Published in final edited form as:
Nature. 2010 March 4; 464(7285): 59–65. doi:10.1038/nature08821.

A human gut microbial gene catalog 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,†], and **Jun Wang**[1,13,†]

[1]BGI-Shenzhen, Shenzhen 518083, China.

[2]European Molecular Biology Laboratory, 69118 Heidelberg, Germany.

[3]address: 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.

[†]Correspondence and requests for materials should be addressed to Ju. W. (wangj@genomics.org.cn) and S. D. E. (dusko.ehrlich@jouy.inra.fr). .
[*]These authors contributed equally to this work.
[§]The list of additional MetaHIT Consortium members is appended at the end of the article.

Author Contributions So. L., H. Y., Je. W. J. D., F. G., K. K., O. P., S.B., J. P., J. 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., H. Z., K. T., D. 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., S. L., D. L., J. L., J. X., Y. X., H. Z., M. B., H. B. N, T. S. P, C. Y., S. L., T. Y., N. P., J. M. B., D. 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.

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

Additional MetaHIT Consortium members Maria Antolin[5], François Artiguenave[10], Hervé Blottiere[6], Natalia Borruel[5], Thomas Bruls[10], Francesc Casellas[5], Christian Chervaux[15], Antonella Cultrone[6], Christine Delorme[6], Gérard Denariaz[15], Rozenn Dervyn[6], Miguel Forte[16], Carsten Friss[9], Maarten van de Guchte[6], Eric Guedon[6], Florence Haimet[6], Alexandre Jamet[6], Catherine Juste[6], Ghalia Kaci[6], Michiel Kleerebezem[17], Jan KNOL[15], Michel Kristensen[4], Severine Layec[6], Karine Le Roux[6], Marion Leclerc[6], Emmanuelle Maguin[6], Raquel Melo Minardi[10], Raish Oozeer[15], Maria Rescigno[18], Nicolas Sanchez[6], Sebastian Tims[17], Toni Torrejon[5], Encarna Varela[5], Willem de Vos[17], Yohanan Winogradsky[6], Erwin Zoetendal[17].
[15] Danone Research, Palaiseau, France.
[16] UCB Pharma SA, Madrid, Spain.
[17] Wageningen Unviersiteit, The Netherlands.
[18] Istituto Europeo di Oncologia, Mila, Italy.

The raw Illumina reads 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_2009/) and BGI (http://gutmeta.genomics.org.cn) websites.

The authors declare no competing financial interests.

Europe PMC Funders Author Manuscripts

[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.

Abstract

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 nonredundant microbial genes, derived from 576.7 Gb 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 microbial genes of the cohort and likely 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, suggesting that the entire cohort harbours between 1000 and 1150 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 encoded by the gene set.

Introduction

It has been estimated that the microbes in our bodies collectively make up to 100 trillion cells, ten-fold the number of human cells, and suggested that they encode 100-fold more unique genes than our own genome[1]. The majority of microbes resides 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 dramatic among infants[12], later in life the gut microbiome converges to more similar phyla.

Metagenomic sequencing represents a powerful alternative to rRNA sequencing for analyzing complex microbial communities[13-15]. Applied to the human gut, such studies have already generated some 3 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 Analyzer (GA) technology to carry out deep

Europe PMC Funders Author Manuscripts

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 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 characterisation 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 individuals human adults, as well as the inflammatory disease 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 to capture most of the novelty (Supplementary 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 the Illumina GA sequence data. Since a high diversity between individuals is expected[8,16,17], we first assembled each sample independently (Supplementary Fig. 3). As many as 42.7% of the Illumina GA reads were assembled into a total of 6.58 million contigs of a length >500 bp, giving a total contig length of 10.3 Gb, with 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). 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 454 contigs that were also generated 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 Mb of Illumina- and 454-based contigs, respectively, (see Supplementary 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 370Mb 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 Illumina GA sequences could be aligned to the contigs at a threshold of 90% identity, allowing to accommodate sequencing errors and strain variability in the gut (Fig. 1), almost twice the 42.7% sequences that were assembled into contigs by SOAPdenovo, because assembly uses more stringent criteria. This indicates that a vast majority of the Illumina sequences are 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 Roche/454 sequencing reads from 18 US adults[8], and 0.79 Gb Sanger reads from 13 Japanese adults and infants[17]), using the 90% identity threshold. 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.

Only 31.0-48.8% of the reads from the two previous and the present studies 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 orthologs but avoids inflation of the dataset due to possible sequencing errors (see Supplementary 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 Supplementary Methods). The minimal relative abundance of the prevalent genes was $\sim 6\times10^{-7}$, 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, $3\times0.85\times10^{3}/4.5\times10^{9}$).

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 2 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 likely is 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 (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%, while 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 (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 the cohort. For this purpose, we used a nonredundant set of 650 sequenced bacterial and archaeal genomes (see Supplementary 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 suggests 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 visualise 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 2187-fold (Fig. 3). As expected[10,23], Bacteroidetes and Firmicutes had the highest abundance.

A complex pattern of species relatedness, characterised by clusters at the genus and family levels, emerges from the analysis of covariance 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 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 suggests 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 UC from the CD patients (Fig. 4), confirming our hypothesis.

Functions encoded by the prevalent gene set

We classified the predicted genes by aligning them to the integrated NR database, the genes in the KEGG (Kyoto Encyclopedia of Genes and Genomes)[24] pathways, 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). Almost all (99.96%) of the phylogenetically assigned genes belonged to bacteria and archaea, reflecting their predominance in the gut. Genes that were not mapped to orthologous groups were clustered into gene families (see Supplementary 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 to 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 functionalities.

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 (i.e. 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 normalisation for gene length and copy number (Supplementary Fig. 10 a,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], hypothesising 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 ("house-keeping") and those potentially specific for the gut. Among many examples of the first category are the functions that are part of main metabolic pathways (eg. 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 decorate 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. 80% 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 e.g. mannose, fructose, cellulose and sucrose is also part of the minimal metagenome. Together, these emphasize the strong dependence of the gut ecosystem on complex sugar degradation for its functioning.

Functional complementarities of the gut metagenome and human genome

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 (SCFA), 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 while, in addition, butyrate is important for enterocytes[32]. Beyond SCFA, a number of amino-acids are indispensable to humans[33] and can be provided by bacteria[34]. Similarly, bacteria can contribute certain vitamins[3] (e.g. biotin, phylloquinone) to the host. All of the steps of biosynthesis of these molecules are encoded by the minimal metagenome.

Gut bacteria appear 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 (Co-A) ligation pathway, encoded in the minimal metagenome, leads to pameloyl-Co-A, which is a precursor of biotin, suggesting 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 nonredundant 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 harbored by our cohort. It is likely that the catalogue contains a large majority of prevalent intestinal microbial genes in the human population, as: (i) over 70% of the metagenomic reads from three previous studies, including American and Japanese individuals[8,16,17], can be mapped on our contigs; (ii) 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 catalog 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 catalog 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 therefore the cohort harbors 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 harbors 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 further explore 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 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 inter-personal diversity of innumerable bacterial species in the gut, arising from most previous studies using the 16 S 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 indeed a very high variability of abundance (12- to 2200-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 likely 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 catalog 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. MetaGgene 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 < 1e-5) of the NCBI-nr database and 126 known gut bacteria genomes. Gene functional annotations were made by BLASTP search (e-value < 1e-5) with eggNOG and KEGG (v48.2) databases. The total and shared number of orthologous groups and/or gene families were computed using random combination of n individuals (with n=2 to 124, 100 replicates per bin).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

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), and 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).

References

1. Ley RE, Peterson DA, Gordon JI. Ecological and evolutionary forces shaping microbial diversity in the human intestine. Cell. 2006; 124:837–848. doi:S0092-8674(06)00192-9 [pii]10.1016/j.cell.2006.02.017. [PubMed: 16497592]

2. Backhed F, Ley RE, Sonnenburg JL, Peterson DA, Gordon JI. Host-bacterial mutualism in the human intestine. Science. 2005; 307:1915–1920. doi:307/5717/1915 [pii]10.1126/science.1104816. [PubMed: 15790844]

3. Hooper LV, Midtvedt T, Gordon JI. How host-microbial interactions shape the nutrient environment of the mammalian intestine. Annu Rev Nutr. 2002; 22:283–307. doi:10.1146/annurev.nutr.22.011602.092259011602.092259 [pii]. [PubMed: 12055347]

4. Ley RE, Turnbaugh PJ, Klein S, Gordon JI. Microbial ecology: human gut microbes associated with obesity. Nature. 2006; 444:1022–1023. doi:4441022a [pii]10.1038/4441022a. [PubMed: 17183309]

5. Turnbaugh PJ, et al. An obesity-associated gut microbiome with increased capacity for energy harvest. Nature. 2006; 444:1027–1031. doi:nature05414 [pii]10.1038/nature05414. [PubMed: 17183312]

6. Ley RE, et al. Obesity alters gut microbial ecology. Proc Natl Acad Sci U S A. 2005; 102:11070–11075. doi:0504978102 [pii]10.1073/pnas.0504978102. [PubMed: 16033867]

7. Zhang H, et al. Human gut microbiota in obesity and after gastric bypass. Proc Natl Acad Sci U S A. 2009; 106:2365–2370. doi:0812600106 [pii]10.1073/pnas.0812600106. [PubMed: 19164560]

8. Turnbaugh PJ, et al. A core gut microbiome in obese and lean twins. Nature. 2009; 457:480–484. doi:nature07540 [pii]10.1038/nature07540. [PubMed: 19043404]

9. Zoetendal EG, Akkermans AD, De Vos WM. Temperature gradient gel electrophoresis analysis of 16S rRNA from human fecal samples reveals stable and host-specific communities of active bacteria. Appl Environ Microbiol. 1998; 64:3854–3859. [PubMed: 9758810]

10. Eckburg PB, et al. Diversity of the human intestinal microbial flora. Science. 2005; 308:1635–1638. doi:1110591 [pii]10.1126/science.1110591. [PubMed: 15831718]

11. Ley RE, Lozupone CA, Hamady M, Knight R, Gordon JI. Worlds within worlds: evolution of the vertebrate gut microbiota. Nat Rev Microbiol. 2008; 6:776–788. doi:nrmicro1978 [pii]10.1038/nrmicro1978. [PubMed: 18794915]

12. Palmer C, Bik EM, Digiulio DB, Relman DA, Brown PO. Development of the Human Infant Intestinal Microbiota. PLoS Biol. 2007; 5:e177. doi:07-PLBI-RA-0129 [pii]10.1371/journal.pbio.0050177. [PubMed: 17594176]

13. Riesenfeld CS, Schloss PD, Handelsman J. Metagenomics: genomic analysis of microbial communities. Annu Rev Genet. 2004; 38:525–552. doi:10.1146/annurev.genet.38.072902.091216. [PubMed: 15568985]

14. von Mering C, et al. Quantitative phylogenetic assessment of microbial communities in diverse environments. Science. 2007; 315:1126–1130. doi:1133420 [pii]10.1126/science.1133420. [PubMed: 17272687]

15. Tringe SG, Rubin EM. Metagenomics: DNA sequencing of environmental samples. Nat Rev Genet. 2005; 6:805–814. doi:nrg1709 [pii]10.1038/nrg1709. [PubMed: 16304596]

16. Gill SR, et al. Metagenomic analysis of the human distal gut microbiome. Science. 2006; 312:1355–1359. doi:312/5778/1355 [pii]10.1126/science.1124234. [PubMed: 16741115]

17. Kurokawa K, et al. Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. DNA Res. 2007; 14:169–181. doi:dsm018 [pii]10.1093/dnares/dsm018. [PubMed: 17916580]

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. 1999; 65:4799–4807. [PubMed: 10543789]

19. Li R, Zhu H. De novo assembly of the human genomes with massively parallel short read sequencing. Genome Res. 2009

20. Noguchi H, Park J, Takagi T. MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. Nucleic Acids Res. 2006; 34:5623–5630. doi:gkl723 [pii]10.1093/nar/gkl723. [PubMed: 17028096]

21. Colwell, RK. EstimateS: Statistical estimation of species richness and shared species from samples. 2005.

22. Manichanh C, et al. Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. Gut. 2006; 55:205–211. doi:gut.2005.073817 [pii]10.1136/gut.2005.073817. [PubMed: 16188921]

23. Wang X, Heazlewood SP, Krause DO, Florin TH. Molecular characterization of the microbial species that colonize human ileal and colonic mucosa by using 16S rDNA sequence analysis. J Appl Microbiol. 2003; 95:508–520. doi:2005 [pii]. [PubMed: 12911699]
24. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Res. 2004; 32:D277–280. doi:10.1093/nar/gkh06332/suppl_1/D277 [pii]. [PubMed: 14681412]
25. Tatusov RL, et al. The COG database: an updated version includes eukaryotes. BMC Bioinformatics. 2003; 4:41. doi:10.1186/1471-2105-4-411471-2105-4-41 [pii]. [PubMed: 12969510]
26. Jensen LJ, et al. eggNOG: automated construction and annotation of orthologous groups of genes. Nucleic Acids Res. 2008; 36:D250–254. doi:gkm796 [pii]10.1093/nar/gkm796. [PubMed: 17942413]
27. Kobayashi K, et al. Essential Bacillus subtilis genes. Proc Natl Acad Sci U S A. 2003; 100:4678–4683. doi:10.1073/pnas.0730515100730515100 [pii]. [PubMed: 12682299]
28. Baba T, et al. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol. 2006; 2 2006 0008. doi:msb4100050 [pii]10.1038/msb4100050.
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. 2000; 66:1321–1327. [PubMed: 10742206]
30. Cummings JH, Macfarlane GT. The control and consequences of bacterial fermentation in the human colon. J Appl Bacteriol. 1991; 70:443–459. [PubMed: 1938669]
31. Wong JM, de Souza R, Kendall CW, Emam A, Jenkins DJ. Colonic health: fermentation and short chain fatty acids. J Clin Gastroenterol. 2006; 40:235–243. doi:00004836-200603000-00015 [pii]. [PubMed: 16633129]
32. Hamer HM, et al. Review article: the role of butyrate on colonic function. Aliment Pharmacol Ther. 2008; 27:104–119. doi:APT3562 [pii]10.1111/j.1365-2036.2007.03562.x. [PubMed: 17973645]
33. Elango R, Ball RO, Pencharz PB. Amino acid requirements in humans: with a special emphasis on the metabolic availability of amino acids. Amino Acids. 2009; 37:19–27. doi:10.1007/s00726-009-0234-y. [PubMed: 19156481]
34. Metges CC. Contribution of microbial amino acids to amino acid homeostasis of the host. J Nutr. 2000; 130:1857S–1864S. [PubMed: 10867063]
35. Tap J, et al. Towards the human intestinal microbiota phylogenetic core. Environ Microbiol. 2009; 11:2574–2584. doi:EMI1982 [pii]10.1111/j.1462-2920.2009.01982.x. [PubMed: 19601958]
36. Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res. 2002; 12:656–664. doi:10.1101/gr.229202. Article published online before March 2002. [PubMed: 11932250]
37. Altschul SF, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25:3389–3402. doi:gka562 [pii]. [PubMed: 9254694]
38. Letunic I, Yamada T, Kanehisa M, Bork P. iPath: interactive exploration of biochemical pathways and networks. Trends Biochem Sci. 2008; 33:101–103. doi:S0968-0004(08)00023-6 [pii]10.1016/j.tibs.2008.01.001. [PubMed: 18276143]
39. von Mering C, et al. STRING 7--recent developments in the integration and prediction of protein interactions. Nucleic Acids Res. 2007; 35:D358–362. doi:gkl825 [pii]10.1093/nar/gkl825. [PubMed: 17098935]

Europe PMC Funders Author Manuscripts

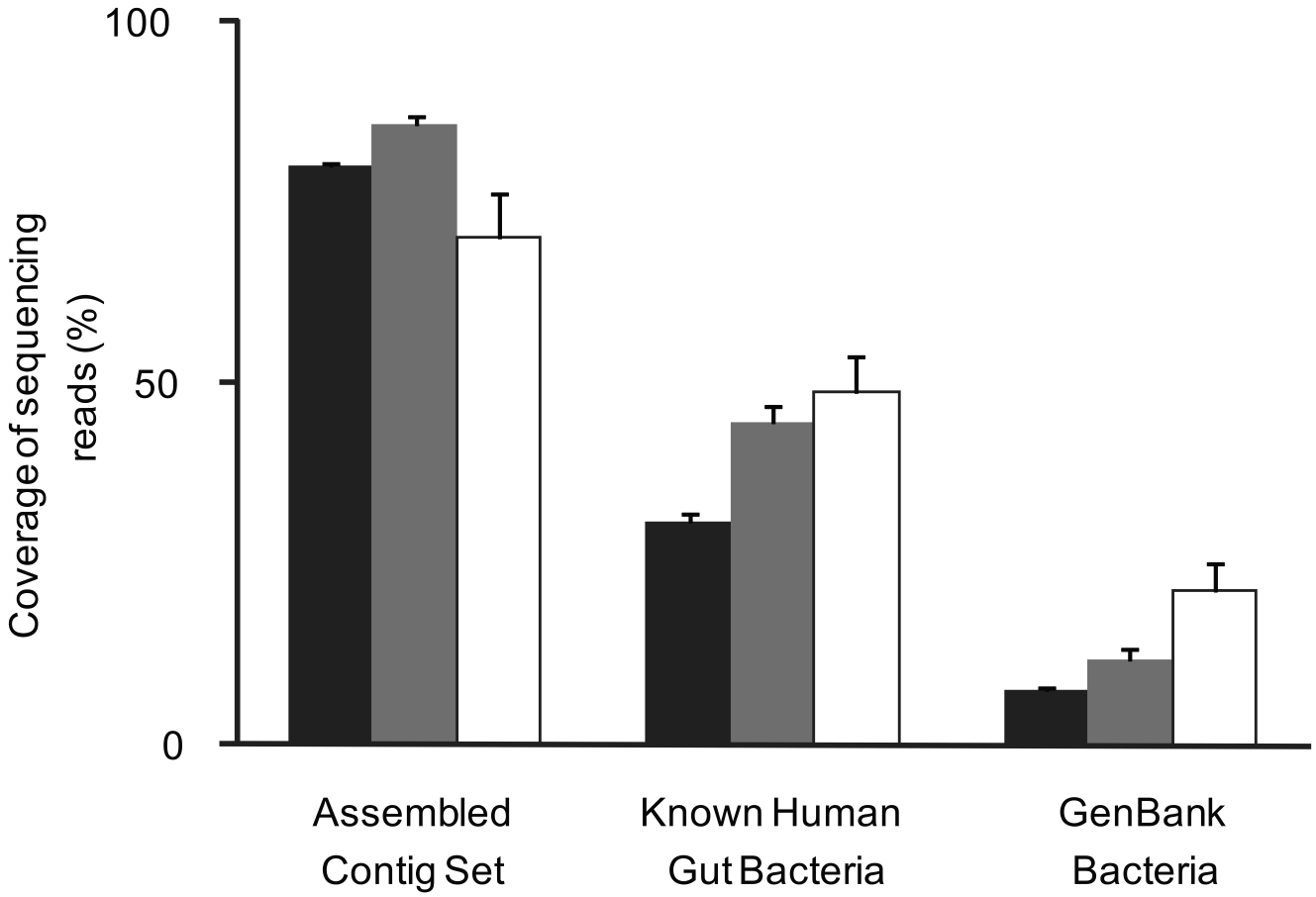

Figure 1. Coverage of human gut microbiome

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.

Europe PMC Funders Author Manuscripts

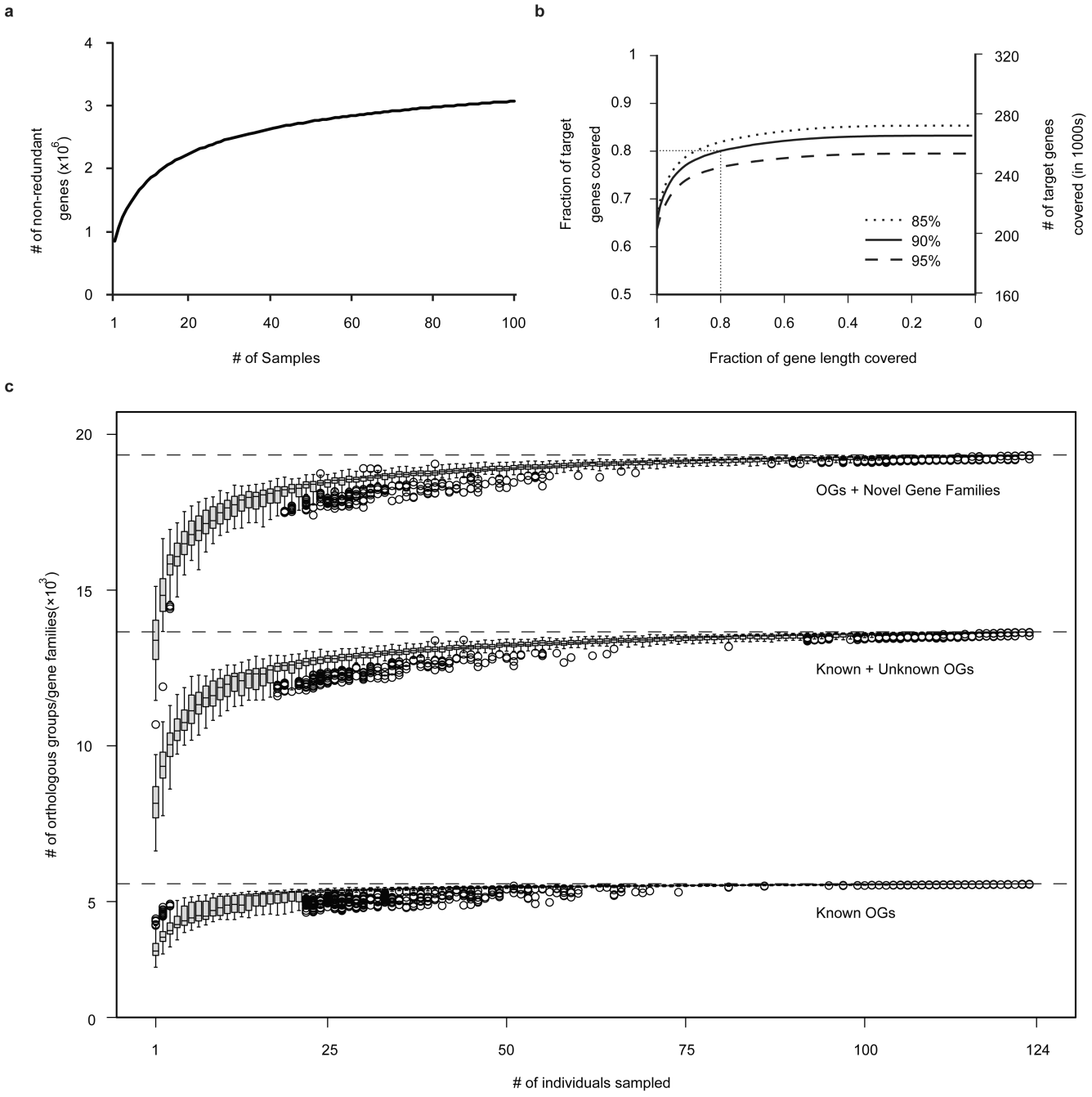

Figure 2. Predicted ORFs in the human gut microbiomes

a, Number of unique genes as function of the extent of sequencing. The gene accumulation curve corresponds to the Sobs (Mao Tau) values, 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 upon known (well characterized) orthologous groups (OGs; bottom), known+unknown orthologous groups (including e.g. putative, predicted, conserved hypothetical functions; center) and OGs+novel gene families (>20 proteins) recovered from the metagenome (top).

Europe PMC Funders Author Manuscripts

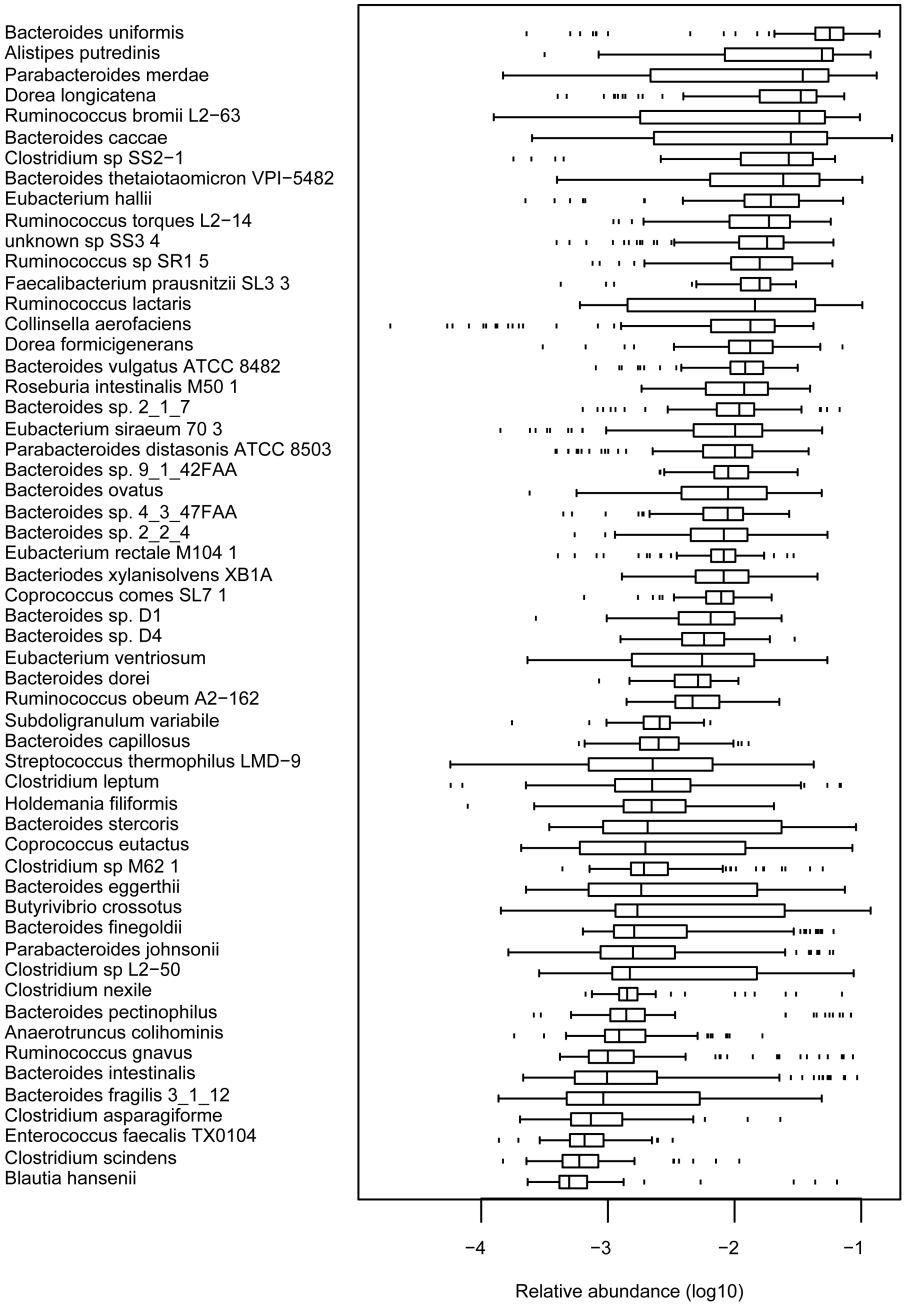

Figure 3. Relative abundance of frequent microbial genomes among individuals of the cohort

Boxes denote 25% and 75% percentiles, the black line in the box corresponds to the median, the “whiskers” indicate the interquartile range from either or both ends of the box, the dots show the outliers, beyond the ends of the whiskers (See supplementary Methods for computation).

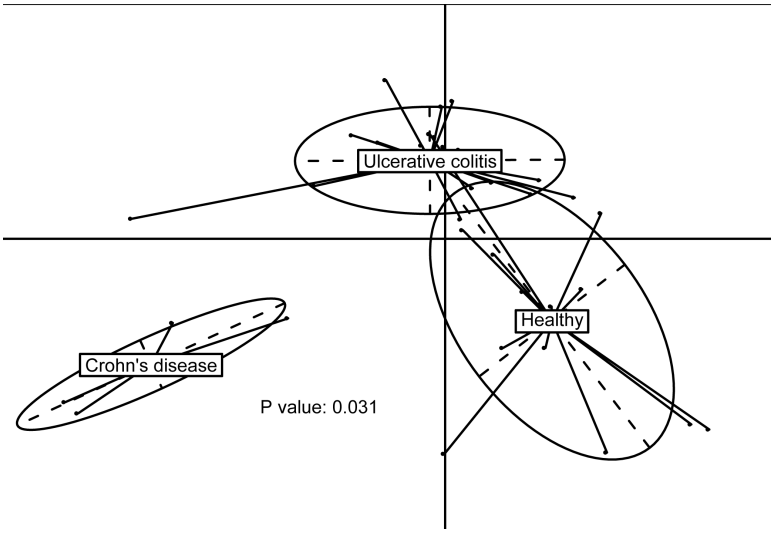

Figure 4. Bacterial species abundance differentiates IBD patients and healthy individuals

Principal component analysis 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 from Spain.

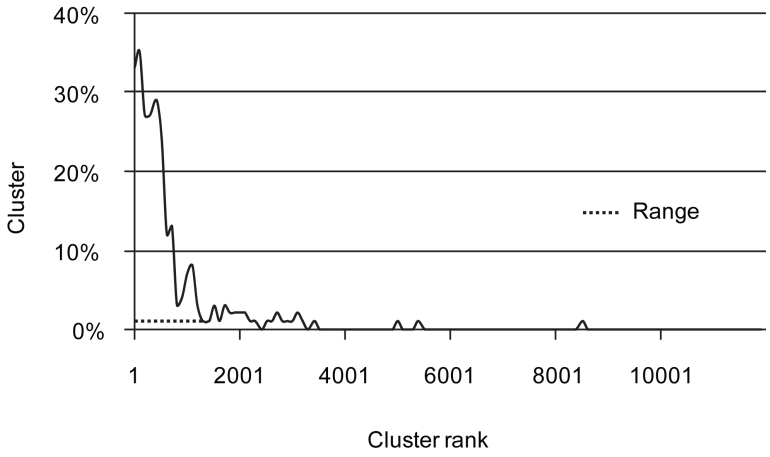

Figure 5. Clusters that contain the *B. subtilis* essential genes

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.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

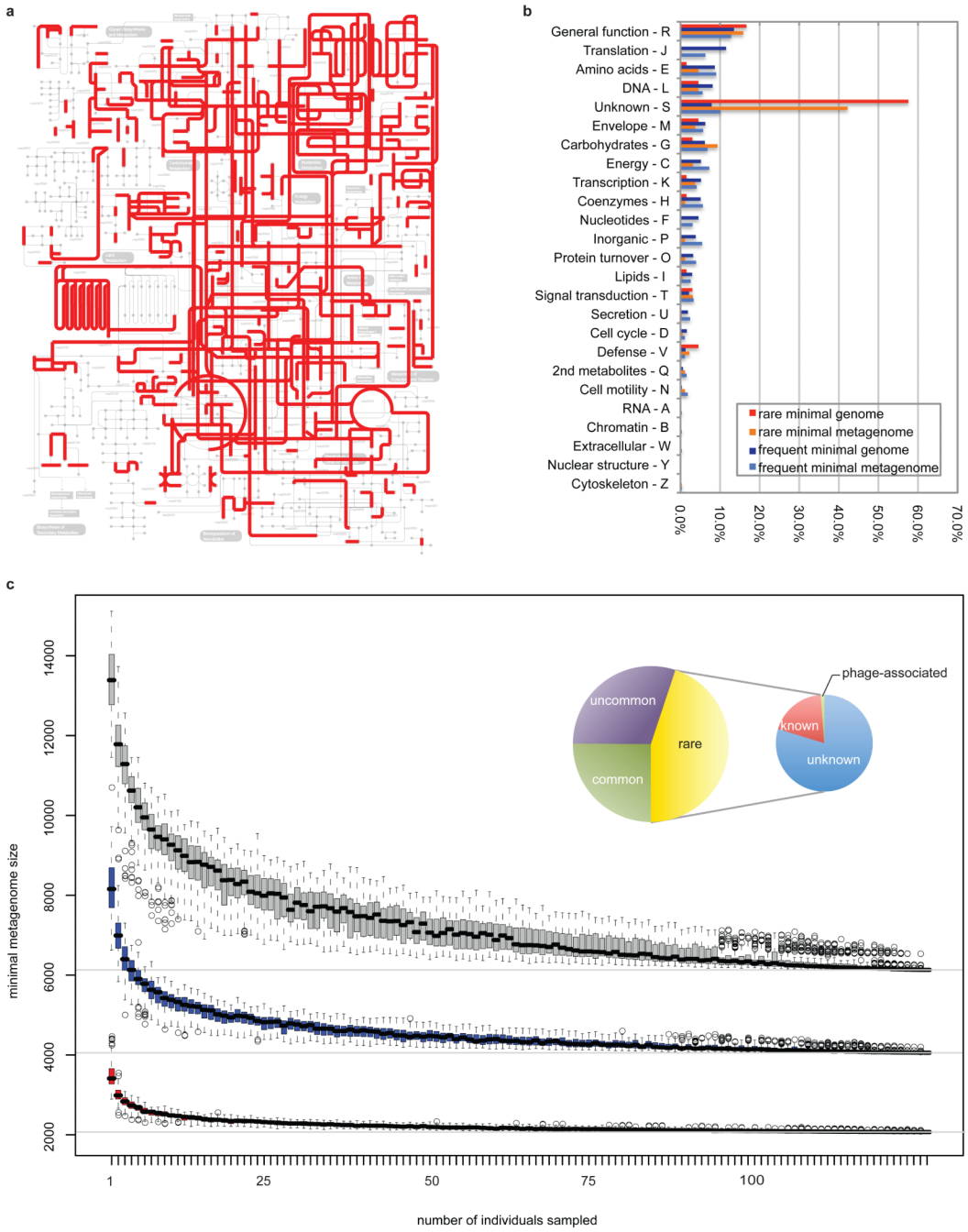

Figure 6. Characterization of the minimal gut genome and metagenome

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*. **c,** Estimation of the minimal gut metagenome size. Known orthologous groups (OGs; red), known+unknown OGs (blue) and OGs+novel gene families (>20 proteins; grey). Inset: Composition of the gut minimal microbiome. Large circle: Classification in the minimal metagenome according to OG occurrence in STRING7[39] bacterial genomes. Common (25%), uncommon (35%) and rare (45%) are present in >50%, <50% but >10% and <10% of genomes, respectively. Small circle: composition of the rare OGs. Unknown (80%) have no annotation or are poorly

characterized, while known bacterial (19%) and phage-related (1%) OGs have functional description.

Europe PMC Funders Author Manuscripts