ARTICLE

doi:10.1038/nature12506

Richness of human gut microbiome correlates with metabolic markers

Emmanuelle Le Chatelier[1]*, Trine Nielsen[2]*, Junjie Qin[3]*, Edi Prifti[1]*, Falk Hildebrand[4,5], Gwen Falony[4,5], Mathieu Almeida[1], Manimozhiyan Arumugam[2,3,6], Jean-Michel Batto[1], Sean Kennedy[1], Pierre Leonard[1], Junhua Li[3,7], Kristoffer Burgdorf[2], Niels Grarup[2], Torben Jørgensen[8,9,10], Ivan Brandslund[11,12], Henrik Bjørn Nielsen[13], Agnieszka S. Juncker[13], Marcelo Bertalan[13], Florence Levenez[1], Nicolas Pons[1], Simon Rasmussen[13], Shinichi Sunagawa[6], Julien Tap[1,6], Sebastian Tims[14], Erwin G. Zoetendal[14], Søren Brunak[13], Karine Clément[15,16,17], Joël Doré[1,18], Michiel Kleerebezem[14], Karsten Kristiansen[19], Pierre Renault[18], Thomas Sicheritz-Ponten[13], Willem M. de Vos[14,20], Jean-Daniel Zucker[15,16,21], Jeroen Raes[4,5], Torben Hansen[2,22], MetaHIT consortium†, Peer Bork[6], Jun Wang[3,19,23,24,25], S. Dusko Ehrlich[1] & Oluf Pedersen[2,26,27,28]

We are facing a global metabolic health crisis provoked by an obesity epidemic. Here we report the human gut microbial composition in a population sample of 123 non-obese and 169 obese Danish individuals. We find two groups of individuals that differ by the number of gut microbial genes and thus gut bacterial richness. They contain known and previously unknown bacterial species at different proportions; individuals with a low bacterial richness (23% of the population) are characterized by more marked overall adiposity, insulin resistance and dyslipidaemia and a more pronounced inflammatory phenotype when compared with high bacterial richness individuals. The obese individuals among the lower bacterial richness group also gain more weight over time. Only a few bacterial species are sufficient to distinguish between individuals with high and low bacterial richness, and even between lean and obese participants. Our classifications based on variation in the gut microbiome identify subsets of individuals in the general white adult population who may be at increased risk of progressing to adiposity-associated co-morbidities.

Modern living with a sedentary everyday life, a constant boom of easily accessible and energy-dense food, and exposure to additional 'obesogenic' environmental factors together with extended life expectancy has resulted in an epidemic of metabolic disorders characterized by a core of excessive body fat accumulation. Projection estimations predict that, on a global scale, cases will rise from 400 million obese adults in 2005 to more than 700 million in 2015, and this trend will continue towards 2030 (refs 1, 2). Some individuals seem to be more susceptible to the obesogenic environment of modern living than others, suggesting an important inherited component, supported by several twin, family and adoption studies, with heritability estimates ranging from 40% to 70% (refs 3–5). Studies of variation in the human genome have so far resulted in the discovery of more than 50 validated genome-wide significant loci associated with overall adiposity and body composition[6–12]. Yet, despite a reasonable number of obesity susceptibility variants identified, the proportion of explained genetic variance of body mass index (BMI) remains low, that is, a few per cent (ref. 6). Emerging evidence suggests, however, that variation in our 'other genome'—the collective genome of the microorganisms inhabiting our body, known as the microbiome—may have an even greater role than human genome variation in the pathogenesis of obesity given its direct interaction with environmental factors. Recent studies show that the human gut microbiota may be altered in obese relative to lean individuals, even if inconsistent changes have been reported. An increase in the phylum Firmicutes and a decrease in Bacteroidetes associated with obesity was observed in some[13,14], but not all, studies[15], with the inverse also reported[16]. An increase of Actinobacteria in obese individuals was also reported[17]. Mouse gut microbiota obesity-related alterations are characterized by changes in the Firmicutes to Bacteroidetes ratio, which is increased in the obese animals[18,19]. These changes are probably not a mere consequence of obesity, because the obese phenotype can be transmitted by gut microbiota transplantation in mice, indicating that gut microbial populations may have an active role in obesity pathogenesis[20,21]. Establishment of a catalogue of bacterial genes from the human gut[22] encouraged us to address the hypothesis that variation in the gut microbiome at gene and species levels defines subsets of individuals in the adult population who are at increased risk of obesity-related metabolic disorders.

[1]INRA, Institut National de la Recherche Agronomique, US1367 Metagenopolis, 78350 Jouy en Josas, France. [2]The Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, DK-2200 Copenhagen, Denmark. [3]BGI-Shenzhen, Shenzhen 518083, China. [4]Department of Structural Biology, VIB, Pleinlaan 2, 1050 Brussels, Belgium. [5]Department of Bioscience Engineering, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium. [6]European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. [7]School of Bioscience and Biotechnology, South China University of Technology, Guangzhou 510006, China. [8]Research Centre for Prevention and Health, Glostrup University Hospital, DK-2900 Glostrup, Denmark. [9]Faculty of Health and Medical Sciences, University of Copenhagen, DK-2200 Copenhagen, Denmark. [10]Institute of Public Health, Faculty of Medicine, University of Aalborg, DK-9100 Aalborg, Denmark. [11]Department of Clinical Biochemistry, Vejle Hospital, DK-7100 Vejle, Denmark. [12]Institute of Regional Health Research, University of Southern Denmark, DK-8200 Odense, Denmark. [13]Center for Biological Sequence Analysis & Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, DK-2800 Kongens Lyngby, Denmark. [14]Laboratory of Microbiology, Wageningen University, 6710BA Ede, The Netherlands. [15]Institut National de la Santé et de la Recherche Médicale, U 872, Nutriomique, Équipe 7, Centre de Recherches des Cordeliers, 75006 Paris, France. [16]Université Pierre et Marie-Curie-Paris VI, 75006 Paris, France. [17]Assistance Publique-Hôpitaux de Paris, Institute of Cardiometabolism and Nutrition, CRNH-Ile de France, Pitié-Salpêtrière, 75013 Paris, France. [18]INRA, Institut National de la Recherche Agronomique, UMR 14121 MICALIS, 78350 Jouy en Josas, France. [19]Department of Biology, Ole Maaløes Vej 5, University of Copenhagen, DK-2200 Copenhagen, Denmark. [20]Department of Bacteriology and Immunology, University of Helsinki, FIN-00014 Finland. [21]Institut de Recherche pour le Développement, UMI 209, Unité de modélisation mathématique et informatique des Systèmes Complexes, F-93143 Bondy, France. [22]Faculty of Health Sciences, University of Southern Denmark, DK-8200 Odense, Denmark. [23]King Abdulazziz University, Jeddah 21589, Saudi Arabia. [24]Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, DK-2200 Copenhagen, Denmark. [25]Center for Sequencing Aarhus University, DK-8000 Aarhus C, Denmark. [26]Hagedorn Research Institute, DK-2820 Gentofte, Denmark. [27]Institute of Biomedical Science, Faculty of Health and Medical Sciences, University of Copenhagen, DK-2200 Copenhagen, Denmark. [28]Faculty of Health, Aarhus University, DK-8000 Aarhus, Denmark.

*These authors contributed equally to this work.

†A list of authors and affiliations appears at the end of the paper.

The abundance of known intestinal bacteria can be assessed by the mapping of a large number of sequencing reads from total faecal DNA onto a reference set of their genomes[23]. This approach, which we term quantitative metagenomics, was extended here to assess the abundance of genes from the reference catalogue in a cohort of 292 non-obese and obese individuals.

Bimodal distribution of microbial genes

Comparison of gene number across the total study sample of 292 individuals showed a bimodal distribution of bacterial genes (Fig. 1a and Supplementary Table 1). A similar distribution was detected in obese French individuals using a different sequencing technology[24]. As the number of genes detected had some dependence on the number of matched reads (Supplementary Fig. 1), we downsized the data set to 11 million reads, thus excluding 15 individuals and the bimodal distribution was again observed (Fig. 1b). We term hereafter the individuals with $<$480,000 genes 'low gene count' (LGC) and others 'high gene count' (HGC). They had, on average, 380,000 and 640,000 genes, a difference of some 40% and harboured less or more rich microbiota, respectively, as shown by scoring several single copy marker genes (Supplementary Fig. 2). Human intestinal tract chip (HITChip) analysis[25], based on the widely accepted 16S ribosomal DNA phylogenetic marker, confirmed the bimodal distribution and the difference of richness of microbial communities between the LGC and HGC individuals (Supplementary Fig. 3 and Supplementary Table 1).

Low richness of gut microbiota has been reported in patients with inflammatory bowel disorder (IBD)[22,26,27], elderly patients with inflammation[28] and in obese individuals[17], but the differences of richness within these groups or among non-obese individuals were not previously detected. As the composition of the gut microbiota seems to be rather stable over long periods of adulthood[29], its richness may well be a characteristic feature of an individual. In mice, the richness seems to be affected by repeated antibiotic treatments[30], and host genetics could also have a role[31]. Also notable diversity differences were observed between the urban US population and rural populations from two developing countries[32]. Further studies, focusing specifically on the richness of the gut microbiota across broad cohorts, might help to determine the causes for its variation.

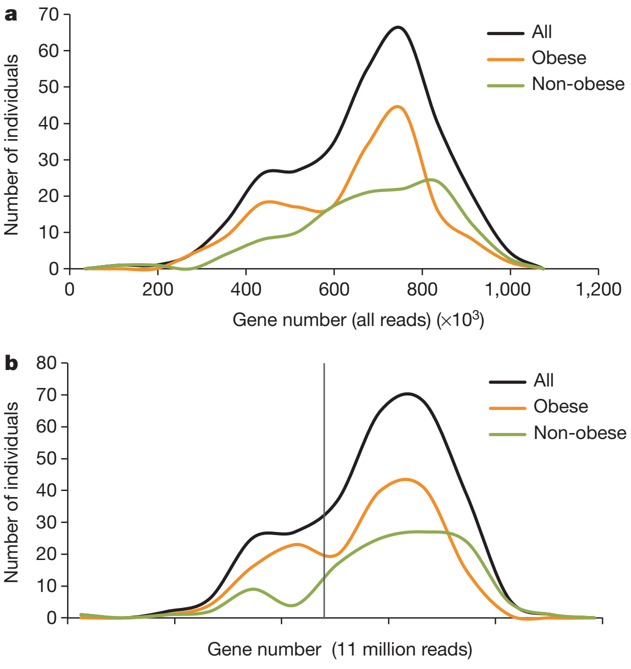

Figure 1 | Distribution of low and high gene count individuals ($n = 292$). **a**, Gene counts from all uniquely matched reads. **b**, Gene counts adjusted to 11 million uniquely mapped reads per individual. Vertical line indicates the threshold of the LGC and the HGS individuals; the observed bimodal distribution was not statistically significant by the dip-test.

Microbial species of LGC and HGC groups

The differences in gene numbers indicate that the LGC and HGC individuals contain different microbial communities. To assess the difference in phylogenetic composition between the two, we combined reference genome mapping with gene abundance data at the phylum, genus and species level.

We first examined the general phylogenetic composition at higher taxonomic levels based on sample-wise rarefied read abundances that were mapped on publicly available reference genomes and binned at the genus and phylum level. Forty-six genera differed significantly in abundance between the HGC and LGC individuals (Supplementary Table 2). Although *Bacteroides*, *Parabacteroides*, *Ruminococcus* (specifically *R. torques* and *R. gnavus*), *Campylobacter*, *Dialister*, *Porphyromonas*, *Staphylococcus* and *Anaerostipes* were more dominant in LGC, 36 genera, including *Faecalibacterium*, Bifidobacterium, *Lactobacillus*, *Butyrivibrio*, *Alistipes*, *Akkermansia*, *Coprococcus* and *Methanobrevibacter*, were significantly associated with HGC. For 33 of these, genera probes were present on the HITChip and their abundance was determined (Supplementary Table 2); the correlation with read mapping was very high (Pearson's r of 0.85 and 0.90 in HGC and LGC individuals, respectively). At the phylum level, this phylogenetic shift resulted in a higher abundance of Proteobacteria and Bacteroidetes in LGC individuals versus increased populations of Verrucomicrobia, Actinobacteria and Euryarchaeota in HGC individuals (Supplementary Fig. 4).

Next, we studied the specific species that were differentially abundant between LGC and HGC individuals. In this approach, we identified the genes that were significantly different between the LGC and HGC individuals by the Wilcoxon rank-sum test, comparing 204 (70% of total) randomly chosen individuals 30 times. We similarly compared 126 'extreme' individuals, containing $<$400,000 genes or $>$600,000 genes. A total of 120,723 genes were found in all 60 tests at $P < 0.0001$ and were analysed further.

We searched for genes from the same species by comparison with all sequenced genomes (Supplementary Materials). At a threshold of 95% identity (the species-level cut-off[23]) over at least 90% of the gene length, 10,225 genes (8.5%) were assigned to a total of 97 genomes representing some 73 species (Supplementary Table 3). However, a vast majority (93.4%) belonged to only 9 species, which varied significantly in abundance between the LGC and HGC individuals, as illustrated in Fig. 2a (upper part), where the presence and abundance of 50 arbitrarily chosen 'tracer' genes from each species in the individuals of the cohort is displayed. As expected, these genes have a sharply bimodal distribution: 71% of individuals had either all or none of the genes from a species and thus harboured or lacked that species, at the present depth of analysis. The first five species were more frequent in LGC individuals, whereas the last four species were more frequent in the HGC group.

Taken together, our analyses highlight the contrast between the distribution of anti-inflammatory species, such as *Faecalibacterium prausnitzii*[33,34], which are more prevalent in HGC individuals and potentially pro-inflammatory, *Bacteroides* and *R. gnavus*[35], associated with IBD[36,37] and found to be more frequent in LGC individuals.

However, a vast majority ($>$90%) of the 120,723 genes with significantly differing abundances between the LGC and HGC gene individuals could not be assigned to a known bacterial genome. These genes must also belong to bacterial species that are present at different abundances in the two groups of individuals. We thus attempted to cluster the genes from the same species using a gene abundance-based approach.

We proposed that the genes of a given bacterial species should be present at a similar abundance in an individual but should display large variations across a cohort, as species abundance is known to vary immensely among individuals (10–10,000-fold)[22]. The genes that vary in abundance in a coordinated way are thus likely to be from the same

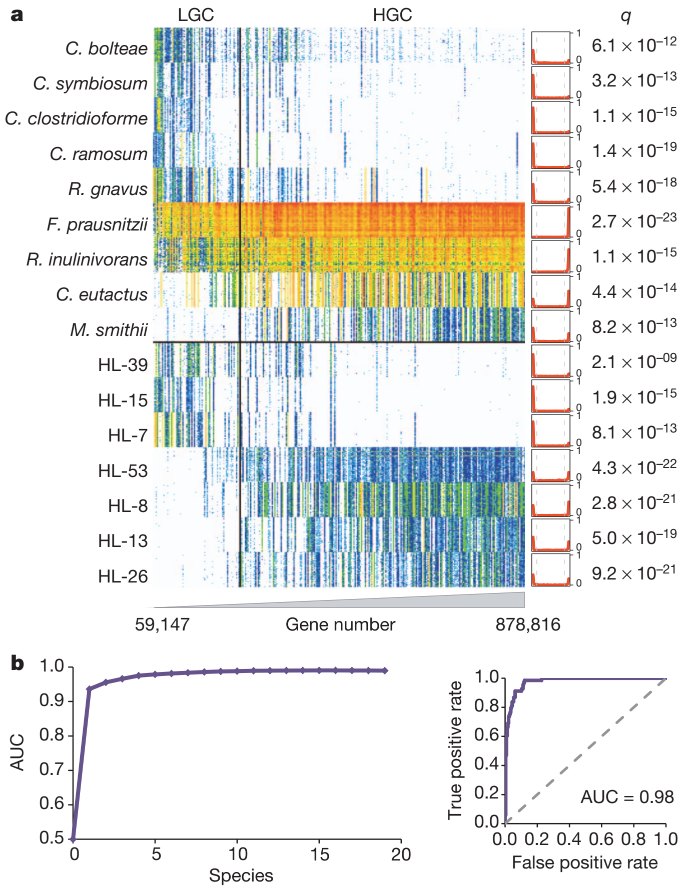

Figure 2 | Bacterial species with different distribution among 292 HGC and LGC individuals. **a**, Presence and abundance of 50 'tracer' genes for representative species differentially abundant in LGC and HGC groups; Mann–Whitney probability (*q* value, FDR adjusted[40]) is given. Genes are in rows, frequency is indicated by colour gradient (white, not detected; red, most abundant); individuals, ordered by increasing gene number, are in columns, proportion of the tracer genes in individuals is shown on the right. Top, known species; bottom, unknown species. *C. bolteae, Clostridium bolteae*; *C. clostridioforme, Clostridium clostridioforme*; *C. eutactus*, *Coprococcus eutactus*; *C. ramosum*, *Clostridium ramosum*; *C. symbiosum*, *Clostridium symbiosum*; *F. prausnitzii, Faecalibacterium prausnitzii*; *M. smithii, Methanobrevibacter smithii*; *R. gnavus, Ruminococcus gnavus*; *R. inulinivorans*, *Roseburia inulinivorans*. **b**, Left, AUC values for the best combinations of species in a ROC analysis. Right, AUC for the best combination of four species.

species. We tested this hypothesis for the 10,225 taxonomically assigned genes that differ significantly between LGC and HGC individuals, by computing the Spearman correlation coefficients for each gene with all the other genes and grouping those that were correlated above a given threshold. A large majority (8,125; 79.4%) clustered into only 8 groups that included the 9 most highly represented species shown in Fig. 2 (*Clostridium bolteae* and *C. clostridioforme* genes were in the same group). The specificity and the sensitivity of clustering were very high (average of 97.8% and 91.8%, respectively, for 7 homogeneous groups), indicating that the approach is efficient and can be used to cluster all the significantly different genes.

In total, 76,564 genes (63% of 120,723) were grouped into 1,440 clusters of two or more genes at a threshold of rho > 0.85, used to favour the specificity of clustering. Some 58 clusters contained ≥ 75 genes and included 90% of the genes; 52 contained genes from previously unknown species (Supplementary Table 4). Genes from a cluster originated from the same species in most cases, as shown by (1) coherence of the BLASTP taxonomic assignments; (2) homogeneity of abundance and abundance variation; (3) homogeneity of tetramer composition[38]; and (4) significant physical linkage (Supplementary Figs 5–8). This conclusion was further supported by the correlation of the abundance of the 16S rRNA gene sequences represented on the HITChip for 27 of the 58 clusters (Supplementary Table 4; it is possible that the HITChip resolution of closely related genomes may have been insufficient or that the corresponding 16S sequences were lacking for 31 clusters). We conclude that the clustering procedure grouped the genes of the same species, even if in some cases genes from more than a single species were grouped (*C. bolteae* and *C. clostridioforme*, or probably prophages present in genomes from several species; Supplementary Fig. 9). A similar approach was applied in a recent paper, published while the present manuscript was under revision[39].

Analyses of the 50 genes from the clusters with known taxonomic assignment have shown that they are present on cognate genomes only and on all cognate genomes (Supplementary Table 5). By extrapolation, we suggest that the same holds true for unknown groups and that the cluster genes can be used as tracers for the species they derive from. Average abundance of the tracer genes was thus equated to the average abundance of the cognate species.

Distribution of unknown species across LGC and HGC individuals of the cohort was clearly biased (Fig. 2a, lower part, and Supplementary Fig. 10). Genes for 10 of the species and the Bacteroides genus were present on the metagenomic arrays (Methods); in all cases the HGC/LGC bias found by sequencing was also detected by the arrays (Fig. 2 and Supplementary Table 4). BLASTP analysis indicated that 37% and 92% of the clusters had at least 80% of the genes with the same taxonomy at a genus and phylum level, respectively, a value similar to that observed for all clusters analysed in this paper (Supplementary Fig. 11). HITChip and BLASTP taxonomic assignments were not fully overlapping but whenever both were available they were congruent; when combined, up to 63% of the species could be assigned to a genus. However, there was no obvious stratification of the species prevalent in LGC and HGC individuals at this taxonomic level (Supplementary Table 4).

To test whether LGC and HGC individuals could be distinguished by the bacterial species they contain, we performed a receiver operating characteristic (ROC) analysis. First, we estimated the abundance of 58 species that were significantly different in abundance between LGC and HGC individuals (Supplementary Table 6). For each individual, we used these values to compute a score, named decisive bacterial abundance (DBA) score, equal to the sum of abundances of the species more frequent in HGC individuals subtracted by the sum of the abundances of species more frequent in LGC individuals. The DBA scores were calculated exhaustively for all combinations of up to 19 species and were used in the ROC analysis; the area under curve (AUC) values for the best combinations are shown in Fig. 2b, left. The best combination of four species gave an AUC value of 0.98 (Fig. 2b, right); in a tenfold cross-validation test[40] with 90% of randomly chosen individuals the AUC value of 0.976 ± 0.02 (mean ± s.d.) was obtained for the groups of the remaining 10%, indicating the robustness of the analysis. Selection of the most distinctive species on the entire cohort does not seem to lead to a significant over-fit, as the algorithm established with one cohort gives comparably high AUC values with an unrelated cohort[24]. Future work, searching for correlations of gene abundances and gene counts without separation of individuals into the LGC and HGC groups, may allow identifying additional species that explain variation of gene numbers.

Microbial metabolism of LGC and HGC groups

Through the functional annotation of the reference gene catalogue to KEGG Orthology (KO) groups, abundances of KO groups were determined for the LGC and HGC individuals as in ref. 22. Using the enzyme annotations of the different KO groups, 51 manually defined gut metabolic pathway modules (Supplementary Table 7) differed significantly in abundance between both groups (Methods). LGC individuals had a higher abundance of peroxidase, catalase and TCA modules, suggesting increased capacity to handle exposure to oxygen/oxidative stress. Furthermore, the genomic potential for production of metabolites with possible deleterious effects on host health (among which pro-carcinogens)—including modules for β-glucuronide degradation, degradation of aromatic amino acids, and dissimilatory nitrate reduction—was significantly higher in LGC participants. Many of the significantly increased modules could be due to the increased abundance of *Bacteroides* spp. (for example, pectin degradation). By contrast, HGC individuals were characterized by a potentially increased production of organic acids—including lactate, propionate and butyrate—combined with a higher hydrogen production potential. Concerning hydrogen removal, a shift from a methanogenic/acetogenic ecosystem in HGC individuals towards a sulphate-reducing one in LGC individuals might take place. The functional capacity of the microbiota in LGC and HGC individuals, when combined with the phylogenetic signal, leads to several interesting observations: in the former group we see (1) a reduction of butyrate-producing bacteria; (2) increased mucus degradation potential combined with a decreased *Akkermansia* to *R. torque/gnavus* ratio); (3) reduced hydrogen and methane production potential combined with increased hydrogen sulphide formation potential; (4) an increase in *Campylobacter/Shigella* abundance; and (5) an increased potential to manage oxidative stress (peroxidase). Overall, this suggests that LGC individuals harbour an inflammation-associated microbiota (Fig. 3).

Phenotypes of the HGC and LGC groups

Characteristics of study materials are given in Supplementary Table 8. We performed an anthropometric and biochemical phenotyping of multiple interrelated features of LGC and HGC individuals, and identified significant differences between them at a false discovery rate (FDR)[41] of up to 10% (Table 1 and Supplementary Table 9). This value was used to avoid missing significant associations; a less stringent level, up to 25%, was chosen in a recent and comparable study design[42]. The LGC individuals, who represented 23% of the total study population, included a significantly higher proportion of obese participants (Fig. 4a; the difference is significant for men and a trend is detected for women), and were as a group characterized by a more marked adiposity, as reflected by an increase in fat mass percentage and body weight (Table 1). The adiposity phenotype of LGC people was associated with increased serum leptin, decreased serum adiponectin, insulin resistance, hyperinsulinaemia, increased levels of triglycerides and free fatty acids, decreased HDL-cholesterol and a more marked inflammatory phenotype (increased highly sensitive C-reactive protein (hsCRP) and higher white blood cell counts) than seen in HGC individuals (Table 1). We further tested the significance of our observations by treating the gene counts as a continuous variable and examining its correlation with the anthropometric and biochemical variables. All but two (BMI and weight) of the observed differences between LGC and HGC individuals were found significantly associated with the gene counts (Table 1). Together, these analyses suggest that the LGC individuals are featured by metabolic disturbances known to bring them at increased risk of pre-diabetes, type-2-diabetes and ischaemic cardiovascular disorders[43,44]. Similar abnormalities were found in the accompanying paper[24].

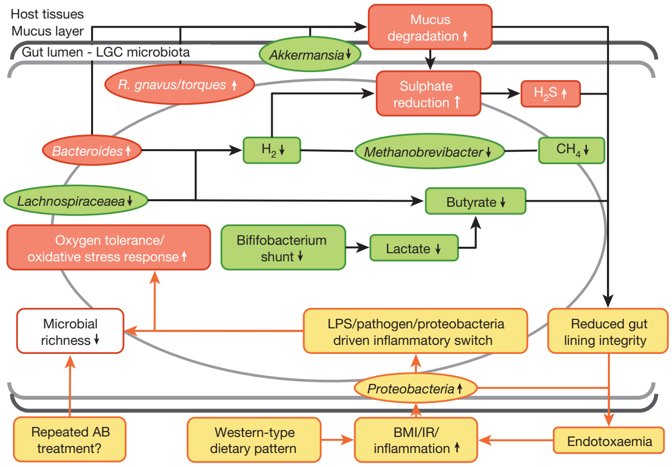

Figure 3 | Functional and phylogenetic shifts in the LGC microbiome. Top, observed increase (red) or decrease (green) of functions and phylogenetic groups. Bottom, potential drivers (yellow) of inflammation related to decreased richness. Left, antibiotic-mediated perturbation of the richness; Right, proteobacterial lipopolysaccharide-mediated perturbation of the richness. AB, antibiotic; IR, insulin resistance.

We propose that an imbalance of potentially pro- and anti-inflammatory bacterial species triggers low-grade inflammation and insulin resistance (Fig. 3). In parallel, we suggest that an altered gut microbiota of LGC individuals induces the noted increase in levels of serum fasting induced adipose factor (FIAF, also known as ANGPLT4), eliciting an increased release of triglycerides and free fatty acids (Table 1), as evidenced by studies in rodent models[45–47]. Broad spectra antibiotics may improve glycaemic regulation and change the hormonal, inflammatory and metabolic status in obese mice. Possibly, the reduction of diversity mediated by the antibiotic treatments has a different effect in mice[48] and man, or, alternatively, is counterbalanced by the restoration of the pro- and anti-inflammatory species balance, which may have been altered in the obese animals. Antibiotic use in early childhood, which may have affected the richness, led to an increased risk of overweight[49].

Interestingly, obese LGC individuals gained on average significantly more weight than HGC individuals during the past 9 years (Fig. 4b); the BMI change was significant without and with linear adjustment for baseline BMI and age. We searched for species associated to the BMI change among the 58 species that differed significantly between LGC and HGC individuals (Supplementary Table 6) and found eight (Fig. 4c). The average weight gain of individuals with the lowest or undetectable levels of a species was in all cases greater than that of their counterparts with the highest species levels; all eight species were more abundant in HGC than in LGC individuals. These species may therefore protect against weight gain. All but one (*Methanobrevibacter smithii*) lack species-level taxonomic assignment, but four could be assigned at a genus level (*Anareotruncus colihominis*, *Butyrivibrio crossotus* and *Faecalibacterium*; Supplementary Table 4). All are butyrate producers, in agreement with the general tendency of lower butyrate producers among the LGC individuals.

Gut microbes of lean and obese individuals

We also attempted to assess the difference in bacterial species between the lean ($\mathrm{BMI} < 25\,\mathrm{kg\,m^{-2}}$, $n = 96$) and obese ($\mathrm{BMI} > 30\,\mathrm{kg\,m^{-2}}$, $n = 169$) individuals by the approach used for LGC/HGC individuals (Methods). Only 15,894 significantly different genes ($P < 0.05$) were found, indicating that the gut microbiota of lean and obese individuals differs less than that of the LGC and HGC individuals. The genes were attributed to 18 species by the covariance-based clustering (Supplementary Fig. 12 and Supplementary Tables 10 and 11). To test whether lean and obese individuals can be distinguished by these species, we carried out an exhaustive ROC analysis, with tenfold cross validation (Supplementary Fig. 12). The best AUC, of 0.78, was reached with nine species. This accuracy, albeit lower than that for the separation of LGC and HGC individuals, is substantially better than an AUC of 0.58, achieved by ROC analysis of 32 human genome loci associated with adiposity measures[6]. Accordingly, we suggest that the obesity-associated signal in the human gut microbiome may be much stronger than that presently known in the human genome. This view is supported by efficient discrimination of lean and obese individuals in a previous study, in which an AUC of 0.88 was reached with a combination of 50 16S-defined operational taxonomic units (OTUs), separated at the 92% homology level[50].

Discussion

Contemporary lifestyle is associated with a tide of metabolic abnormalities characterized by a core of excessive body fat accumulation. However, obesity is not just obesity. Some obese individuals seem to

Table 1 | Characteristics of 292 participants stratified by low and high gene counts

	LGC	HGC	LGC versus HGC		Gene count	
			P	*q*	*P*	*q*
N (men/women)	68 (23/45)	224 (113/111)			277 (133/144)*	
Age (years)	55 (50–62)	57 (50–61)	0.86	0.89	0.81	0.84
BMI (kg m^{-2})	32 (29–34)	30 (23–33)	0.035	0.059	0.11	0.18
Weight (kg)	95 (75–103)	86 (71–102)	0.019	0.037	0.12	0.18
Whole body fat (%)	37 (29–42)	31 (25–39)	0.0069	0.022	0.0024	0.014
S-insulin (pmol l^{-1})	50 (35–91)	44 (26–66)	0.0095	0.023	0.0052	0.018
HOMA-IR	1.9 (1.2–3.3)	1.6 (0.9–2.6)	0.012	0.027	0.0059	0.018
P-triglycerides (mmol l^{-1})	1.32 (0.97–1.76)	1.15 (0.82–1.57)	0.0014	0.013	0.00073	0.0062
P-free fatty acids (mmol l^{-1})	0.55 (0.39–0.70)	0.48 (0.35–0.60)	0.014	0.029	0.00042	0.0062
P-ALT (U l^{-1})	20 (14–30)	19 (15–26)	0.22	0.31	0.029	0.06
S-leptin (μg l^{-1})	17.0 (6.7–32.6)	8.3 (3.4–26.4)	0.0036	0.019	0.00058	0.0062
S-adiponectin (mg l^{-1})	7.5 (5.5–12.9)	9.6 (6.7–13.7)	0.006	0.022	0.016	0.036
B-leucocytes (10^{9} l^{-1})	6.4 (5.2–7.8)	5.6 (4.8–6.9)	0.0021	0.014	0.0026	0.014
B-lymphocytes (10^{9} l^{-1})	2.1 (1.6–2.3)	1.8 (1.5–2.1)	0.00082	0.012	0.0037	0.015
P-hsCRP (mg l^{-1})	2.3 (1.1–5.7)	1.4 (0.6–2.7)	0.00088	0.012	0.0038	0.015
S-FIAF (μg l^{-1})	88 (72–120)	78 (60–101)	0.0047	0.021	0.0088	0.023

Descriptive data are reported as median and interquartile range. To test for differences between the HGC and LGC group, a linear model adjusting for age and sex (*P*) was applied. In the analysis of plasma triglycerides, treatment for lipid lowering medications was added as a covariate to the linear model. Benjamini–Hochberg method was used for multiple testing corrections setting the FDR at 10% (*q*). A similar model was applied to test for associations with gene counts. The P-, S- and B- prefixes denote plasma, serum and blood. ALT, alanin aminotransferase; HOMA-IR, homeostatic model assessment of insulin resistance.
*$n = 277$ owing to downsizing of the reads to 11 million.

have a benign prognosis, whereas others progress to co-morbidities such as type-2 diabetes, ischaemic cardio- and cerebrovascular disorders, and non-alcoholic liver disorders. It is also recognized that human obesity in the context of pathogenesis, pathophysiology and therapeutic responsiveness is a heterogeneous condition. The present report provides evidence that studies of alterations in our other genome— the microbial gut metagenome—may define subsets of adult individuals with different metabolic risk profiles and thereby contribute to resolve some of the heterogeneity associated with adiposity-related phenotypes.

We demonstrate that an almost perfect stratification of LGC and HGC individuals can be achieved with very few bacterial species, suggesting that simple molecular diagnostic tests, based on our other genome, can be developed to identify individuals at risk of common morbidities. Therefore, focus on our other genome may spearhead development of stratified approaches for treatment and prevention of widespread chronic disorders. Beyond metabolic dysfunctions, low-grade inflammation is associated with a plethora of chronic diseases. Whether a low gut bacterial richness is common to many or even all of those, as already reported for IBD[22,26,27], could be revealed by exploring gut microbiota at a deep metagenomic level in a broad variety of these afflictions.

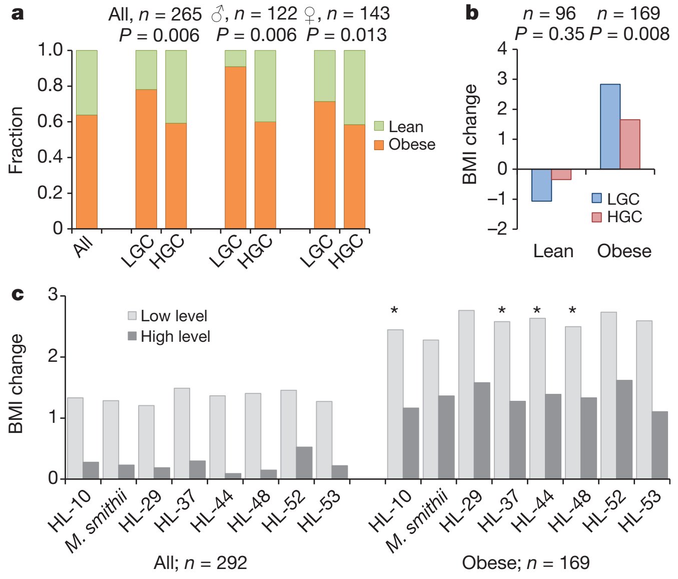

Figure 4 | Evolution of BMI in obese and non-obese LGC and HGC individuals ($n = 265$). **a**, LGC individuals are more frequently obese. **b**, LGC obese individuals gained more weight over 9 years. **c**, Bacterial species associated with weight change. BMI change was computed for at least 125 (all) or 60 (obese) individuals having undetectable or lowest level of a species (‘Low level’) and for at least 40 (all) or 30 (obese) individuals having the highest abundance of the same species (‘High level’). Average BMI change of low and high level groups was significantly different ($P < 0.05$ or $*P < 0.01$), with the exception of HL-52 for all individuals ($P = 0.052$); for obese, BMI and age were adjusted.

METHODS SUMMARY

Informed consent was obtained from all 292 volunteers from the Ethical Committees of the Capital Region of Denmark before participation in the study. All individuals were examined after an overnight fast with blood sampling and anthropometric measurements. Faeces samples were collected from all volunteers and frozen immediately; DNA was extracted and sequenced as described[22]. Gene frequency profiles were established by matching the sequencing reads from an individual sample onto the gut bacterial catalogue[22], genes significantly different by frequency between the groups of individuals were identified by Wilcoxon rank-sum test. Genes were clustered by abundance covariance at a selected Spearman correlation coefficient. Taxonomic gene assignments were carried out using BLASTN against an in-house reference catalogue of 3,340 complete and draft microbial genomes downloaded from NCBI databanks at a cut-off of 95% identity[23]. Functional annotation was carried out by BLASTP against the eggNOG and KEGG databases as reported[22]. Abundance of a given species in each individual was estimated as a mean abundance of 50 tracer genes for that species, a value very close to that of the mean abundance of all genes of that species. ROC analysis was based on a combination of bacterial species. For each combination, only a single model was considered, based on the DBA score, computed as the sum of abundances of the species more frequent in one group of individuals subtracted by the sum of the abundances of species more frequent in another group. To select the best models, we used the cross-validated AUC criterion, well adapted to classification models for binary outcome data. Association of microbial composition and metabolic traits were analysed as cross-sectional or continuous variable, upon appropriate normalization; the Benjamini–Hochberg[41] method to correct for multiple testing was used, setting the FDR at 10%.

Full Methods and any associated references are available in the online version of the paper.

Received 10 April 2012; accepted 26 July 2013.

1. World Health Organization. Obesity and overweight. Fact sheet no. 311; http://www.who.int/mediacentre/factsheets/fs311/en/ (2006).
2. Kelly, T., Yang, W., Chen, C. S., Reynolds, K. & He, J. Global burden of obesity in 2005 and projections to 2030. *Int. J. Obes.* **32,** 1431–1437 (2008).
3. Stunkard, A. J., Harris, J. R., Pedersen, N. L. & McClearn, G. E. The body-mass index of twins who have been reared apart. *N. Engl. J. Med.* **322,** 1483–1487 (1990).
4. Allison, D. B. *et al.* The heritability of body mass index among an international sample of monozygotic twins reared apart. *Int. J. Obes. Relat. Metab. Disord.* **20,** 501–506 (1996).
5. Maes, H. H., Neale, M. C. & Eaves, L. J. Genetic and environmental factors in relative body weight and human adiposity. *Behav. Genet.* **27,** 325–351 (1997).

6. Speliotes, E. K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genet.* **42,** 937–948 (2010).
7. Frayling, T. M. *et al.* A common variant in the *FTO* gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* **316,** 889–894 (2007).
8. Loos, R. J. *et al.* Common variants near *MC4R* are associated with fat mass, weight and risk of obesity. *Nature Genet.* **40,** 768–775 (2008).
9. Willer, C. J. *et al.* Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nature Genet.* **41,** 25–34 (2009).
10. Thorleifsson, G. *et al.* Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. *Nature Genet.* **41,** 18–24 (2009).
11. Heid, I. M. *et al.* Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nature Genet.* **42,** 949–960 (2010).
12. Lindgren, C. M. *et al.* Genome-wide association scan meta-analysis identifies three Loci influencing adiposity and fat distribution. *PLoS Genet.* **5,** e1000508 (2009).
13. Ley, R. E., Turnbaugh, P. J., Klein, S. & Gordon, J. I. Microbial ecology: human gut microbes associated with obesity. *Nature* **444,** 1022–1023 (2006).
14. Furet, J. P. *et al.* Differential adaptation of human gut microbiota to bariatric surgery-induced weight loss: links with metabolic and low-grade inflammation markers. *Diabetes* **59,** 3049–3057 (2010).
15. Duncan, S. H. *et al.* Human colonic microbiota associated with diet, obesity and weight loss. *Int. J. Obes.* **32,** 1720–1724 (2008).
16. Schwiertz, A. *et al.* Microbiota and SCFA in lean and overweight healthy subjects. *Obesity* **18,** 190–195 (2010).
17. Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457,** 480–484 (2009).
18. Backhed, F. *et al.* The gut microbiota as an environmental factor that regulates fat storage. *Proc. Natl Acad. Sci. USA* **101,** 15718–15723 (2004).
19. Ley, R. E. *et al.* Obesity alters gut microbial ecology. *Proc. Natl Acad. Sci. USA* **102,** 11070–11075 (2005).
20. Turnbaugh, P. J., Backhed, F., Fulton, L. & Gordon, J. I. Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome. *Cell Host Microbe* **3,** 213–223 (2008).
21. Turnbaugh, P. J. *et al.* An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444,** 1027–1031 (2006).
22. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464,** 59–65 (2010).
23. Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473,** 174–180 (2011).
24. Cotillard, A. *et al.* Dietary intervention impact on gut microbial gene richness. *Nature* http://dx.doi.org/10.1038/nature12480 (this issue).
25. Rajilić-Stojanovic, M. *et al.* Development and application of the human intestinal tract chip, a phylogenetic microarray: analysis of universally conserved phylotypes in the abundant microbiota of young and elderly adults. *Environ. Microbiol.* **11,** 1736–1751 (2009).
26. Manichanh, C. *et al.* Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. *Gut* **55,** 205–211 (2006).
27. Lepage, P. *et al.* Twin study indicates loss of interaction between microbiota and mucosa of patients with ulcerative colitis. *Gastroenterology* **141,** 227–236 (2011).
28. Claesson, M. J. *et al.* Gut microbiota composition correlates with diet and health in the elderly. *Nature* **488,** 178–184 (2012).
29. Costello, E. K. *et al.* Bacterial community variation in human body habitats across space and time. *Science* **326,** 1694–1697 (2009).
30. Cho, I. *et al.* Antibiotics in early life alter the murine colonic microbiome and adiposity. *Nature* **488,** 621–626 (2012).
31. Vijay-Kumar, M. *et al.* Metabolic syndrome and altered gut microbiota in mice lacking Toll-like receptor 5. *Science* **328,** 228–231 (2010).
32. Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* **486,** 222–227 (2012).
33. Sokol, H. *et al.* *Faecalibacterium prausnitzii* is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. *Proc. Natl Acad. Sci. USA* **105,** 16731–16736 (2008).
34. Devillard, E., McIntosh, F. M., Duncan, S. H. & Wallace, R. J. Metabolism of linoleic acid by human gut bacteria: different routes for biosynthesis of conjugated linoleic acid. *J. Bacteriol.* **189,** 2566–2570 (2007).
35. Png, C. W. *et al.* Mucolytic bacteria with increased prevalence in IBD mucosa augment *in vitro* utilization of mucin by other bacteria. *Am. J. Gastroenterol.* **105,** 2420–2428 (2010).
36. Swidsinski, A., Weber, J., Loening-Baucke, V., Hale, L. P. & Lochs, H. Spatial organization and composition of the mucosal flora in patients with inflammatory bowel disease. *J. Clin. Microbiol.* **43,** 3380–3389 (2005).
37. Joossens, M. *et al.* Dysbiosis of the faecal microbiota in patients with Crohn's disease and their unaffected relatives. *Gut* **60,** 631–637 (2011).
38. Yang, B. *et al.* Unsupervised binning of environmental genomic fragments based on an error robust selection of l-mers. *BMC Bioinformatics* **11** (suppl. 2), S5 (2010).
39. Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490,** 55–60 (2012).
40. Chevaleyre, Y., Koriche, F. & Zucker, J.-D. Rounding methods for discrete linear classification. *Proc. 30th Int. Conf. Machine Learning (ICML-13)* 651–659 (2013).
41. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57,** 289–300 (1995).
42. Wu, G. D. *et al.* Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334,** 105–108 (2011).
43. Ouchi, N., Parker, J. L., Lugus, J. J. & Walsh, K. Adipokines in inflammation and metabolic disease. *Nature Rev. Immunol.* **11,** 85–97 (2011).
44. Shoelson, S. E., Lee, J. & Goldfine, A. B. Inflammation and insulin resistance. *J. Clin. Invest.* **116,** 1793–1801 (2006).
45. 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).
46. Backhed, F., Manchester, J. K., Semenkovich, C. F. & Gordon, J. I. Mechanisms underlying the resistance to diet-induced obesity in germ-free mice. *Proc. Natl Acad. Sci. USA* **104,** 979–984 (2007).
47. Mandard, S. *et al.* The fasting-induced adipose factor/angiopoietin-like protein 4 is physically associated with lipoproteins and governs plasma lipid levels and adiposity. *J. Biol. Chem.* **281,** 934–944 (2006).
48. Membrez, M. *et al.* Gut microbiota modulation with norfloxacin and ampicillin enhances glucose tolerance in mice. *FASEB J.* **22,** 2416–2426 (2008).
49. Ajslev, T. A., Andersen, C. S., Gamborg, M., Sorensen, T. I. & Jess, T. Childhood overweight after establishment of the gut microbiota: the role of delivery mode, pre-pregnancy weight and early administration of antibiotics. *Int. J. Obes.* **35,** 522–529 (2011).
50. Sun, Y. *et al.* Advanced computational algorithms for microbial community analysis using massive 16S rRNA sequence data. *Nucleic Acids Res.* **38,** e205 (2010).

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

Acknowledgements The authors wish to thank A. Forman, T. Lorentzen, B. Andreasen, G. J. Klavsen and M. M. Andersen for technical assistance; A. L. Nielsen, G. Lademann and M. M. H. Kristensen for management assistance, K. Kiil for discussions and assistance, and A. Walker for comments on the manuscript. This research has received funding from European Community's Seventh Framework Program (FP7/2007-2013): MetaHIT, grant agreement HEALTH-F4-2007-201052. Additional funding came from The Lundbeck Foundation Centre for Applied Medical Genomics in Personalized Disease Prediction, Prevention and Care (LuCamp, http://www.lucamp.org), ANR MicroObes, the Metagenopolis grant ANR-11-DPBS-0001, Region Ile de France (CODDIM) and Fondacoeur. The Novo Nordisk Foundation Center for Basic Metabolic Research is an independent Research Center at the University of Copenhagen partially funded by an unrestricted donation from the Novo Nordisk Foundation (http://www.metabol.ku.dk).

Author Contributions O.P. and S.D.E. designed the study, O.P., S.D.E., P.B., W.J., S.B., K.C., J.D., M.K., P.R., T.S.-P., W.M.d.V., T.H., J.R. and K.K. managed the study. T.N., K.B., T.H., N.G., T.J., I.B. and O.P. carried out patient phenotyping and clinical data analyses. T.N., K.B. and F.L. performed sample collection and DNA extraction. J.Q. and J.L. supervised DNA sequencing and gene profiling. S.D.E. and O.P. designed and supervised the data analyses. E.L.C., E.P., T.N., N.G., G.F., F.H., M.Al., M.Ar., J.-M.B., S.K., P.L., N.P., S.S., J.T., J.Q., J.L., J.-D.Z., S.R. and S.D.E. performed the data analyses. S.T. and E.G.Z. carried out HITChip analysis. M.B., A.S.J., H.B.N. and T.S.-P. carried out metagenomic array analyses. S.D.E., O.P., J.R. and P.B. wrote the paper. MetaHIT consortium members provided creative environment and constructive criticism throughout the study.

Author Information The raw Illumina read data for all samples has been deposited in the EBI European Nucleotide Archive under the accession number ERP003612. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to P.B. (bork@embl.de, for bioinformatics), W.J. (wangj@genomics.org.cn, for sequencing), S.D.E. (dusko.ehrlich@jouy.inra.fr, for any aspect of this work) or O.P. (oluf@sund.ku.dk, for physiological aspects of this work).

MetaHIT consortium additional members

Eric Guedon[1], Christine Delorme[1], Séverine Layec[1], Ghalia Khaci[1], Maarten van de Guchte[1], Gaetana Vandemeulebrouck[1], Alexandre Jamet[1], Rozenn Dervyn[1], Nicolas Sanchez[1], Emmanuelle Maguin[1], Florence Haimet[2], Yohanan Winogradski[1], Antonella Cultrone[1], Marion Leclerc[1], Catherine Juste[1], Hervé Blottière[1,2], Eric Pelletier[3,4,5], Denis LePaslier[3,4,5], François Artiguenave[3,4,5], Thomas Bruls[3,4,5], Jean Weissenbach[3,4,5], Keith Turner[6], Julian Parkhill[6], Maria Antolin[7], Chaysavanh Manichanh[7], Francesc Casellas[7], Natalia Boruel[7], Encarna Varela[7], Antonio Torrejon[7], Francisco Guarner[7], Gérard Denariaz[8], Muriel Derrien[8], Johan E. T. van Hylckama Vlieg[8], Patrick Veiga[8], Raish Oozeer[9], Jan Knol[9], Maria Rescigno[10], Christian Brechot[11], Christine M'Rini[11], Alexandre Mérieux[11] & Takuji Yamada[12]

[1]INRA, Institut National de la Recherche Agronomique, UMR 14121 MICALIS, 78350 Jouy en Josas, France. [2]INRA, Institut National de la Recherche Agronomique, US1367 Metagenopolis, 78350 Jouy en Josas, France. [3]Commissariat à l'Energie Atomique, Genoscope, 91000 Evry, France. [4]Centre National de la Recherche Scientifique, UMR8030, 91000 Evry, France. [5]Evry, France, Université d'Evry Val d'Essone. 91000 Evry, France. [6]The Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK. [7]Digestive System Research Unit, University Hospital Vall d'Hebron, Ciberehd, 08035 Barcelona, Spain. [8]Danone Research, 91120 Palaiseau, France. [9]Gut Biology & Microbiology, Danone Research, Centre for Specialized Nutrition, Bosrandweg 20, 6704 PH Wageningen, The Netherlands. [10]Istituto Europeo di Oncologia, 20100 Milan, Italy. [11]Institut Mérieux, 17 rue Burgelat, 69002 Lyon, France. [12]European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany.

METHODS

Study population. The study participants were recruited from the Inter99 study population. The Inter99 study is a randomized, non-pharmacological intervention study for the prevention of ischaemic heart disease, and was conducted at the Research Centre for Prevention and Health in Glostrup, Denmark, between 1999 and 2006 (clinicalTrials.gov: NCT00289237)[51]. The participants in the Inter99 study were examined at baseline, after 1, 3 and 5 years depending on the type of intervention.

For the present study individuals with BMI $< 25\,\mathrm{kg\,m^{-2}}$ or BMI $> 30\,\mathrm{kg\,m^{-2}}$ at year five in the Inter99 study were randomly selected from track records. They had no known gastrointestinal disease, no previously bariatric surgery, no medications known to affect the immune system, and no antibiotics 2 months before faecal sample collection. Individuals with type-2 diabetes at the day of examination were excluded. In total, 292 non-diabetic individuals were included in the protocol. All had North European ethnicity. At the time of the current physical examination, 96 (33%) of study volunteers were lean with BMI $< 25\,\mathrm{kg\,m^{-2}}$, 27 (9%) were overweight with BMI between 25 and $30\,\mathrm{kg\,m^{-2}}$, and 169 (58%) were obese with BMI $> 30\,\mathrm{kg\,m^{-2}}$, according to World Health Organization definitions[52]. The study was approved by the local Ethical Committees of the Capital Region of Denmark (HC-2008-017), and was in accordance with the principals of the Declaration of Helsinki. All individuals gave written informed consent before participation in the study.

Phenotyping. The participants were examined on two different dates, approximately 14 days apart. On the first day, participants were examined in the morning after an overnight fast. Height was measured without shoes to the nearest 0.5 cm, and weight was measured without shoes and wearing light clothes to the nearest 0.1 kg. Hip and waist circumference were recorded using a non-expandable measuring tape to the nearest 0.5 cm. Waist circumference was measured midway between the lower rib margin and the iliac crest. Hip circumference was measured as the largest circumference between the waist and the thighs. On the second day of examination, all participants delivered a stool sample collected at home and dual-emission X-ray absorptiometry (DXA) was performed. Analyses of data from the DXA scan were conducted with the integrated software (Hologic Discovery A, Santax). Sagittal height was measured at the time of the DXA scan with the use of the Holtain–Kahn abdominal caliper at the highest point of the abdomen with the participant supine and while breathing out. Participant receiving statins, fibrates and/or ezetimibe were reported as receiving lipid-lowering medication.

Derived anthropometrical measure and index of insulin resistance. Intra-abdominal adipose tissue (cm^2) was calculated using data from DXA scans and anthropometry using the equation[53]: $y = -208.2 + 4.62$ (sagittal diameter, cm) + 0.75 (age, years) + 1.73 (waist, cm) + 0.78 (trunk fat, %). HOMA-IR was calculated as: (fasting plasma glucose ($\mathrm{mmol\,l^{-1}}$) × fasting serum insulin ($\mathrm{mU\,l^{-1}}$))/22.5 (ref. 54).

Biochemical measurements. All analyses were performed on blood samples drawn in the morning after an overnight fast from at least 22:00 the previous evening.

Plasma glucose was analysed by a glucose oxidase method (Granutest, Merck) with a detection limit of $0.11\,\mathrm{mmol\,l^{-1}}$ and intra- and inter-assay coefficients of variation (CV) of <0.8 and $<1.4\%$, respectively. HbA1c was measured on TOSOH G7 by ion-exchange high performance liquid chromatography.

Serum insulin (excluding des-31,32-proinsulin and intact proinsulin) was measured using the AutoDELFIA insulin kit (Perkin-Elmer, Wallac) with a detection limit of $3\,\mathrm{pmol\,l^{-1}}$ and with intra- and inter-assay CV of $<3.2\%$ and $<4.5\%$, respectively. Plasma total cholesterol, plasma HDL-cholesterol and plasma triglycerides were all measured on Vitros 5600 using reflect-spectrophotometrics. Blood leucocytes and white blood cell differential count were measured on Sysmex XS 1000i using flow cytometrics. Plasma ALT and plasma total free fatty acids were analysed using standard biochemical methods (Modular Evo). Plasma hsCRP was analysed by a particle-enhanced immunoturbidmetric assay on MODULAR Evo using CRPL3 kit (Roche) with a detection limit of $0.3\,\mathrm{mg\,l^{-1}}$ and intra- and inter-assay CV of $<4.0\%$ and $<6.2\%$, respectively.

Serum adiponectin was analysed using a two-site-sandwich ELISA kit for measuring total human adiponectin (TECO). Detection limit was $0.6\,\mathrm{ng\,ml^{-1}}$ and intra- and inter-assay CV were $<4.66\%$ and $<6.72\%$, respectively. Serum FIAF was measured using a quantitative sandwich ELISA (Adipo Bioscience). Detection limit was $0.6\,\mu\mathrm{g\,l^{-1}}$ and the intra- and inter-assay CV of 4% and 8%, respectively. Serum lipopolysaccharide binding protein was analysed by a solid phase sandwich ELISA kit (Abnova) with intra- and inter-assay CV of $<6.1\%$ and $<17.8\%$, respectively. Serum IL-6 and serum TNF-α were analysed by Luminex using the Bio-Plex Pro cytokine assay (Bio-Rad), whereas serum leptin was measured using the Bio-Plex Pro diabetes assay.

Faecal sampling. Stool samples were obtained at the homes of each participant and samples were immediately frozen in their home freezer. Frozen samples were delivered to Steno Diabetes Center using insulating polystyrene foam containers, and stored at $-80\,^{\circ}\mathrm{C}$ until analysis. The time span from sampling to delivery at the Steno Diabetes Center was aimed to be as short as possible and no more than 48 h.

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[26]. The DNA concentration and its molecular size were estimated by nanodrop (Thermo Scientific) and on agarose gel electrophoresis.

Illumina sequencing. DNA library preparation followed the manufacturer's instruction (Illumina). We used the workflow indicated by the provider to perform cluster generation, template hybridization, isothermal amplification, linearization, blocking and denaturing 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 base pairs (bp)) for each of the first batch of 15 samples; two libraries with different clone insert sizes (135 and 400 bp) for each of the second batch of 70 samples, and one library (350 bp) for each of the third batch of 207 samples.

After sequencing, we performed quality control and screened human genome contaminant. Finally, we generated 26.0 million–186.1 million high-quality reads for the 292 samples, with an average of 68.2 million high-quality reads. Sequencing read length of the first batch of 15 samples was 44 bp, the second batch was 75 bp, and the third batch was 75 bp and 90 bp.

Microbial gene abundance profiling by quantitative metagenomics. An average of 34.1 million paired-end reads were produced for each sample and, after removing human contamination ($\sim$0.1%, on average), 19.9 ± 6.7 million reads were mapped at a unique position of the reference catalogue of 3.3 million genes, using SOAP2.21 (ref. 55) by allowing at most two mismatches in the first 35-bp region and 90% identity over the read sequence; reads mapping at multiple positions (13.4%, on average) were discarded. The abundance of a gene in a sample was estimated by dividing the number of reads that uniquely mapped to that gene by the gene length and by the total number of reads from the sample that uniquely mapped to any gene in the catalogue. The resulting set of gene abundances, termed a microbial gene profile of an individual, was used for further analyses; Illumina and Solid sequencing platforms gave highly similar gene profiles for the same DNA samples[24].

Gene counting. On the basis of the pair-oriented counting result of each samples, we selected the threshold of one read for gene identification to include the rare genes into the analysis. We identified 91,032–1,005,488 genes for the 292 samples, with an average of 670,528 genes.

Read downsizing. To eliminate the influence of sequencing depth fluctuation, we sampled the alignment results and downsized the number of mapped pairs to 11 million for each sample. After that, we found 59,147–878,816 genes for the 292 samples, with an average of 578,512 genes.

Diversity estimate by single copy gene scoring. Genes belonging to the orthologous groups COG0085, COG052, and COG0090 from 3,515 prokaryotic genomes were clustered to OTUs at 95% identity by UCLUST and used as a reference database. Paired-end Illumina reads from 292 metagenomic samples were mapped at 95% identity cut-off using soap2.21 (ref. 55). The numbers of fragments that were assigned to the reference sequences were counted so that each fragment's weight equals 1, that is, a fragment assigned to *N* different reference sequences contributes $1/N$ to each reference sequence. Fragment counts of reference sequences were grouped to yield OTU counts. Samples with low sampling effort, that is, with less than 3,000 fragments mapped to reference genes, were removed leaving 229 samples for comparative analyses. OTU counts were normalized by gene length, scaled by the maximum count across all marker genes, and down-sampled using the vegan package to the minimum sum of OTU counts across all samples to compare species richness between high and low gene content groups.

Phylogenetic microarray analysis. HITChip microarray analyses were performed as described previously[25]. In short, 16S rRNA genes were amplified the T7prom-Bact-27-for and Uni-1492-rev primers from 10 ng fecal DNA extracts. On these amplicons an *in vitro* transcription and subsequent labelling with Cy3 and Cy5 dyes were performed. Labelled RNA was fragmented and hybridized on the arrays at $62.5\,^{\circ}\mathrm{C}$ for 16 h in a rotation oven (Agilent Technologies). The arrays were washed, dried, scanned and the signal intensity data was extracted as described (http://www.agilent.com). Microarray data normalization and analysis were carried out with a set of R-based scripts (http://r-project.org), while making use of a custom designed database, which operates under the MySQL database management system (http://www.mysql.com).

From the 3,699 unique HITChip probes, we selected the probes that accounted for the top 99.9% of the total signal. These probes were counted for each sample to measure richness, which was between 713 and 1,597 probes per sample. The probes that accounted for the lowest 0.1% of the total signal were regarded as background noise and were not taken into account for further analysis. Probe signal values were used to calculate the inverse Simpson's diversity index for each sample.

HITChip probes specificity can be assigned to three phylogenetic levels based on 16S rRNA gene sequence similarity: order-like groups, genus-like groups (sequence similarity >90%), and phylotype-like groups (sequence similarity >98%)[24]. Relative abundances were calculated for each specificity level by summing all signal values of the probes targeting a group and dividing by the total of all probe signals for the corresponding sample. All comparisons between the HGC and LGC individuals were assessed with dependent two-group Wilcoxon signed rank tests. When statistical tests were performed on a large number of variables the obtained P values were adjusted by a Bonferroni correction. To place the gene count and BMI marker species (HL and oble, referring to species significantly different between HGC/LGC and obese/lean individuals, respectively) in HITChip phylogeny, Spearman correlation coefficients were calculated between the metagenomic profiling frequencies and relative abundances of the phylotype-like across 251 samples. A threshold of 0.7 was used to associate 16S to a species.

The Pearson correlation coefficient between the log(gene count) and log(probe count) was 0.8 and the concordance between assignments of individuals to a low or high richness class by the two techniques (gene counts or probe counts) was 88%.

Metagenomic microarray analysis. A 2.1-million-feature custom Roche NimbleGen microarray targeting a 700,000 genes subset of the MetaHit human gut gene catalogue[22] was designed and manufactured. The subset of genes was prioritized for genes that were observed in more than 20 of the 124 gene catalogue samples. DNA extracted from faecal samples were labelled and hybridized according to standard NimbleGen protocols. Data was pre-processed and Shannon diversity index calculated using the RMA implementation under the 'oligo' package and the vegan package, respectively, both available in the statistical programming environment R.

To validate the observed biomarkers for low/high gene counts found by sequencing, we compared the data to DNA microarray signals for the same samples and individuals. Thus, the tracer genes for known and unknown species indicated in Fig. 2 were compared to a microarray gene set comprising more than 700,000 gut-associated genes selected from the MetaHit Gene Catalog[8] in addition to reference genomes. Perfect matches were found for 129 tracer genes on the DNA microarray. To test whether a similar discrimination could be obtained from the microarray data, the samples were divided into low and high diversity sets using the Shannon diversity index. Using this index, 90 samples were categorized as low diversity, and 70 were categorized as high. Differences in DNA abundance signals between low and high diversity samples were tested for the 129 matching genes (t-test). Summarized, in terms of species the following groups were associated to low diversity, *C. clostridioforme/C. bolteae*, HL-7, HL-39, *R. gnavus*, HL-15 and *Bacteroides*, whereas HL-53 and *M. smithii* were associated to high diversity. These DNA microarray observations support the quantitative metagenomics results (Fig. 2 and Supplementary Table 4).

Phylogenetic annotation. Taxonomic assignment of predicted genes for global analysis was carried out using BLASTN to assign reads to a reference genome database at a cut-off of 95% sequence identity and >100 bp overlap, unless indicated otherwise. This assignment was used as high confidence assignment on species level. As reference database we used 1,869 available reference genomes from NCBI and the set of draft gastrointestinal genomes from the DACC (http://hmpdacc.org/), both as of the 15 July 2011 version. The assigned reads to each taxonomic group per sample were rarefied to 5.5 million reads (the size of the smallest sample), on this rarefied matrix taxonomic groups were tested for significant differences in abundance using a Wilcoxon rank-sum test. Multiple testing corrections were performed using the Benjamini–Hochberg method[40] ($q < 0.1$). From the same matrix we calculated the genus abundance as the percentage of reads assigned from 5.5 million total reads per sample; this matrix was used to calculate the class-wise means as well as standard deviations expressed in percentages (Supplementary Table 2).

Functional annotation. BLASTP was used to search the protein sequences of the predicted genes in the eggNOG database[56] and KEGG database[57] with $E \leq 10^{-5}$ as described previously[8], and the NOG/KEGG OG of the best hit was assigned to each gene. The genes annotated by COG were classified into the 25 COG categories, and genes that were annotated by KEGG were assigned to a set of manually determined gut metabolic modules (G.W. *et al.*, manuscript in preparation). The relative pathway/module abundance of higher order functional categories were calculated from rarefied KO abundances. Modules were deemed present when ≥30% of the enzymes were recovered, after manual removing of overly 'promiscuous' enzymes (that is, present in multiple modules) before abundance calculation. For higher-level functional assignments, KO abundances were summed and distributed evenly when KO groups appeared in multiple categories. Functional differences were calculated with a Wilcoxon rank-sum test and multiple testing corrections were performed using the Benjamini–Hochberg method[40] ($q < 0.05$).

Genes significantly different in groups of individuals. Genes significantly different in abundance between groups of individuals were identified by the Wilcoxon rank-sum test coupled to a bootstrapping approach.

Approximately 70% of the whole cohort (204 individuals) was randomly chosen and genes differentially abundant between LGC and HGC individuals were identified at $P \leq 0.0001$ as threshold. This test was repeated 30 times. We also composed 30 groups of randomly chosen 'extreme' individuals that had <400,000 or >600,000 genes and applied the same test. Genes common to all 60 tests were further analysed.

For lean and obese individuals we used a similar approach by randomly choosing 70% of individuals 30 times and using Wilcoxon rank-sum test at $P \leq 0.05$.

Gene clustering method. We used an unsupervised strategy to cluster genes of the same species. Such genes are expected to be present at a similar abundance in an individual but at different abundances in different individuals. The genes that vary in abundance in a coordinated way are thus likely to be from the same species. The clustering algorithm developed in Delphi6 programming language consists of two steps: (1) Spearman correlation coefficients were determined for all pairs of genes of each gene set, using the covariance abundance of the genes among the 292 individuals, and the SpearmanRankCorrelation function included in the free delphi correlation.pas library of the ALGLIB project (see http://www.alglib.net/). (2) All the genes correlated above a given threshold were assigned to the same cluster. If a gene of a given cluster was correlated with a gene of another cluster above the threshold, the genes of the two clusters were merged (single linkage algorithm).

A similar clustering approach was recently described, using Kendall tau instead of Spearman R to compute correlation coefficients and double linkage (all genes of a cluster are correlated above the threshold) rather than single linkage[38]. We have not compared the performance of the two methods, but rather characterized the outcome of the one we used.

Cluster characterization. To test whether the tracer genes originate from the same species we carried out a number of analyses, using as control 135 clusters composed of randomly chosen catalogue genes.

(1) BLASTN. Some 17 clusters contained genes that matched a reference genome at a threshold of 95% identity over 90% of gene length when mapped against a collection of 6,006 genomes (the available reference genomes from NCBI and the set of draft gastrointestinal genomes from the DACC and MetaHIT as of the 3 August 2012 version. The taxonomic assignment was highly coherent (Supplementary Fig. 5). We denote these clusters hereafter as 'known', referring to their taxonomy at the above-mentioned thresholds. Similarly, for clusters containing genes that did not reach these thresholds, the genes were uniformly not assigned to known genomes. We denote these clusters as 'unknown'. A single exception to the rule was a cluster having a high identity (97–99%, on average) with four different genomes, *Pseudoflavonifractor capilosus*, *F. prausnitzii L2-6*, *F. prausnitzii A2-16* and *Subdoligranulum variabile DSM16176*. The genes of this cluster were present in the same order on all genomes (Supplementary Fig. 9); we suggest that they derive from a (possibly defective) prophage or a conjugative transposon integrated in the chromosome (BLASTP analysis reveals functions involved in DNA metabolism and transfer, such as DNA helicases or tra genes). A similar grouped localization of the cluster genes was observed for the *Roseburia inulinivorans* genome revealed in the accompanying manuscript (cluster MO-HL-16)[24], suggesting that they also originate from a prophage or a transposon. These observations point to the capacity of our clustering method to group genes not only of bacteria but also of mobile and promiscuous elements, and indicate a potential source of false positives, as regards grouping of genes from a single species. Identification of all the catalogue genes that co-vary by abundance with the tracer genes could probably help to differentiate clusters that represent bacterial species from those that originate from promiscuous genetic elements, as the former should include substantially more genes than the latter (species encode mostly >500 genes while the promiscuous genetic elements encode generally <100). Such systematic analysis is, however beyond the scope of the current work.

(2) Abundance covariance of the genes of each cluster was computed across all individuals of our cohort and also across the French cohort from the accompanying manuscript[24]. The genes of both known and unknown clusters co-varied by abundance in the two data sets (Danish and French, Supplementary Fig. 6; the values were higher and more homogeneous in the former than the latter, presumably reflecting the data set size: 292 Danes and 49 French). For the randomly constituted groups pairwise correlation values were close to zero (0.01 ± 0.006).

(3) Abundance similarity. Abundance covariance does not imply that the genes of a cluster have a similar abundance in an individual, as only the abundance ratios could be constant. Nevertheless it is expected that the genes belonging to the same species show similar abundance profiles as they are carried by the same DNA molecule; this is a basic postulate of our approach to gene clustering. To estimate abundance similarity we computed for each cluster a score of gene

abundance variation for each individual as a ratio of the standard deviation of the gene abundance over the mean in each individual. The known and unknown clusters show very small and similar scores indicating the homogeneity of the abundance signal. That is not the case for the randomly constituted groups (Supplementary Fig. 7).

(4) k-mer composition. This composition is similar across a given genome and different from that of unrelated genomes[37]. We computed first the tetranucleotide profile of each gene of a cluster and then the pairwise correlation of the profiles of all genes of a cluster. Genes that belong to the known and unknown clusters have similar highly correlated 4-mer profiles, whereas the genes of the control random groups have a much lower correlation (Supplementary Fig. 8, left). Because the cluster gene size (1,135 ± 446 bp) is relatively low, we retrieved from the catalogue the contigs that carry the genes of each group. Indeed, about 2,000 bp is estimated to be necessary for accurate measure of the 135 non-redundant 4-mer combinations[37]. The contigs are longer than genes (4,084 ± 3,678 bp) and have a higher correlation of 4-mer profiles (Supplementary Fig. 8, right).

(5) Physical linkage of the genes. We retrieved from the gene catalogue the scaffolds (composed of contigs that are 'bridged' by a paired-end sequence; one end is in one contig and the other in another, but the two sequences are not contiguous) that bear the tracer genes and searched for the presence of more than one gene of a cluster on a scaffold. Most of the known and unknown groups had at least some genes on the same scaffold (on average, about 18%) whereas randomly composed clusters had none, a difference of high significance ($P < 2 \times 10^{-16}$).

We conclude that the genes from a cluster originate, as expected, from the same genome.

As only a minority of the tracer gene clusters (15%) could be assigned to species-level taxonomy by BLASTN, we used BLASTP against either a collection of 6,006 available genomes or the non-redundant sequences databases available at NCBI to assess the taxonomy of other clusters. On the basis of the criterion of the homogeneity of the best-hit taxonomic assignment (with the threshold of at least 80% of a cluster genes having the same taxonomic best hit assignment), 91.9% and 45.2% of the clusters could be assigned at a phylum and genus level, respectively (Supplementary Fig. 11 and Supplementary Tables 4 and 11). In all cases the higher-level taxonomic assignments were congruent with the lower level ones.

Species abundance determination. We assessed whether the genes of a cluster can be used as tracers for a species they originate from, using the known clusters as benchmark. We examined whether the genes are either homologous to only the cognate genome of a species when a single strain of a species was sequenced, or homologous to all cognate genomes when multiple strains of a species were sequenced. The results are summarized in Supplementary Table 5. Eight clusters were in the first category and five in the second. Three of the latter matched two genomes of inconsistent species annotation, but which belong to the same species as deduced from the BLASTN comparison of the two genomes, which had an average identity of the reciprocal best hit genes comprised between 99.5% and 99.8%. A single cluster matched two of the five sequenced *F. prausnitzii* genomes with 98% average identity and three other genomes with an identity of 80–84%. However, the three last genomes had an average reciprocal best-hit gene identity with the two well-matched genomes of only 80.4–86% and therefore should probably be considered as belonging to a different species. We conclude that, for the known groups, the cluster genes are present (1) on cognate genomes only, and (2) on all cognate genomes. By extrapolation, we suggest that the same holds true for unknown groups and that the cluster genes can be used as tracers for the species they derive from.

Abundance of a given species in each individual was estimated as a mean abundance of 50 tracer genes of each cluster. The values were very close to the mean frequency of all the genes of a cluster.

ROC analysis. The analyses were carried out to distinguish between HGC and LGC individuals or lean and obese individuals by a combination of bacterial species. For each combination, only a single decision model was considered. In this very specific regression model weights are only allowed to take the values in (0, −1, 1). More precisely, the weight of each species in a given combination that belong to the set of the species more frequent in one group is equal to 1, whereas that of the species that belong to the set of species more frequent in the other group is equal to −1. The weight of each species that is outside of the combination is 0. For each individual, this model yields a DBA score. As opposed to the infinite number of regression models, such ternary models are finite and can be exhaustively explored. To select the best models, we used the cross-validated AUC criterion[39] well adapted to classification models for binary outcome data.

Species correlated with the BMI change. For the entire cohort of 292 individuals, 40 individuals (14%) having the highest abundance of a species were compared with at least 125 individuals (42%) having the lowest abundance (all individuals lacking a species were included, when more numerous than 125); these numbers were chosen to allow contrasting the extremes of the distribution while keeping the sample size high enough to reduce the probability of a fortuitous difference in BMI change. For the 169 obese individuals, 30 (18%) having the highest abundance of a species were compared with at least 60 individuals (36%) having the lowest abundance (all individuals lacking a species were included, when more numerous than 60). The differences were calculated with a Student's t-test, the BMI changes being normally distributed, and multiple testing correction was performed using the Benjamini–Hochberg method[41] ($q < 0.05$).

Association of microbial composition and metabolic traits. We analysed the association of (1) the high gene and low gene group, and (2) gene count as a continuous trait to quantitative traits applying a linear model adjusting for age and sex. Plasma triglycerides, plasma HDL cholesterol, serum insulin, plasma ALT, serum leptin and serum adiponectin and HOMA-IR were log transformed, whereas blood leucocytes, lymphocytes, monocytes, neutrophilocytes, plasma hsCRP, serum FIAF, plasma free fatty acids, serum TNF-α, serum interleukin (IL)-6, serum lipopolysaccharide binding protein and BMI were rank normalized before analyses in the linear model. In the analyses of triglycerides, treatment with lipid lowering medications was added as a covariate to the linear model. We corrected for multiple testing by the Benjamini–Hochberg method[41], setting the FDR at 10%.

Bacterial species that discriminate between lean and obese individuals. To assess the difference in bacterial species between the lean (BMI $< 25\,\mathrm{kg\,m^{-2}}$, $n = 96$) and obese (BMI $>30\,\mathrm{kg\,m^{-2}}$, $n = 169$) individuals we searched for the genes significantly different by abundance. Some 15,894 were found at $P < 0.05$, a value lower than that for the LGC and HGC individuals, indicating that the gut microbiota of lean and obese individuals differs less. In total, 14,149 could be clustered into 187 groups by covariance, at a Spearman threshold of rho > 0.75, and 90% of these (12,753) were found in only 18 groups, ranging from 2,507 to 68 genes (Supplementary Table 10); four were correlated with a 16S rRNA gene by a co-variance-based HITChip analysis (Supplementary Table 11). The species represented by the 18 clusters had a significantly different distribution among the lean and obese individuals (Supplementary Fig. 12 and Supplementary Table 10); 4 were more frequent among the obese and 14 among the lean individuals.

51. Jørgensen, T. *et al.* A randomized non-pharmacological intervention study for prevention of ischaemic heart disease: baseline results Inter99. *Eur. J. Cardiovasc. Prev. Rehabil.* **10,** 377–386 (2003).
52. World Health Organization. Preventing and managing the globalepidemic. Report of a WHO consultation. *World Health Organ. Tech. Rep. Ser.* **894,** 1–253 (2000).
53. Treuth, M. S., Hunter, G. R. & Kekes-Szabo, T. Estimating intraabdominal adipose tissue in women by dual-energy X-ray absorptiometry. *Am. J. Clin. Nutr.* **62,** 527–532 (1995).
54. Matthews, D. R. *et al.* Homeostasis model assessment: insulin resistance and beta-cell function from fasting plasma glucose and insulin concentrations in man. *Diabetologia* **28,** 412–419 (1985).
55. Li, R. *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25,** 1966–1967 (2009).
56. Jensen, L. J. *et al.* eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res.* **36,** D250–D254 (2008).
57. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **40,** D109–D114 (2012).