Supplementary Methods

Human faecal samples collection. Danish individuals were from the Inter-99 cohort[40], varying in phenotypes according to BMI and status towards obesity/diabetes while 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 -80ºC 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 Inc.). We used the same workflow as described elsewhere to perform cluster generation, template hybridization, isothermal amplification, linearization, blocking, and denaturisation 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 311.7Mb generated from the same two samples (156.9 and 154.7 Mb respectively, Supplementary Table 2), using the SOAP[41] program 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 January 10th 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/HumanGutMicrobiome/), 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 includes: 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 Bork's group at EMBL (http://www.bork.embl.de). The integrated NR database we constructed in this study included GenBank 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 firstly 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-frequent sequences were very unlikely to be assembled, while 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 loaded into memory 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' was used for 44 bp and 75 bp read, 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 (156.9Mb and 154.7Mb respectively) to Illumina contigs (single best hit longer than 75bp 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 30bp from each other were merged. An error was discarded if there exists a Sanger read that agrees with the contig structure for 60bp on both sides of the error. For comparison, we repeated this on a Newbler2 assembly of 454 Titanium reads from MH0006 (550 Mb reads). Fig. 4a 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 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] (Short Oligonucleotide Alignment Program) 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 1E-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 same samples by SOAP.

Gene prediction and construction of the non-redundant gene set. We use MetaGene[20], which utilize 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 gene 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 1E-5 were filtered, and for each gene the significant matches which were defined by e-values $\leq$ 10 × e-value of the top hit, were retained to distinguish taxonomic groups. Then we determined the taxonomical level of each gene by the 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 lowest common ancestor (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 $\leq$ 1E-05. 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 which 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 normalised 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 (OGs; i.e. 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 OGs; (3) all OGs + gene families constructed from remaining genes not assigned to the two above categories. Families were clustered from all-against-all BLASTP results using MCL (van Dongen, Graph Clustering by Flow Simulation. PhD thesis, University of Utrecht, May 2000) with an inflation factor of 1.1 and a bitscore 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) while 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 cut-off 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 of the cohort. 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 nonredundant 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 pairwise 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[43] displaying the average genome coverage of each species as node size in the graph.

Supplementary Methods References

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*, doi:S0091-7435(08)00273-9 [pii]10.1016/j.ypmed.2008.05.013 (2008).

41 Li, R. *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, doi:btp336 [pii]10.1093/bioinformatics/btp336 (2009).

42 Huson, D. H., Auch, A. F., Qi, J. & Schuster, S. C. MEGAN analysis of metagenomic data. *Genome Res* **17**, 377-386, doi:gr.5969107 [pii]10.1101/gr.5969107 (2007).

43 Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-2504, doi:10.1101/gr.123930313/11/2498 [pii] (2003).