the line inside denotes the median. Whiskers denote the lowest and highest values within 1.5 times IQR from the first and third quartiles, respectively. Circles denote outliers beyond the whiskers.

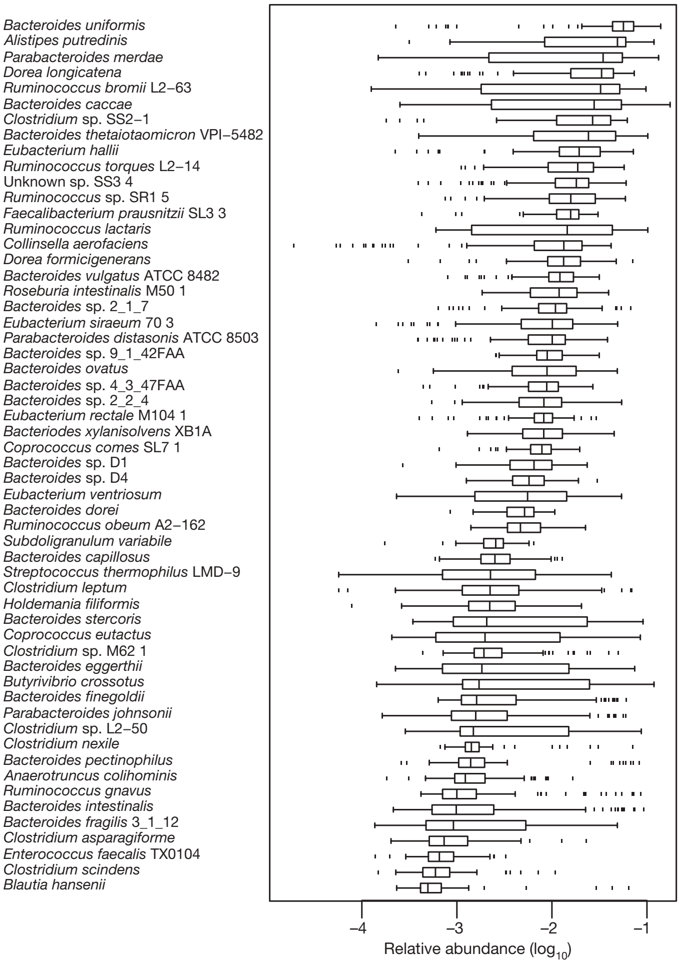

Figure 3 | Relative abundance of 57 frequent microbial genomes among individuals of the cohort. See Fig. 2c for definition of box and whisker plot. See Methods for computation.

the cohort. For this purpose, we used a non-redundant set of 650 sequenced bacterial and archaeal genomes (see Methods). We aligned the Illumina GA reads of each human gut microbial sample onto the genome set, using a 90% identity threshold, and determined the proportion of the genomes covered by the reads that aligned onto only a single position in the set. At a 1% coverage, which for a typical gut bacterial genome corresponds to an average length of about 40 kb, some 25-fold more than that of the 16S gene generally used for species identification, we detected 18 species in all individuals, 57 in ≥90% and 75 in ≥50% of individuals (Supplementary Table 8). At 10% coverage, requiring ~10-fold higher abundance in a sample, we still found 13 of the above species in ≥90% of individuals and 35 in ≥50%.

When the cumulated sequence length increased from 3.96 Gb to 8.74 Gb and from 4.41 Gb to 11.6 Gb, for samples MH0006 and MH0012, respectively, the number of strains common to the two at the 1% coverage threshold increased by 25%, from 135 to 169. This indicates the existence of a significantly larger common core than the one we could observe at the sequence depth routinely used for each individual.

The variability of abundance of microbial species in individuals can greatly affect identification of the common core. To visualize this variability, we compared the number of sequencing reads aligned to different genomes across the individuals of our cohort. Even for the most common 57 species present in ≥90% of individuals with genome coverage >1% (Supplementary Table 8), the inter-individual variability was between 12- and 2,187-fold (Fig. 3). As expected[10,23], Bacteroidetes and Firmicutes had the highest abundance.

A complex pattern of species relatedness, characterized by clusters at the genus and family levels, emerges from the analysis of the network based on the pair-wise Pearson correlation coefficients of 155 species present in at least one individual at ≥1% coverage (Supplementary Fig. 9). Prominent clusters include some of the most abundant gut species, such as members of the Bacteroidetes and *Dorea*/*Eubacterium*/*Ruminococcus* groups and also bifidobacteria, Proteobacteria and streptococci/lactobacilli groups. These observations indicate that similar constellations of bacteria may be present in different individuals of our cohort, for reasons that remain to be established.

The above result indicates that the Illumina-based bacterial profiling should reveal differences between the healthy individuals and patients. To test this hypothesis we compared the IBD patients and healthy controls (Supplementary Table 1), as it was previously reported that the two have different microbiota[22]. The principal component analysis, based on the same 155 species, clearly separates patients from healthy individuals and the ulcerative colitis from the Crohn's disease patients (Fig. 4), confirming our hypothesis.

Functions encoded by the prevalent gene set

We classified the predicted genes by aligning them to the integrated NCBI-NR database of non-redundant protein sequences, the genes in the KEGG (Kyoto Encyclopedia of Genes and Genomes)[24] pathways, and COG (Clusters of Orthologous Groups)[25] and eggNOG[26] databases. There were 77.1% genes classified into phylotypes, 57.5% to eggNOG clusters, 47.0% to KEGG orthology and 18.7% genes assigned to KEGG pathways, respectively (Supplementary Table 9).

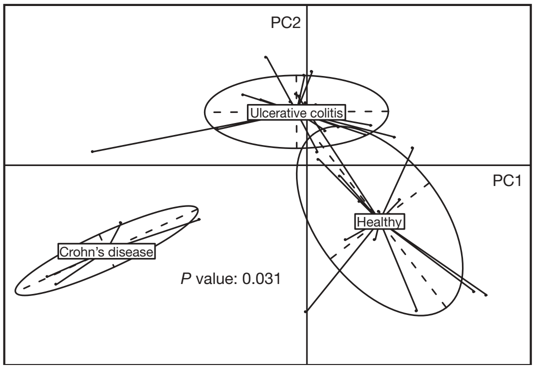

Figure 4 | Bacterial species abundance differentiates IBD patients and healthy individuals. Principal component analysis with health status as instrumental variables, based on the abundance of 155 species with ≥1% genome coverage by the Illumina reads in at least 1 individual of the cohort, was carried out with 14 healthy individuals and 25 IBD patients (21 ulcerative colitis and 4 Crohn's disease) from Spain (Supplementary Table 1). Two first components (PC1 and PC2) were plotted and represented 7.3% of whole inertia. Individuals (represented by points) were clustered and centre of gravity computed for each class; *P*-value of the link between health status and species abundance was assessed using a Monte-Carlo test (999 replicates).

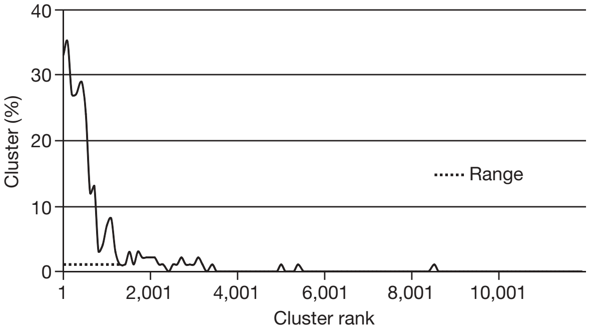

Figure 5 | Clusters that contain the *B. subtilis* essential genes. The clusters were ranked by the number of genes they contain, normalized by average length and copy number (see Supplementary Fig. 10), and the proportion of clusters with the essential *B. subtilis* genes was determined for successive groups of 100 clusters. Range indicates the part of the cluster distribution that contains 86% of the *B. subtilis* essential genes.

Almost all (99.96%) of the phylogenetically assigned genes belonged to the Bacteria and Archaea, reflecting their predominance in the gut. Genes that were not mapped to orthologous groups were clustered into gene families (see Methods). To investigate the functional content of the prevalent gene set we computed the total number of orthologous groups and/or gene families present in any combination of n individuals (with $n = 2$–124; see Fig. 2c). This rarefaction analysis shows that the 'known' functions (annotated in eggNOG or KEGG) quickly saturate (a value of 5,569 groups was observed): when sampling any subset of 50 individuals, most have been detected. However, three-quarters of the prevalent gut functionalities consists of uncharacterized orthologous groups and/or completely novel gene families (Fig. 2c). When including these groups, the rarefaction curve only starts to plateau at the very end, at a much higher level (19,338 groups were detected), confirming that the extensive sampling of a large number of individuals was necessary to capture this considerable amount of novel/unknown functionality.

Bacterial functions important for life in the gut

The extensive non-redundant catalogue of the bacterial genes from the human intestinal tract provides an opportunity to identify bacterial functions important for life in this environment. There are functions necessary for a bacterium to thrive in a gut context (that is, the 'minimal gut genome') and those involved in the homeostasis of the whole ecosystem, encoded across many species (the 'minimal gut metagenome'). The first set of functions is expected to be present in most or all gut bacterial species; the second set in most or all individuals' gut samples.

To identify the functions encoded by the minimal gut genome we use the fact that they should be present in most or all gut bacterial species and therefore appear in the gene catalogue at a frequency above that of the functions present in only some of the gut bacterial species. The relative frequency of different functions can be deduced from the number of genes recruited to different eggNOG clusters, after normalization for gene length and copy number (Supplementary Fig. 10a, b). We ranked all the clusters by gene frequencies and determined the range that included the clusters specifying well-known essential bacterial functions, such as those determined experimentally for a well-studied firmicute, *Bacillus subtilis*[27], hypothesizing that additional clusters in this range are equally important. As expected, the range that included most of *B. subtilis* essential clusters (86%) was at the very top of the ranking order (Fig. 5). Some 76% of the clusters with essential genes of *Escherichia coli*[28] were within this range, confirming the validity of our approach. This suggests that 1,244 metagenomic clusters found within the range (Supplementary Table 10; termed 'range clusters' hereafter) specify functions important for life in the gut.

We found two types of functions among the range clusters: those required in all bacteria (housekeeping) and those potentially specific for the gut. Among many examples of the first category are the functions that are part of main metabolic pathways (for example, central carbon metabolism, amino acid synthesis), and important protein complexes (RNA and DNA polymerase, ATP synthase, general secretory apparatus). Not surprisingly, projection of the range clusters on the KEGG metabolic pathways gives a highly integrated picture of the global gut cell metabolism (Fig. 6a).

The putative gut-specific functions include those involved in adhesion to the host proteins (collagen, fibrinogen, fibronectin) or in harvesting sugars of the globoseries glycolipids, which are carried on blood and epithelial cells. Furthermore, 15% of range clusters encode functions that are present in <10% of the eggNOG genomes (see Supplementary Fig. 11) and are largely (74.3%) not defined (Fig. 6b). Detailed studies of these should lead to a deeper comprehension of bacterial life in the gut.

To identify the functions encoded by the minimal gut metagenome, we computed the orthologous groups that are shared by individuals of our cohort. This minimal set, of 6,313 functions, is much larger than the one estimated in a previous study[8]. There are only 2,069 functionally annotated orthologous groups, showing that they gravely underestimate the true size of the common functional complement among individuals (Fig. 6c). The minimal gut metagenome includes a considerable fraction of functions (~45%) that are present in <10% of the sequenced bacterial genomes (Fig. 6c, inset). These otherwise rare functionalities that are found in each of the 124 individuals may be necessary for the gut ecosystem. Eighty per cent of these orthologous groups contain genes with at best poorly characterized function, underscoring our limited knowledge of gut functioning.

Of the known fraction, about 5% codes for (pro)phage-related proteins, implying a universal presence and possible important ecological role of bacteriophages in gut homeostasis. The most striking secondary metabolism that seems crucial for the minimal metagenome relates, not unexpectedly, to biodegradation of complex sugars and glycans harvested from the host diet and/or intestinal lining. Examples include degradation and uptake pathways for pectin (and its monomer, rhamnose) and sorbitol, sugars which are omnipresent in fruits and vegetables, but which are not or poorly absorbed by humans. As some gut microorganisms were found to degrade both of them[29,30], this capacity seems to be selected for by the gut ecosystem as a non-competitive source of energy. Besides these, capacity to ferment, for example, mannose, fructose, cellulose and sucrose is also part of the minimal metagenome. Together, these emphasize the

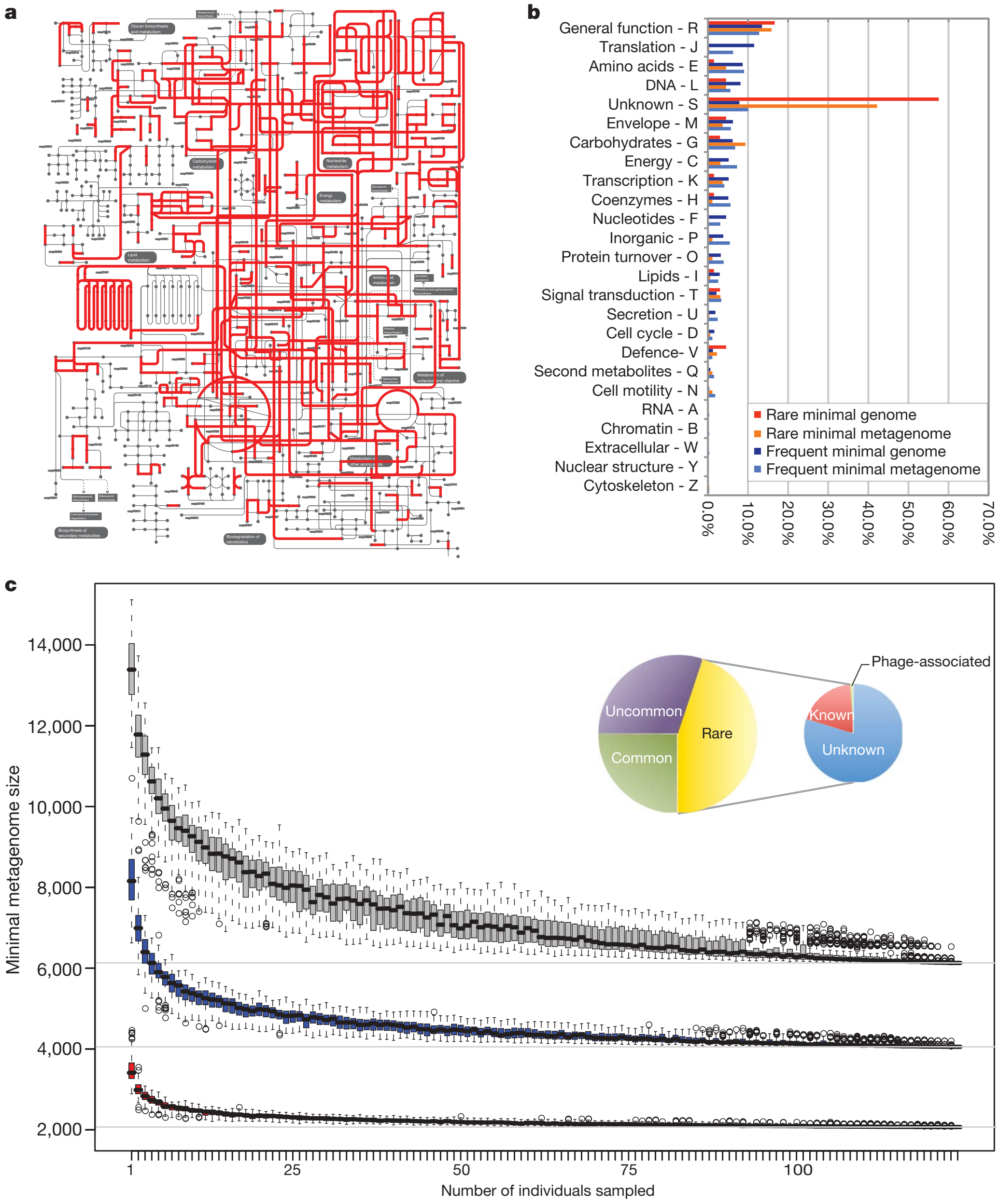

Figure 6 | Characterization of the minimal gut genome and metagenome. **a**, Projection of the minimal gut genome on the KEGG pathways using the iPath tool[38]. **b**, Functional composition of the minimal gut genome and metagenome. Rare and frequent refer to the presence in sequenced eggNOG genomes. **c**, Estimation of the minimal gut metagenome size. Known orthologous groups (red), known plus unknown orthologous groups (blue) and orthologous groups plus novel gene families (>20 proteins; grey) are shown (see Fig. 2c for definition of box and whisker plot). The inset shows composition of the gut minimal microbiome. Large circle: classification in the minimal metagenome according to orthologous group occurrence in STRING7[39] bacterial genomes. Common (25%), uncommon (35%) and rare (45%) refer to functions that are present in >50%, <50% but >10%, and <10% of STRING bacteria genomes, respectively. Small circle: composition of the rare orthologous groups. Unknown (80%) have no annotation or are poorly characterized, whereas known bacterial (19%) and phage-related (1%) orthologous groups have functional description.

strong dependence of the gut ecosystem on complex sugar degradation for its functioning.

Functional complementarities of the genome and metagenome

Detailed analysis of the complementarities between the gut metagenome and the human genome is beyond the scope of the present work. To provide an overview, we considered two factors: conservation of the functions in the minimal metagenome and presence/absence of functions in one or the other (Supplementary Table 11). Gut bacteria use mostly fermentation to generate energy, converting sugars, in part, to short-chain fatty acid, that are used by the host as energy source. Acetate is important for muscle, heart and brain cells[31], propionate is used in host hepatic neoglucogenic processes, whereas, in addition, butyrate is important for enterocytes[32]. Beyond short-chain fatty acid, a number of amino acids are indispensable to humans[33] and can be provided by bacteria[34]. Similarly, bacteria can contribute certain vitamins[3] (for example, biotin, phylloquinone) to the host. All of the steps of biosynthesis of these molecules are encoded by the minimal metagenome.

Gut bacteria seem to be able to degrade numerous xenobiotics, including non-modified and halogenated aromatic compounds (Supplementary Table 11), even if the steps of most pathways are not part of the minimal metagenome and are found in a fraction of individuals only. A particularly interesting example is that of benzoate, which is a common food supplement, known as E211. Its degradation by the coenzyme-A ligation pathway, encoded in the minimal metagenome, leads to pimeloyl-coenzyme-A, which is a precursor of biotin, indicating that this food supplement can have a potentially beneficial role for human health.

Discussion

We have used extensive Illumina GA short-read-based sequencing of total faecal DNA from a cohort of 124 individuals of European (Nordic and Mediterranean) origin to establish a catalogue of non-redundant human intestinal microbial genes. The catalogue contains 3.3 million microbial genes, 150-fold more than the human gene complement, and includes an overwhelming majority (>86%) of prevalent genes harboured by our cohort. The catalogue probably contains a large majority of prevalent intestinal microbial genes in the human population, for the following reasons: (1) over 70% of the metagenomic reads from three previous studies, including American and Japanese individuals[8,16,17], can be mapped on our contigs; (2) about 80% of the microbial genes from 89 frequent gut reference genomes are present in our set. This result represents a proof of principle that short-read sequencing can be used to characterize complex microbiomes.

The full bacterial gene complement of each individual was not sampled in our work. Nevertheless, we have detected some 536,000 prevalent unique genes in each, out of the total of 3.3 million carried by our cohort. Inevitably, the individuals largely share the genes of the common pool. At the present depth of sequencing, we found that almost 40% of the genes from each individual are shared with at least half of the individuals of the cohort. Future studies of world-wide span, envisaged within the International Human Microbiome Consortium, will complete, as necessary, our gene catalogue and establish boundaries to the proportion of shared genes.

Essentially all (99.1%) of the genes of our catalogue are of bacterial origin, the remainder being mostly archaeal, with only 0.1% of eukaryotic and viral origins. The gene catalogue is therefore equivalent to that of some 1,000 bacterial species with an average-sized genome, encoding about 3,364 non-redundant genes. We estimate that no more than 15% of prevalent genes of our cohort may be missing from the catalogue, and suggest that the cohort harbours no more than ~1,150 bacterial species abundant enough to be detected by our sampling. Given the large overlap between microbial sequences in this and previous studies we suggest that the number of abundant intestinal bacterial species may be not much higher than that observed in our cohort. Each individual of our cohort harbours at least 160 such bacterial species, as estimated by the average prevalent gene number, and many must thus be shared.

We assigned about 12% of the reference set genes (404,000) to the 194 sequenced intestinal bacterial genomes, and can thus associate them with bacterial species. Sequencing of at least 1,000 human-associated bacterial genomes is foreseen within the International Human Microbiome Consortium, via the Human Microbiome Project and MetaHIT. This is commensurate with the number of dominant species in our cohort and expected more broadly in human gut, and should enable a much more extensive gene to species assignment. Nevertheless, we used the presently available sequenced genomes to explore further the concept of largely shared species among our cohort and identified 75 species common to >50% of individuals and 57 species common to >90%. These numbers are likely to increase with the number of sequenced reference strains and a deeper sampling. Indeed, a 2–3-fold increase in sequencing depth raised by 25% the number of species that we could detect as shared between two individuals. A large number of shared species supports the view that the prevalent human microbiome is of a finite and not overly large size.

How can this view be reconciled with that of a considerable interpersonal diversity of innumerable bacterial species in the gut, arising from most previous studies using the 16S RNA marker gene[4,8,10,11]? Possibly the depth of sampling of these studies was insufficient to reveal common species when present at low abundance, and emphasized the difference in the composition of a relatively few dominant species. We found a very high variability of abundance (12- to 2,200-fold) for the 57 most common species across the individuals of our cohort. Nevertheless, a recent 16S rRNA-based study concluded that a common bacterial species 'core', shared among at least 50% of individuals under study, exists[35].

Detailed comparisons of bacterial genes across the individuals of our cohort will be carried out in the future, within the context of the ongoing MetaHIT clinical studies of which they are part. Nevertheless, clustering of the genes in families allowed us to capture a virtually full functional potential of the prevalent gene set and revealed a considerable novelty, extending the functional categories by some 30% in regard to previous work[8]. Similarly, this analysis has revealed a functional core, conserved in each individual of the cohort, which reflects the full minimal human gut metagenome, encoded across many species and probably required for the proper functioning of the gut ecosystem. The size of this minimal metagenome exceeds several-fold that of the core metagenome reported previously[8]. It includes functions known to be important to the host–bacterial interaction, such as degradation of complex polysaccharides, synthesis of short-chain fatty acids, indispensable amino acids and vitamins. Finally, we also identified functions that we attribute to a minimal gut bacterial genome, likely to be required by any bacterium to thrive in this ecosystem. Besides general housekeeping functions, the minimal genome encompasses many genes of unknown function, rare in sequenced genomes and possibly specifically required in the gut.

Beyond providing the global view of the human gut microbiome, the extensive gene catalogue we have established enables future studies of association of the microbial genes with human phenotypes and, even more broadly, human living habits, taking into account the environment, including diet, from birth to old age. We anticipate that these studies will lead to a much more complete understanding of human biology than the one we presently have.

METHODS SUMMARY

Human faecal samples were collected, frozen immediately and DNA was purified by standard methods[22]. For all 124 individuals, paired-end libraries were constructed with different clone insert sizes and subjected to Illumina GA sequencing. All reads were assembled using SOAPdenovo[19], with specific parameter '−M 3' for metagenomics data. MetaGene was used for gene prediction. A non-redundant gene set was constructed by pair-wise comparison of all genes, using BLAT[36] under the criteria of identity >95% and overlap >90%. Gene taxonomic assignments were made on the basis of BLASTP[37] search (*e*-value $<1 \times 10^{-5}$) of the NCBI-NR database and 126 known gut bacteria genomes. Gene functional annotations were made by BLASTP search (*e*-value $<1 \times 10^{-5}$) with eggNOG and KEGG (v48.2) databases. The total and shared number of orthologous groups and/or gene families were computed using a random combination of *n* individuals (with $n = 2$ to 124, 100 replicates per bin).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 14 August; accepted 23 December 2009.

1. Ley, R. E., Peterson, D. A. & Gordon, J. I. Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell* **124**, 837–848 (2006).
2. Backhed, F., Ley, R. E., Sonnenburg, J. L., Peterson, D. A. & Gordon, J. I. Host-bacterial mutualism in the human intestine. *Science* **307**, 1915–1920 (2005).
3. Hooper, L. V., Midtvedt, T. & Gordon, J. I. How host-microbial interactions shape the nutrient environment of the mammalian intestine. *Annu. Rev. Nutr.* **22**, 283–307 (2002).
4. Ley, R. E., Turnbaugh, P. J., Klein, S. & Gordon, J. I. Microbial ecology: human gut microbes associated with obesity. *Nature* **444**, 1022–1023 (2006).
5. Turnbaugh, P. J. *et al.* An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444**, 1027–1031 (2006).
6. Ley, R. E. *et al.* Obesity alters gut microbial ecology. *Proc. Natl Acad. Sci. USA* **102**, 11070–11075 (2005).
7. Zhang, H. *et al.* Human gut microbiota in obesity and after gastric bypass. *Proc. Natl Acad. Sci. USA* **106**, 2365–2370 (2009).
8. Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457**, 480–484 (2009).
9. Zoetendal, E. G., Akkermans, A. D. & De Vos, W. M. Temperature gradient gel electrophoresis analysis of 16S rRNA from human fecal samples reveals stable and host-specific communities of active bacteria. *Appl. Environ. Microbiol.* **64**, 3854–3859 (1998).
10. Eckburg, P. B. *et al.* Diversity of the human intestinal microbial flora. *Science* **308**, 1635–1638 (2005).

11. Ley, R. E., Lozupone, C. A., Hamady, M., Knight, R. & Gordon, J. I. Worlds within worlds: evolution of the vertebrate gut microbiota. *Nature Rev. Microbiol.* **6,** 776–788 (2008).
12. Palmer, C., Bik, E. M., Digiulio, D. B., Relman, D. A. & Brown, P. O. Development of the human infant intestinal microbiota. *PLoS Biol.* **5,** e177 (2007).
13. Riesenfeld, C. S., Schloss, P. D. & Handelsman, J. Metagenomics: genomic analysis of microbial communities. *Annu. Rev. Genet.* **38,** 525–552 (2004).
14. von Mering, C. *et al.* Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science* **315,** 1126–1130 (2007).
15. Tringe, S. G. & Rubin, E. M. Metagenomics: DNA sequencing of environmental samples. *Nature Rev. Genet.* **6,** 805–814 (2005).
16. Gill, S. R. *et al.* Metagenomic analysis of the human distal gut microbiome. *Science* **312,** 1355–1359 (2006).
17. Kurokawa, K. *et al.* Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res.* **14,** 169–181 (2007).
18. Suau, A. *et al.* Direct analysis of genes encoding 16S rRNA from complex communities reveals many novel molecular species within the human gut. *Appl. Environ. Microbiol.* **65,** 4799–4807 (1999).
19. Li, R. & Zhu, H. *De novo* assembly of the human genomes with massively parallel short read sequencing. *Genome Res.* doi:10.1101/gr.097261.109 (17 December 2009).
20. Noguchi, H., Park, J. & Takagi, T. MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res.* **34,** 5623–5630 (2006).
21. Colwell, R. K. EstimateS: Statistical estimation of species richness and shared species from samples, version 8.2. ⟨http://viceroy.eeb.uconn.edu/estimates⟩ (1997).
22. Manichanh, C. *et al.* Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. *Gut* **55,** 205–211 (2006).
23. Wang, X., Heazlewood, S. P., Krause, D. O. & Florin, T. H. Molecular characterization of the microbial species that colonize human ileal and colonic mucosa by using 16S rDNA sequence analysis. *J. Appl. Microbiol.* **95,** 508–520 (2003).
24. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32,** D277–D280 (2004).
25. Tatusov, R. L. *et al.* The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4,** 41 (2003).
26. Jensen, L. J. *et al.* eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res.* **36,** D250–D254 (2008).
27. Kobayashi, K. *et al.* Essential *Bacillus subtilis* genes. *Proc. Natl Acad. Sci. USA* **100,** 4678–4683 (2003).
28. Baba, T. *et al.* Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* **2,** doi: 10.1038/msb4100050 (2006).
29. Dongowski, G., Lorenz, A. & Anger, H. Degradation of pectins with different degrees of esterification by *Bacteroides thetaiotaomicron* isolated from human gut flora. *Appl. Environ. Microbiol.* **66,** 1321–1327 (2000).
30. Cummings, J. H. & Macfarlane, G. T. The control and consequences of bacterial fermentation in the human colon. *J. Appl. Bacteriol.* **70,** 443–459 (1991).
31. Wong, J. M., de Souza, R., Kendall, C. W., Emam, A. & Jenkins, D. J. Colonic health: fermentation and short chain fatty acids. *J. Clin. Gastroenterol.* **40,** 235–243 (2006).
32. Hamer, H. M. *et al.* The role of butyrate on colonic function. *Aliment. Pharmacol. Ther.* **27,** 104–119 (2008).
33. Elango, R., Ball, R. O. & Pencharz, P. B. Amino acid requirements in humans: with a special emphasis on the metabolic availability of amino acids. *Amino Acids* **37,** 19–27 (2009).
34. Metges, C. C. Contribution of microbial amino acids to amino acid homeostasis of the host. *J. Nutr.* **130,** 1857S–1864S (2000).
35. Tap, J. *et al.* Towards the human intestinal microbiota phylogenetic core. *Environ. Microbiol.* **11,** 2574–2584 (2009).
36. Kent, W. J. BLAT–the BLAST-like alignment tool. *Genome Res.* **12,** 656–664 (2002).
37. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).
38. Letunic, I., Yamada, T., Kanehisa, M. & Bork, P. iPath: interactive exploration of biochemical pathways and networks. *Trends Biochem Sci.* **33,** 101–103 (2008).
39. von Mering, C. *et al.* STRING 7—recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.* **35,** D358–D362 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are indebted to the faculty and staff of Beijing Genomics Institute at Shenzhen, whose names were not included in the author list, but who contributed to large-scale sequencing of this team work. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013): MetaHIT, grant agreement HEALTH-F4-2007-201052, the Ole Rømer grant from the Danish Natural Science Research Council, the Solexa project (272-07-0196), the Shenzhen Municipal Government of China, the National Natural Science Foundation of China (30725008), the International Science and Technology Cooperation Project (0806), China (CXB200903110066A; ZYC200903240076A), the Danish Strategic Research Council grant no 2106-07-0021 (Seqnet), and the Lundbeck Foundation Centre for Applied Medical Genomics in Personalised Disease Prediction, Prevention and Care. Ciberehd is funded by Instituto de Salud Carlos III (Spain). We also thank X. Wang from the School of Biosciences and Bioengineering, South China University of Technology, for his coordination on the Innovative Program for Undergraduate Students in which J.L. and Y.X. joined.

Author Contributions All authors are members of the Metagenomics of the Human Intestinal Tract (MetaHIT) Consortium. So.L., H.Y., Je.W., J.D., F.G., K.K., O.P., S.B., J.P., Ji.W., S.D.E. and Ju.W. managed the project. T.N., T.H. and K.S.B. performed clinical analyses; F.L. and C.M. performed DNA extraction. X.Z., B.W., J.C., H.L., Hu.Z., K.T., D.L.P., E.P. and M.J. performed sequencing. Ju.W., S.D.E, P.B., R.L., J.R., M.A. and J.Q. designed the analyses. J.Q., Sha.L., D.L., J.L., J.X., Y.X., Ho.Z., M.B., H.B.N., T.S.-P., C.Y., She.L., T.Y., N.P., J.-M.B., P.L., D.R.M., S.D.E. and Y.Z. performed the data analyses. S.D.E., P.B., J.R., J.Q., R.L. and Ju.W. wrote the paper. J.T., A.L., P.R., Y.L. and N.Q. revised the paper. The MetaHIT Consortium members contributed to design and execution of the study.

Author Information The raw Illumina read data of all 124 samples has been deposited in the EBI, under the accession ERA000116. The contigs and gene set are available to download from the EMBL (http://www.bork.embl.de/∼arumugam/Qin_et_al_2010/) and BGI (http://gutmeta.genomics.org.cn) websites. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. This paper is distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share-Alike license, and is freely available to all readers at www.nature.com/nature. Correspondence and requests for materials should be addressed to Ju.W. (wangj@genomics.org.cn) or S.D.E. (dusko.ehrlich@jouy.inra.fr).

MetaHIT Consortium (additional members)

Maria Antolin[1], François Artiguenave[2], Hervé Blottiere[3], Natalia Borruel[1], Thomas Bruls[2], Francesc Casellas[1], Christian Chervaux[4], Antonella Cultrone[3], Christine Delorme[3], Gérard Denariaz[4], Rozenn Dervyn[3], Miguel Forte[5], Carsten Friss[6], Maarten van de Guchte[3], Eric Guedon[3], Florence Haimet[3], Alexandre Jamet[3], Catherine Juste[3], Ghalia Kaci[3], Michiel Kleerebezem[7], Jan Knol[4], Michel Kristensen[8], Severine Layec[3], Karine Le Roux[3], Marion Leclerc[3], Emmanuelle Maguin[3], Raquel Melo Minardi[2], Raish Oozeer[4], Maria Rescigno[9], Nicolas Sanchez[3], Sebastian Tims[7], Toni Torrejon[1], Encarna Varela[1], Willem de Vos[7], Yohanan Winogradsky[3] & Erwin Zoetendal[7]

[1]Hospital Universitari Val d'Hebron, Ciberehd, 08035 Barcelona, Spain. [2]Commissariat à l'Energie Atomique, Genoscope, 91000 Evry, France. [3]Institut National de la Recherche Agronomique, 78350 Jouy en Josas, France. [4]Danone Research, 91120 Palaiseau, France. [5]UCB Pharma SA, 28046 Madrid, Spain. [6]Center for Biological Sequence Analysis, Technical University of Denmark, DK-2800 Kongens Lyngby, Denmark. [7]Wageningen Unviersiteit, 6710BA Ede, The Netherlands. [8]Hagedorn Research Institute, DK 2820 Copenhagen, Denmark. [9]Istituto Europeo di Oncologia, 20100 Mila, Italy.

METHODS

Human faecal sample collection. Danish individuals were from the Inter-99 cohort[40], varying in phenotypes according to BMI and status towards obesity/diabetes, whereas Spanish individuals were either healthy controls or patients with chronic inflammatory bowel diseases (Crohn's disease or ulcerative colitis) in clinical remission.

Patients and healthy controls were asked to provide a frozen stool sample. Fresh stool samples were obtained at home, and samples were immediately frozen by storing them in their home freezer. Frozen samples were delivered to the Hospital using insulating polystyrene foam containers, and then they were stored at −80 °C until analysis.

DNA extraction. A frozen aliquot (200 mg) of each faecal sample was suspended in 250 µl of guanidine thiocyanate, 0.1 M Tris (pH 7.5) and 40 µl of 10% N-lauroyl sarcosine. Then, DNA extraction was conducted as previously described[22]. The DNA concentration and its molecular size were estimated by nanodrop (Thermo Scientific) and agarose gel electrophoresis.

DNA library construction and sequencing. DNA library preparation followed the manufacturer's instruction (Illumina). We used the same workflow as described elsewhere to perform cluster generation, template hybridization, isothermal amplification, linearization, blocking and denaturization and hybridization of the sequencing primers. The base-calling pipeline (version IlluminaPipeline-0.3) was used to process the raw fluorescent images and call sequences.

We constructed one library (clone insert size 200 bp) for each of the first 15 samples, and two libraries with different clone insert sizes (135 bp and 400 bp) for each of the remaining 109 samples for validation of experimental reproducibility.

To estimate the optimal return between the generation of novel sequence and sequencing depth, we aligned the Illumina GA reads from samples MH0006 and MH0012 onto 468,335 Sanger reads totalling to 311.7 Mb generated from the same two samples (156.9 and 154.7 Mb, respectively, Supplementary Table 2), using the Short Oligonucleotide Alignment Program (SOAP)[41] and a match requirement of 95% sequence identity. With about 4 Gb of Illumina sequence, 94% and 89% of the Sanger reads (for MH0006 and MH0012, respectively) were covered. Further extensive sequencing, to 12.6 and 16.6 Gb for MH0006 and MH0012, respectively, brought only a moderate increase of coverage to about 95% (Supplementary Fig. 1). More than 90% of the Sanger reads were covered by the Illumina sequences to a very high and uniform level (Supplementary Fig. 2), indicating that there is little or no bias in the Illumina GA sequence. As expected, a large proportion of Illumina sequences (57% and 74% for M0006 and M0012, respectively) was novel and could not be mapped onto the Sanger reads. This fraction was similar at the 4 and 12–16 Gb sequencing levels, confirming that most of the novelty was captured already at 4 Gb.

We generated 35.4–97.6 million reads for the remaining 122 samples, with an average of 62.5 million reads. Sequencing read length of the first batch of 15 samples was 44 bp and the second batch was 75 bp.

Public data used. The sequenced bacteria genomes (totally 806 genomes) deposited in GenBank were downloaded from NCBI database (http://www.ncbi.nlm.nih.gov/) on 10 January 2009. The known human gut bacteria genome sequences were downloaded from HMP database (http://www.hmpdacc-resources.org/cgi-bin/hmp_catalog/main.cgi), GenBank (67 genomes), Washington University in St Louis (85 genomes, version April 2009, http://genome.wustl.edu/pub/organism/Microbes/Human_Gut_Microbiome/), and sequenced by the MetaHIT project (17 genomes, version September 2009, http://www.sanger.ac.uk/pathogens/metahit/). The other gut metagenome data used in this project include: (1) human gut metagenomic data sequenced from US individuals[8], which was downloaded from NCBI with the accession SRA002775; (2) human gut metagenomic data from Japanese individuals[17], which was downloaded from P. Bork's group at EMBL (http://www.bork.embl.de). The integrated NR database we constructed in this study included NCBI-NR database (version April 2009) and all genes from the known human gut bacteria genomes.

Illumina GA short reads *de novo* assembly. High-quality short reads of each DNA sample were assembled by the SOAPdenovo assembler[19]. In brief, we first filtered the low abundant sequences from the assembly according to 17-mer frequencies. The 17-mers with depth less than 5 were screened in front of assembly, for these low-frequency sequences were very unlikely to be assembled, whereas removing them would significantly reduce memory requirement and make assembly feasible in an ordinary supercomputer (512 GB memory in our institute).

Then the sequences were processed one by one and the de Bruijn graph data format was used to store the overlap information among the sequences. The overlap paths supported by a single read were unreliable and removed. Short low-depth tips and bubbles that were caused by sequencing errors or genetic variations between microbial strains were trimmed and merged, respectively. Read paths were used to solve the tiny repeats.

Finally, we broke the connections at repeat boundaries, and outputted the continuous sequences with unambiguous connections as contigs. The metagenomic special model was chosen, and parameters '−K 21' and '−K 23' were used for 44 bp and 75 bp reads, respectively, to indicate the minimal sequence overlap required.

After *de novo* assembly for each sample independently, we merged all the unassembled reads together and performed assembly for them, as to maximize the usage of data and assemble the microbial genomes that have low frequency in each read set, but have sufficient sequence depth for assembly by putting the data of all samples together.

Validating Illumina contigs using Sanger reads. We used BLASTN (WU-BLAST 2.0) to map Sanger reads from samples MH0006 and MH0012 (156.9 Mb and 154.7 Mb, respectively) to Illumina contigs (single best hit longer than 75 bp and over 95% identity) from the same samples. Each alignment was scanned for breakage of collinearity where both sequences have at least 50 bases left unaligned at one end of the alignment. Each such breakage was considered an assembly error in the Illumina contig at the location where collinearity breaks. Errors within 30 bp from each other were merged. An error was discarded if there exists a Sanger read that agrees with the contig structure for 60 bp on both sides of the error. For comparison, we repeated this on a Newbler2 assembly of 454 Titanium reads from MH0006 (550 Mb reads). Supplementary Fig. 5a shows the number of errors per Mb of assembled Illumina/454 contigs. We estimate 14.12 errors per Mb of contigs for the Illumina assembly, which is comparable to that of the 454 assembly (20.73 per Mb). 98.7% of Illumina contigs that map at least one Sanger read were collinear over 99.55% of the mapped regions, which is comparable to 97.86% of such 454 contigs being collinear over 99.48% of the mapped regions.

Evaluation of human gut microbiome coverage. The Illumina GA reads were aligned against the assembled contigs and known bacteria genomes using SOAP[41] by allowing at most two mismatches in the first 35-bp region and 90% identity over the read sequence. The Roche/454 and Sanger sequencing reads were aligned against the same reference using BLASTN with 1×10^{-8}, over 100 bp alignment length and minimal 90% identity cutoff. Two mismatches were allowed and identity was set 95% over the read sequence when aligned to the GA reads of MH0006 and MH0012 to Sanger reads from the same samples by SOAP.

Gene prediction and construction of the non-redundant gene set. We use MetaGene[20]—which uses di-codon frequencies estimated by the GC content of a given sequence, and predicts a whole range of ORFs based on the anonymous genomic sequences—to find ORFs from the contigs of each of the 124 samples as well as the contigs from the merged assembly.

The predicted ORFs were then aligned to each other using BLAT[36]. A pair of genes with greater than 95% identity and aligned length covered over 90% of the shorter gene was grouped together. The groups sharing genes were then merged, and the longest ORF in each merged group was used to represent the group, and the other members of the group were taken as redundancy. Therefore, we organized the non-redundant gene set from all the predicted genes by excluding the redundancy. Finally, the ORFs with length less than 100 bp were filtered. We translated the ORFs into protein sequences using the NCBI Genetic Codes11.

Identification of genes. To make a balance between identifying low-abundance genes and reducing the error-rate of identification, we explored the impact of the threshold set for read coverage required to identify a gene in individual microbiomes. The number of genes decreased about twice when the number of reads required for identification was increased from 2 to 6, and changed slowly thereafter (Supplementary Fig. 6a). Nevertheless, to include the rare genes into the analysis, we selected the threshold of 2 reads.

Gene taxonomic assignment. Taxonomic assignment of predicted genes was carried out using BLASTP alignment against the integrated NR database. BLASTP alignment hits with *e*-values larger than 1×10^{-5} were filtered, and for each gene the significant matches which were defined by *e*-values $\leq 10 \times$ *e*-value of the top hit were retained to distinguish taxonomic groups. Then we determined the taxonomical level of each gene by the lowest common ancestor (LCA)-based algorithm that was implemented in MEGAN[42]. The LCA-based algorithm assigns genes to taxa in the way that the taxonomical level of the assigned taxon reflects the level of conservation of the gene. For example, if a gene was conserved in many species, it was assigned to the LCA rather than to a species.

Gene functional classification. We used BLASTP to search the protein sequences of the predicted genes in the eggNOG database[26] and KEGG database[24] with *e*-value $\leq 1 \times 10^{-5}$. The genes were annotated as the function of the NOGs or KEGG homologues with lowest *e*-value. The eggNOG database is an integration of the COG and KOG databases. The genes annotated by COG were classified into the 25 COG categories, and genes that were annotated by KEGG were assigned into KEGG pathways.

Determination of minimal gut bacterial genome. The number of non-redundant genes assigned to the eggNOG clusters was normalized by gene length and cluster copy number (Supplementary Fig. 8). The clusters were ranked by normalized gene number and the range that included the clusters encoding essential *Bacillus subtilis* genes was determined, computing the proportion of these clusters among the successive groups of 100 clusters. Analysis of the range gene clusters involved, besides iPath projections, use of KEGG and manual verification of the completeness of the pathways and protein machineries they encode.

Determination of total functional complement and minimal metagenome. We computed the total and shared number of orthologous groups and/or gene families present in random combinations of n individuals (with $n = 2$ to 124, 100 replicates per bin). This analysis was performed on three groups of gene clusters: (1) known eggNOG orthologous groups (that is, those with functional annotation, excluding those in which the terms [Uu]ncharacteri[sz]ed, [Uu]nknown, [Pp]redicted or[Pp]utative occurred); (2) all eggNOG orthologous groups; (3) all orthologous groups plus gene families constructed from remaining genes not assigned to the two above categories. Families were clustered from all-against-all BLASTP results using MCL[43] with an inflation factor of 1.1 and a bit-score cutoff of 60.

Rarefaction analysis. Estimation of total gene richness was done using EstimateS on 100 randomly picked samples due to memory limitations. Because the CV value was >0.5, both chao2 (classic) and ICE richness estimators were calculated and the larger estimate of the two (ICE) was used. The estimate for this sample size was 3,621,646 genes (ICE) whereas S_{obs} (Mao Tau) was 3,090,575 genes, or 85.3%. The ICE estimator curve did not completely saturate, (data not shown) indicating that additional samples will need to be added to achieve a final, conclusive estimate.

Common bacterial core. To eliminate the influence of very similar strains and assess the presence of known microbial species among the individuals of the cohort, we used 650 sequenced bacterial and archaeal genomes as a reference set. The set was composed from 932 publicly available genomes, which were grouped by similarity, using a 90% identity cutoff and the similarity over at least 80% of the length. From each group only the largest genome was used. Illumina reads from 124 individuals were mapped to the set, for species profiling analysis and the genomes originating from the same species (by differing in size >20%) curated by manual inspection and by using the 16S-based clustering when the sequences were available.

Relative abundance of microbial genomes among individuals. We computed the genome coverage by uniquely mapping Illumina reads and normalized it to 1 Gb of sequence, to correct for different sequencing levels in different individuals. The coverage was summed over all species of the non-redundant bacterial genome set for each individual and the proportion of each species relative to the sum calculated.

Species co-existence network. For the 155 species that had genome coverage by the Illumina reads ≥1% in at least one individual we calculated the pair-wise inter-species Pearson correlations between sequencing depths (abundance) throughout the entire cohort of 124 individuals. From the resulting 11,175 inter-species correlations, correlations less than −0.4 or above 0.4 ($n = 342$) were visualized in a graph using Cytoscape[44] displaying the average genome coverage of each species as node size in the graph.

40. Toft, U. *et al.* The impact of a population-based multi-factorial lifestyle intervention on changes in long-term dietary habits: The Inter99 study. *Prev. Med.* **47,** 378–383 (2008).

41. Li, R. *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25,** 1966–1967 (2009).

42. Huson, D. H., Auch, A. F., Qi, J. & Schuster, S. C. MEGAN analysis of metagenomic data. *Genome Res.* **17,** 377–386 (2007).

43. van Dongen, S. *Graph Clustering by Flow Simulation*. PhD thesis, Univ. Utrecht (2000).

44. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13,** 2498–2504 (2003).