LETTER

doi:10.1038/nature12480

Dietary intervention impact on gut microbial gene richness

Aurélie Cotillard[1,2]*, Sean P. Kennedy[3]*, Ling Chun Kong[1,2,4]*, Edi Prifti[1,2,3]*, Nicolas Pons[3]*, Emmanuelle Le Chatelier[3], Mathieu Almeida[3], Benoit Quinquis[3], Florence Levenez[3,5], Nathalie Galleron[3], Sophie Gougis[4], Salwa Rizkalla[1,2,4], Jean-Michel Batto[3,5], Pierre Renault[5], ANR MicroObes consortium†, Joel Doré[3,5], Jean-Daniel Zucker[1,2,6], Karine Clément[1,2,4] & Stanislav Dusko Ehrlich[3]

Complex gene–environment interactions are considered important in the development of obesity[1]. The composition of the gut microbiota can determine the efficacy of energy harvest from food[2–4] and changes in dietary composition have been associated with changes in the composition of gut microbial populations[5,6]. The capacity to explore microbiota composition was markedly improved by the development of metagenomic approaches[7,8], which have already allowed production of the first human gut microbial gene catalogue[9] and stratifying individuals by their gut genomic profile into different enterotypes[10], but the analyses were carried out mainly in non-intervention settings. To investigate the temporal relationships between food intake, gut microbiota and metabolic and inflammatory phenotypes, we conducted diet-induced weight-loss and weight-stabilization interventions in a study sample of 38 obese and 11 overweight individuals. Here we report that individuals with reduced microbial gene richness (40%) present more pronounced dys-metabolism and low-grade inflammation, as observed concomitantly in the accompanying paper[11]. Dietary intervention improves low gene richness and clinical phenotypes, but seems to be less efficient for inflammation variables in individuals with lower gene richness. Low gene richness may therefore have predictive potential for the efficacy of intervention.

To examine relationships between variations in gut microbiota composition and bioclinical parameters after dietary intervention, we used the approach termed quantitative metagenomics[11]. Forty-nine obese or overweight subjects were recruited and subjected to a 6-week energy-restricted high-protein diet followed by a 6-week weight-maintenance diet (Methods); the compliance was good, as indicated by a principal component analysis (PCA) of 35 nutrients over time (Supplementary Fig. 1). Bioclinical characteristics and detailed qualitative and quantitative features of individuals' food intake were obtained at baseline, 6 and 12 weeks (Supplementary Tables 1 and 2). The 35% decrease in energy intake after the first 6 weeks was associated with a reduction in body-fat mass, adipocyte diameter and improvements in insulin sensitivity and markers of metabolism and inflammation (Supplementary Tables 1 and 3). During the weight-maintenance phase, intake of nutrients tended to return to baseline values, whereas dietary total energy, carbohydrate and lipid intake remained lower than at beginning of the intervention (Supplementary Tables 2 and 3). Serum lipid variables also tended to return to their basal levels as well, while a progressive reduction occurred in systemic inflammation markers.

We first examined the gut microbial composition of the study population at baseline (Methods). A bimodal distribution of bacterial gene number was observed (Fig. 1a), similar to the one found in a cohort of 292 Danish individuals[11], albeit less distinct, possibly owing to a lower cohort size. At a threshold of 480,000 genes, corresponding to that from the accompanying manuscript[11], there were 18 (40%) low gene count (LGC) and 27 (60%) high gene count (HGC) individuals, harbouring on average 379,436 and 561,499 genes respectively, a one-third difference. A difference in diversity between lean and obese individuals was reported previously[12], but the difference among the obese was not described.

We then examined the baseline phenotypes of the study population. The LGC group had significantly higher insulin resistance and fasting serum triglyceride levels, as well as a tendency towards higher LDL cholesterol and inflammation than the HGC group (Fig. 2); as observed in the accompanying paper[11]. Analysing gene richness as a quantitative variable gave similar results (Supplementary Table 4). We conclude that in two European countries, the individuals of the LGC group present phenotypes that expose them to an increased risk of obesity-associated co-morbidities. Antibiotic treatments, which lower the diversity, have

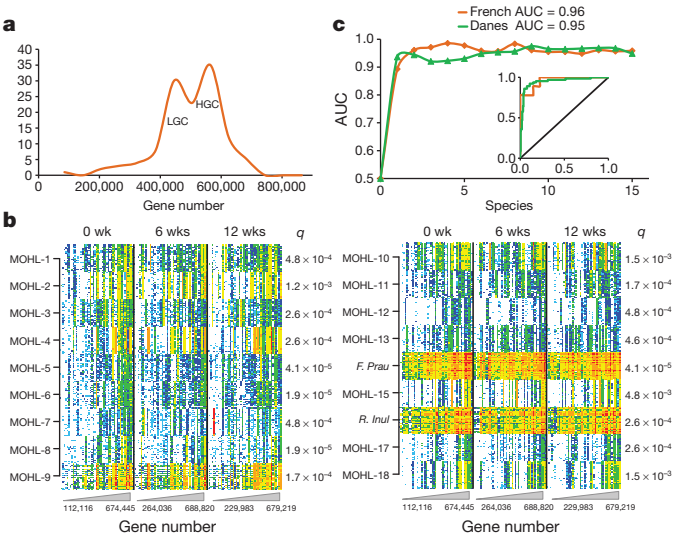

Figure 1 | Gut microbial composition of LGC ($n = 18$) and HGC ($n = 27$) subjects. **a**, Baseline gene count. **b**, Presence and frequency of 25 tracer genes for species differentially abundant in LGC and HGC groups; Mann–Whitney probability (q, false discovery rate (FDR) adjusted) is given. Genes are in rows, frequency is indicated by colour gradient (white, not detected; red, most abundant); individuals, ordered by increasing gene number, are in columns. **c**, Highest AUC values for a combination of a given number of species in a ROC analysis of 45 individuals of our cohort (red) and 292 individuals of the Danish cohort[11]. Inset, AUC for the combination of six species.

[1]Institut National de la Santé et de la Recherche Médicale, U872, Nutriomique, Équipe 7, Centre de Recherches des Cordeliers, Paris 75006, France. [2]Université Pierre et Marie-Curie-Paris 6, Nutriomique, 15 rue de l'Ecole de Medecine, Paris 75006, France. [3]INRA, Institut National de la Recherche Agronomique, Metagenopolis, Jouy en Josas78350, France. [4]Institute of Cardiometabolism and Nutrition, Assistance Publique-Hôpitaux de Paris, CRNH-Ile de France, Pitié-Salpêtrière, Boulevard de l'Hopital, Paris 75013, France. [5]INRA, Institut National de la Recherche Agronomique, UMR 1319 Micalis, Jouy en Josas 78350, France. [6]Institut de Recherche pour le Développement, IRD, UMI 209, UMMISCO, France Nord, Bondy F-93143, France.

*These authors contributed equally to this work.

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

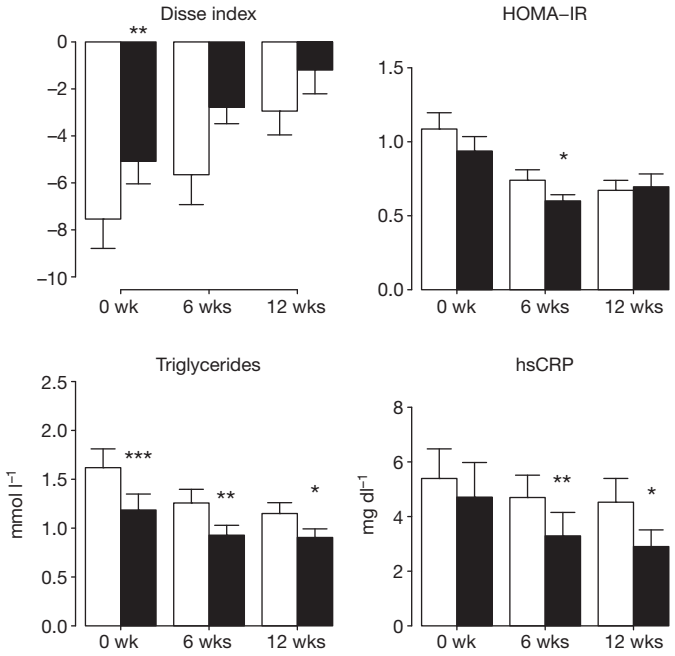

Figure 2 | Differences between LGC and HGC subjects in bioclinical variables. White and black bars refer to LGC ($n = 18$) and HGC ($n = 27$) groups, respectively; error bars denote s.e.m. 0 weeks, baseline; 6 weeks, end of the energy restriction period; and 12 weeks, end of stabilization period. $^*P < 0.1$, $^{**}P < 0.05$, $^{***}P < 0.01$ by Mann–Whitney tests. 'Disse index' is calculated by combining lipid and insulin values (see Supplementary Information). HOMA-IR, homeostatic model assessment of insulin resistance; hsCRP, highly sensitive C-reactive protein.

been reported to improve the hormonal, metabolic and inflammatory status of obese mice; this apparent contradiction may be due to a restoration of a balance of the pro-inflammatory and inflammatory bacterial species in mice. Interestingly, LGC subjects seemed to consume less fruits and vegetables and less fishery products than HGC subjects (Supplementary Tables 4–6), raising the possibility that long-term dietary habits may affect gene richness and the associated phenotypes, as suggested for the elderly[13].

We next searched for bacterial species differentially abundant in the LGC and the HGC groups. To this aim, we first identified the genes that had significantly different frequencies in the LGC and HGC groups and then clustered the genes supposedly from the same species by a frequency-based covariance analysis (Methods). We identified 6,230 genes that were different according to a Mann–Whitney test ($P < 0.0001$); 4,462 (72%) were grouped into 112 clusters containing at least 2 genes with a Spearman correlation coefficient $\rho > 0.85$. A vast majority of these genes (3,966; 89%) were found in only 18 clusters, which originate from species differentially abundant in the LGC and HGC groups (Supplementary Table 7). The relative abundance of the 18 clusters in each individual was computed as a mean frequency of the 25 tracer genes for each cluster; all were significantly more abundant among the HGC individuals (Fig. 1b and Supplementary Table 8).

To test whether the LGC and HGC individuals could be distinguished using the 18 species represented by the tracer genes, we carried out an exhaustive receiver operating characteristic (ROC) analysis of all clusters combinations, with tenfold cross validation, using 90% of individuals for computation and the remaining 10% for test (Methods). The best area under the curve (AUC) values for combination of different numbers of species are shown in Fig. 1c; they ranged between 0.96 and 0.99 for 2 to 9 species combinations, indicating an almost perfect stratification of LGC and HGC individuals. Interestingly, 14 of the 18 species represented by the tracer genes (78%) were also identified as differentially abundant among the LGC and HGC individuals in a larger Danish cohort[11]. Not surprisingly, the combinations yielding the best AUC values for our cohort also efficiently stratified LGC and HGC Danes (Fig. 1c). This indicates that the LGC and HGC individuals from two European countries differ in a similar way, not only by their clinical phenotypes but also by specific features of their gut microbiota.

Very interestingly, gene richness increased significantly in the LGC group after the energy-restricted diet and remained after the stabilization phase higher than at baseline even though a slight downwards trend was apparent, whereas it did not change significantly during intervention in the HGC group (Fig. 3a). We conclude that a dietary intervention can correct a putative loss of richness in the LGC group, albeit partially, as the difference between the LGC and HGC groups remained significant at the end of the intervention.

To investigate the potential effect of the increase in gene richness on patient status we analysed association of the changes of richness and of bioclinical variables. Increase of gene richness was associated with a significant decrease in adiposity measures (hip circumference and total fat mass) and circulating cholesterol as well as a trend towards a decrease in inflammation (highly sensitive C-reactive protein) (Supplementary Table 9). These results suggest that the correction of a putative loss of microbial richness is associated with an improvement of the systemic metabolic status. However, although the inflammation was decreased in all individuals, the difference between LGC and HGC individuals was not attenuated (Fig. 2). Low basal gene richness was also associated with increased adipose tissue inflammatory cells at 6 weeks and increased

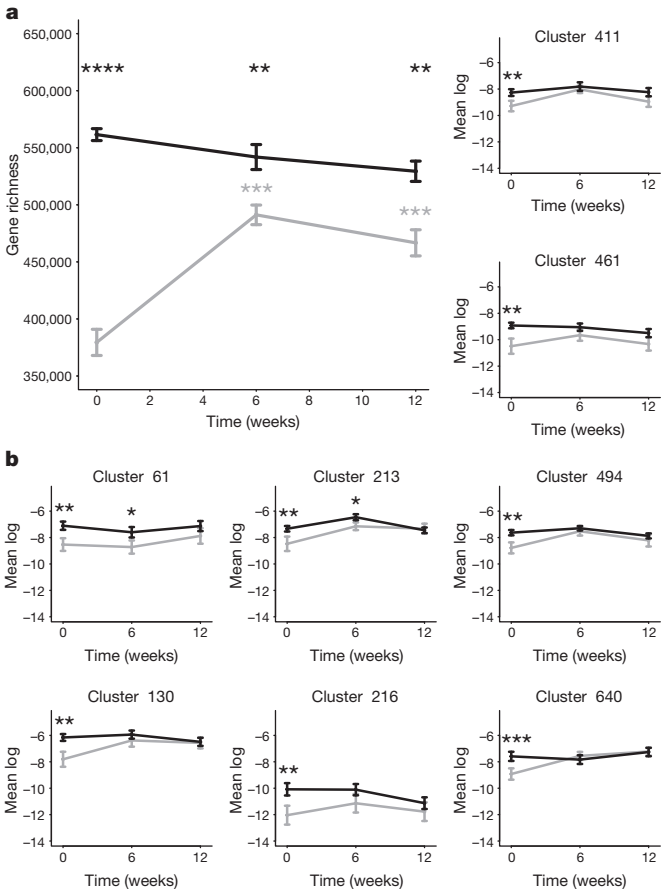

Figure 3 | Gene richness of LGC and HGC groups during the intervention. Data are mean ± s.e.m. Black line, HGC ($n = 27$); grey line, LGC ($n = 18$). Differences between HGC and LGC groups were tested using Mann–Whitney tests (black asterisks). $^*P < 0.1$, $^{**}P < 0.05$, $^{***}P < 0.01$, $^{****}P < 0.001$. **a**, Overall pattern of variation. For each group, differences between one time point and basal state were tested using Wilcoxon signed-rank tests (grey asterisks). **b**, Variation of eight clusters that were significantly different at baseline and modulated by the dietary intervention.

systemic inflammation at 12 weeks (Supplementary Table 4). Furthermore, higher gene richness at baseline was associated with a more marked improvement of adipose tissue and systemic inflammation (delta changes at 6 and 12 weeks, respectively; Supplementary Table 10). Gene richness may therefore help to predict the efficacy of dietary intervention on inflammatory variables in overweight or obese individuals.

To further explore the effects of dietary intervention on gut microbial species we used a gene clustering procedure similar to the one described above for the comparison of LGC and HGC individuals. A set of 213,532 genes that varied significantly in frequency between different time points (Wilcoxon signed-rank test, $P < 0.05$) was first identified. To reduce the complexity of the data set, an entropy-filtering criterion was then applied, removing the genes present in only a few individuals (Supplementary Fig. 2). The remaining 58,109 genes were clustered by frequency covariance (Methods and Supplementary Fig. 3). Some 34,920 genes (60%) were grouped in 39 clusters larger than 100 genes (Supplementary Table 11); a large majority, 72%, were very compact, with a clustering coefficient >0.75 (ref. 14) (Supplementary Information, see cluster sheets for a more detailed description). Of the 39 clusters, 17 had ≥80% of their genes assigned to the same species and 19 to the same genus (the global distribution was 64% Firmicutes, 33% Bacteroidetes, and 3% Actinobacteria; Supplementary Table 11), confirming a species-specific clustering (Methods).

The abundance of the potential species represented by the 39 clusters was computed as the sum of the respective gene frequencies, and variations over time and correlations with bioclinical variables and food items were examined (Methods and Supplementary Tables 12 and 13). We observed that the abundance of 26 clusters varied significantly with time, indicating that a number of bacterial species can be modulated by nutritional intervention; the remaining 13 were not studied further. Only a few of our gene clusters decreased or showed a tendency to decrease during the calorie restriction phase, but one of those was assigned to *Eubacterium rectale* and another one to *Bifidobacterium* spp., in accordance with previous results[6] (Supplementary Table 11 and Supplementary Information).

The main trend after 6 weeks of energy-restricted diet was a significant increase of abundance of most gene clusters ($n = 15$), whereas the trend was opposite after 6 weeks of weight-maintenance diet, as the abundance of 14 species decreased. A total of five different patterns was observed, reflecting combinations of variation during the two periods (Supplementary Table 11), but the overall tendency was to return close to a baseline level by the end of the weight-maintenance phase (illustrated in Supplementary Informatino), suggesting a transient effect of dietary intervention on gut microbiota, as described previously[15]. Interestingly, for 8 of the 26 gene clusters that had a significantly lower abundance in the LGC than HGC individuals at baseline (Supplementary Table 14), the energy-restricted diet led to an increase of abundance in the LGC individuals, bringing them close to the level found in the HGC individuals (Fig. 3b); there was no significant abundance difference between the LGC and HGC individuals upon the stabilization phase. We conclude that the dietary intervention, in spite of its overall transient effect, may lead to more persistent changes of some gut microbial species.

Quantitative metagenomics analysis of the gut microbiome in 3 different samples for each of the 49 French (our study) and in 292 Danish subjects[11] revealed the existence of a high proportion of individuals (23–40%) with low microbial richness. In both study populations, a detailed clinical analysis indicated that these individuals show adiposity associated dyslipidaemia, higher insulin resistance and low-grade inflammation when compared to their higher-gene-diversity counterparts. This deleterious phenotype is known to be associated with increased risk of pre-diabetes, type 2 diabetes, hepatic and cardiovascular disorders as well as some forms of cancer[16–18]. In both study populations, abundance of many gut bacterial species in low-richness individuals was altered in a similar way relative to high richness individuals; this alteration can be accurately detected by combinations of only a few bacterial species. This indicates that simple diagnostic tests, based on our 'other genome' could be developed to identify individuals at a higher risk of obesity-associated co-morbidities. In the context of the current global epidemics of obesity and metabolic disorders, such tests could have a broad usefulness.

The concomitant improvement of gut microbial gene richness and bioclinical variables in LGC individuals by a dietary intervention suggests a possibility to advance from risk detection to risk alleviation, under the assumption that the less rich microbiota are also less healthy (see the accompanying paper[11]). Restoration of gene richness was not achieved fully by our short-term intervention, but seems to be a desirable goal, as decreased gene richness was found to be associated with a less efficient improvement of the inflammatory variables by dietary intervention. Interestingly, increased consumption of fruits and vegetable and thus higher fibre consumption before the intervention seemed to be associated with high bacterial richness. This finding, although exploratory in nature and requiring replication, supports a recently reported link between long-term dietary habits and the structure of gut microbiota[15] and suggests that a permanent change of microbiota may be achieved by appropriate diet. Development of a two-pronged approach, coupling early detection of an impending loss of gut bacterial richness to appropriate nutritional recommendations, which is yet to be established, may help to reach this goal and possibly contribute to diminish the risk of the obesity-linked co-morbidities; stratification by gene richness may have predictive value in respect to the efficacy of a dietary treatment and even guide its choice. However, low-grade inflammation, an important trait related to obesity but also common to many chronic diseases, seemed relatively refractory to dietary intervention in the LGC individuals, suggesting that specific therapeutic actions, aiming at restoring gut microbiota richness and equilibrium in obesity and altered metabolism, may need to be developed as well.

METHODS SUMMARY

Forty-nine obese or overweight subjects were recruited and subjected to a 6-week energy-restricted high-protein diet followed by a 6-week weight-maintenance diet. Bioclinical characteristics, physical activity scores and detailed qualitative and quantitative features of their food intake were obtained at baseline, 6 and 12 weeks (Methods). The clinical trial was registered at http://www.ClinicalTrials.gov under study number NCT01314690. The Ethical Committee of Hôtel Dieu Hospital approved the clinical study and all subjects provided written informed consent. Faecal samples were collected at each time point and analysed with the next generation sequencing SOLiD System. After read mapping a frequency table of microbial genes was obtained (Methods).

Two groups of patients with LGC and HGC were defined using the gene-richness distribution. Differences in terms of food, bioclinical variables and gene abundance were identified by standard statistical methods (Methods). Focusing on the dietary intervention and using a multi-criteria selection to narrow down the number of genes to a few thousands, gene clusters of co-varying microbial genes were constructed. These resulting gene clusters were then analysed for changes over time and correlations with bioclinical markers (Methods).

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

Received 12 April 2012; accepted 17 July 2013.

1. Mutch, D. M. & Clément, K. Unraveling the genetics of human obesity. *PLoS Genet.* **2,** e188 (2006).
2. Bäckhed, F. *et al.* The gut microbiota as an environmental factor that regulates fat storage. *Proc. Natl Acad. Sci. USA* **101,** 15718–15723 (2004).
3. Turnbaugh, P. J. *et al.* An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444,** 1027–1031 (2006).
4. Bäckhed, 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).
5. Ley, R. E., Turnbaugh, P. J., Klein, S. & Gordon, J. I. Microbial ecology: human gut microbes associated with obesity. *Nature* **444,** 1022–1023 (2006).
6. Duncan, S. H. *et al.* Reduced dietary intake of carbohydrates by obese subjects results in decreased concentrations of butyrate and butyrate-producing bacteria in feces. *Appl. Environ. Microbiol.* **73,** 1073–1078 (2007).
7. Riesenfeld, C. S., Schloss, P. D. & Handelsman, J. Metagenomics: genomic analysis of microbial communities. *Annu. Rev. Genet.* **38,** 525–552 (2004).
8. National Research Council. *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet* (The National Academies Press, 2007).

9. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464,** 59–65 (2010).
10. Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473,** 174–180 (2011).
11. Le Chatelier, E. *et al.* Richness of human gut microbiome correlates with metabolic markers. *Nature* http://dx.doi.org/10.1038/nature12506 (this issue).
12. Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457,** 480–484 (2009).
13. Claesson, M. J. *et al.* Gut microbiota composition correlates with diet and health in the elderly. *Nature* **488,** 178–184 (2012).
14. Wasserman, S. & Faust, K. *Social Network Analysis: Methods and Applications.* (Cambridge Univ. Press, 1994).
15. Wu, G. D. *et al.* Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334,** 105–108 (2011).
16. Ouchi, N., Parker, J. L., Lugus, J. J. & Walsh, K. Adipokines in inflammation and metabolic disease. *Nature Rev. Immunol.* **11,** 85–97 (2011).
17. Shoelson, S. E., Lee, J. & Goldfine, A. B. Inflammation and insulin resistance. *J. Clin. Invest.* **116,** 1793–1801 (2006).
18. Renehan, A. G., Tyson, M., Egger, M., Heller, R. F. & Zwahlen, M. Body-mass index and incidence of cancer: a systematic review and meta-analysis of prospective observational studies. *Lancet* **371,** 569–578 (2008).

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

Acknowledgements We are grateful to O. Pedersen (Univ. Copenhagen) for helpful comments on this manuscript and to the MetaHIT consortium for providing the gene profiles of the Danish subjects used to test the ROC models in advance of publication and the DNA samples sequenced on the SOLiD platform for comparison with the Illumina platform used in the accompanying manuscript. We thank C. Baudoin, P. Ancel and V. Pelloux who contributed to the clinical investigation study; S. Fellahi and J.-P. Bastard for analyses of inflammatory markers; D. Bonnefont-Rousselot and R. Bittar for help with the analysis of plasma lipid profile. This work was supported by Agence Nationale de la Recherche (ANR MICRO-Obes, ANR, Nutra2sens, *ANR-10-IAHU-05*), the Metagenopolis grant ANR-11-DPBS-0001, KOT-Ceprodi (Florence Massiera), Danone Research (Damien Paineau) and the associations Fondacoeur, and Louis-Bonduelle. Additional funding came from the European Commission FP7 grant HEALTH-F4-2007-201052 and METACARDIS.

Author Contributions S.D.E., J.D. and K.C. designed the study; S.D.E., J.D., K.C. and P.R. managed the study; K.C. and S.R. designed the clinical research; S.R. and L.C.K. conducted the clinical research and clinical data management; A.C., S.R. and L.C.K. conducted clinical and dietary data analysis; S.G. gave dietary counselling to the patients and carried out analysis of dietary data; F.L. prepared the DNA for sequencing; S.K. managed DNA sequencing, which B.Q. and N.G. carried out; N.P. and J.-M.B. established the sequence analysis pipeline; A.C., J.-D.Z., E.P., N.P., E.L.C., M.A., J.-M.B., S.K. and S.D.E. carried out microbial data analysis; A.C., K.C., L.C.K. and S.D.E. wrote the manuscript.

Author Information The raw solid read data for all samples has been deposited in the European Bioinformatics Institute (EBI) European Nucleotide Archive (ENA) under the accession number ERP003699. 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 S.D.E. (dusko.ehrlich@jouy.inra.fr) or K.C. (karine.clement@psl.aphp.fr).

ANR MicroObes consortium members

Hervé Blottière[1,2], Marion Leclerc[1], Catherine Juste[1], Tomas de Wouters[1], Patricia Lepage[1], Charlene Fouqueray[1], Arnaud Basdevant[3], Cornelieu Henegar[3], Cindy Godard[3], Marine Fondacci[3], Alili Rohia[3], Froogh Hajduch[3], Jean Weissenbach[4], Eric Pelletier[4], Denis Le Paslier[4], Jean-Pierre Gauchi[5], Jean-François Gibrat[6], Valentin Loux[6], Wilfrid Carré[6], Emmanuelle Maguin[1], Maarten van de Guchte[1], Alexandre Jamet[1], Fouad Boumezbeur[1] & Séverine Layec[1]

[1]INRA, Institut National de la Recherche Agronomique, UMR 1319 Micalis, Jouy en Josas 78350, France. [2]INRA, Institut National de la Recherche Agronomique, Metagenopolis, Jouy en Josas 78350, France. [3]Institute of Cardiometabolism and Nutrition, Assistance Publique-Hôpitaux de Paris, CRNH-Ile de France, Pitié-Salpêtrière, Paris 75013, France. [4]Commissariat à l'Energie Atomique, Genoscope, Evry 91000, France. [5]Institut National de la Recherche Agronomique, Mathématiques et Informatique Appliquées, Jouy en Josas 78350, France. [6]Institut National de la Recherche Agronomique, Mathématique, Informatique et Génome, Jouy en Josas 78350, France.

METHODS

Clinical investigation. Obese ($n = 38$) and overweight ($n = 11$) subjects, 8 men and 41 women, were recruited for a 12-week controlled dietary intervention at the Center of Research in Human Nutrition, Pitié-Salpêtrière Hospital, Paris, France. The subjects included in the study had no chronic pathologies except excess body weight. Their body weight was stable within 3 months before the study. None of the participants was undergoing chronic treatment or had been involved in weight-loss programs in the preceding 12 months. No antibiotics or drugs were taken within 2 months before or during the course of the study. The Ethical Committee of Hôtel Dieu Hospital approved the clinical study and subjects provided written informed consent. In the first 6-week phase, subjects consumed an energy-restricted high-protein diet (1,200 kilocalories (kcal) per day for women and 1,500 kcal for men: 35% proteins, 25% lipids, 44% carbohydrates) with low glycaemic index carbohydrates and enrichment with soluble fibres[19]. This phase was followed by a second 6-week body weight stabilization period with 20% increase in total energy intake, above their resting energy metabolic rate. At 0, 6 and 12 weeks, blood and faecal samples were collected and anthropometric measurements were performed. Subjects filled a 7-day dietary record and were interviewed by a registered dietician. On the visit day, the dietician checked the information and clarified any ambiguities regarding detail of food consumed. All records were analysed by the registered dietician using the computer software program PROFILE DOSSIER V3 (Audit Conseil en Informatique Médicale), which has a dietary database initially made up of 400 food items representative of the French diet as described previously[20]. A nutrient analysis was generated for each subject. Body composition was determined by dual-energy X-ray absorptiometry (DEXA). Blood samples were obtained after 12 h of fasting to measure total cholesterol, high-density lipoprotein (HDL) cholesterol, triglycerides, insulin, glucose, and inflammatory markers (hsCRP and interleukin 6 (IL-6)) as described previously[21]. Insulin resistance was estimated using HOMA-IR and Disse index scores[22,23]. Subcutaneous abdominal adipose tissue samples were obtained at all time points by needle biopsy from the periumbilical area under local anaesthesia (1% xylocaine) to measure the adipocytes diameter[24] and for immunohistochemical studies (HAM56+-stained macrophages in adipose tissue). Whole faecal samples were self-collected in sterile boxes and stored at $-20\,^{\circ}C$ within 4 h, sampled (200-mg aliquots) and then stored at $-80\,^{\circ}C$ until analysis. Paired Wilcoxon tests were performed to analyse changes in these variables between various time points ($P < 0.05$). P values were adjusted for multiple testing using the Benjamini–Hochberg procedure[25].

Metagenomic sequencing. Intestinal bacterial gene content of 49 obese and overweight individuals at 3 time-points (baseline, week 6 and week 12) was determined by high-throughput ABI SOLiD sequencing technology of total faecal DNA. An average of 76.5 million ± 36.5 million (mean ± s.d.) 35-base-long single reads were determined for each sample (a total of 393 Gb of sequence) (Supplementary Table 15). By using corona_lite (v4.0r2.0), an average of 24.8 million ± 14.3 million reads per individual were mapped on the reference catalogue of 3.3 million genes[9] with a maximum of 3 mismatches. Reads mapping at multiple positions were discarded and an average of 14.2 million ± 8.1 million uniquely mapped reads per individuals were retained for estimating the abundance of each reference gene by using METEOR[26] software. Abundance of each gene in an individual was normalized with METEOR by dividing the number of reads that uniquely mapped to a gene by its nucleotide length. After that, normalized gene abundances were transformed in frequencies by dividing them with the total number of uniquely mapped reads for a given sample. The resulting set of gene frequencies, termed as a microbial gene profile of an individual, was used for further analyses.

Comparison between SOLiD and Illumina sequencing technologies. Two primary short-read technologies currently exist for quantitative metagenomic analysis; SOLiD and Illumina. To validate data set correspondences and comparisons between results in this study and the accompanying paper[11], 24 samples from the Danish Inter99 cohort, previously sequenced on an Illumina GA platform, were also sequenced and analysed by SOLiD technology. Representative samples for cross-comparison included 14 females and 10 males, 15 obese and 9 lean, and 15 HGC and 9 LGC individuals. Hierarchical clustering demonstrated all samples self-clustered as technology-independent pairs, with the average Pearson correlation coefficient of 0.87 (computed upon log transformation) between the two technologies and increasing concordance associated with increased signal (Supplementary Fig. 4).

Gene-richness analysis. Gene richness was compared between subjects using the same number of mapped reads. Data were downsized to adjust for technical variability linked to different sequencing depths. This downsizing was performed at different levels by randomly selecting 4.5 or 7 million mapped reads for each sample and then computing the mean number of genes over 30 drawings (Supplementary Table 15). The 4.5-million-read downsizing allows keeping more than 90% of the individuals at each time point (required for the quantitative analysis of gene richness), but shrinks the data distribution (Supplementary Fig. 5). The 7-million-read downsizing was used for the analysis of the gene count distribution among the individuals and the enterotypes. The distribution of gene number obtained with the two downsizings is quite similar as shown by Spearman correlation ($\rho > 0.99$) (Supplementary Fig. 5).

Differentially abundant gene clusters between LGC and HGC. Two groups of patients with LGC and HGC were defined using the 480,000-gene threshold, consistent with the accompanying manuscript[11] (Fig. 1a, and main text). Genes significantly different in groups of individuals were identified by Mann–Whitney tests using a P-value threshold of <0.0001. They were clustered by an abundance-based binning strategy, using the covariance of their gene frequency profiles among the individuals of the cohort, as described in the accompanying paper[11]. Abundance of a given cluster in each individual was estimated as a mean abundance of 25 arbitrarily selected 'tracer' genes for each cluster; these values were close to those obtained by using all the genes of a cluster.

ROC analysis. The analyses were carried out to distinguish between HGC and LGC individuals by a combination of gene clusters. For each combination, only a single decision model was considered, computed as the sum of mean abundance of clusters with greater abundance in HGC than in LGC minus the sum of those with greater abundance in LGC than in HGC. As opposed to the infinite number of regression models, such models are finite and can be exhaustively explored. To select the best models, we used the cross-validated area under the ROC curve cross-validated AUC criterion[27] well adapted to classification models for binary outcome data.

Correlations between microbial gene clusters and clinical variables. Mann–Whitney tests were used to compare bioclinical variables, food items and gene clusters between LGC and HGC groups at each time point. Associations between quantitative basal gene richness and bioclinical or food variables, or differences (deltas) in bioclinical or food variables were investigated using linear models. For the associations between deltas of bioclinical parameters and deltas of gene richness, all pairs of deltas were computed (6 weeks–0 weeks, 12 weeks–6 weeks, 12 weeks–0 weeks). Linear mixed models were then fitted using all data. A P-value threshold of 0.05 was applied for statistical significance. Owing to the highly correlated bioclinical and food variables, adjustment for multiple testing is not really adequate, but the false discovery rates (Benjamini–Hochberg[25]) are given for information purposes in Supplementary Table 9.

Taxonomical annotation. The genes from clusters were mapped by BLASTN (BLAST 2.2.24, default parameters) 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 03.08.2012). Following taxonomical assignation parameters described by Arumugam[10], each gene was assigned with the taxonomy of the best-hit covering ≥80% of the gene length and according to the identity threshold for the taxonomic rank (≥65% for phylum, ≥ 85% for genus and ≥ 90% for species). To assess the taxonomy of clusters below these thresholds we used BLASTP against the non-redundant sequences databases available at NCBI. Based on the criterion of the homogeneity of the best hit taxonomic assignment (at least 80% of tracer genes from a cluster having the same taxonomic best hit assignment), 100% and 25% of the clusters could be assigned at a phylum and genus level, respectively (Supplementary Table 7).

Gene clusters affected by the dietary intervention. The analysis was carried out with genes with a potentially dietary linked signal. The first filtering step consisted in selecting the genes whose frequency was modulated significantly by the nutritional intervention during the dietary restriction or the stabilization period with a Wilcoxon signed-rank test ($P < 0.05$). A subset of these genes, with high Shannon entropy[28], was selected in a second filtering step. The entropy distribution of the filtered genes presented a bimodal distribution and the genes corresponding to the highest mode were selected using a threshold estimation on an approximation of its density function[29] (Supplementary Fig. 2). The genes with high entropy were mostly shared among individuals of the cohort. Genes with significantly similar frequency profiles (P divided by number of tests < 0.05) and high Spearman correlation coefficient ($\rho > 0.85$), were clustered in a way similar to the LGC–HGC clusters using single-linkage clustering (Supplementary Fig. 3). The 39 clusters with a size superior to 100 genes were kept for further analyses. The group abundance of each cluster was computed as the sum of the frequencies of its genes, and the data were log-transformed for parametric statistics.

Gene-cluster analysis. Gene clusters were analysed for changes over time and correlations with bioclinical markers using linear mixed models were adjusted for age and sex (Supplementary Tables 12 and 13). The highly correlated data induced P values distributions not adapted to standard procedures for multiple testing adjustments; nevertheless, we provide the false discovery rates using the Benjamini–Hochberg method in Supplementary Tables 12 and 14. All statistical analyses were performed using the R environment[30].

19. Rizkalla, S. W. *et al.* Differential effects of macronutrient content in 2 energy-restricted diets on cardiovascular risk factors and adipose tissue cell size in

moderately obese individuals: a randomized controlled trial. *Am. J. Clin. Nutr.* **95,** 49–63 (2012).
20. Bouché, C. *et al.* Five-week, low-glycemic index diet decreases total fat mass and improves plasma lipid profile in moderately overweight nondiabetic men. *Diabetes Care* **25,** 822–828 (2002).
21. Tordjman, J. *et al.* Structural and inflammatory heterogeneity in subcutaneous adipose tissue: Relation with liver histopathology in morbid obesity. *J. Hepatol.* **56,** 1152–1158 (2012).
22. Disse, E. *et al.* A lipid-parameter-based index for estimating insulin sensitivity and identifying insulin resistance in a healthy population. *Diabetes Metab.* **34,** 457–463 (2008).
23. Antuna-Puente, B. *et al.* Evaluation of insulin sensitivity with a new lipid-based index in non-diabetic postmenopausal overweight and obese women before and after a weight loss intervention. *Eur. J. Endocrinol.* **161,** 51–56 (2009).
24. Prat-Larquemin, L. *et al.* Adipose angiotensinogen secretion, blood pressure, and AGT M235T polymorphism in obese patients. *Obes. Res.* **12,** 556–561 (2004).
25. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B* **57,** 289–300 (1995).
26. Pons, N. *et al.* METEOR, a platform for quantitative metagenomic profiling of complex ecosystems. Journées Ouvertes en Biologie, Informatique et Mathématiques http://www.jobim2010.fr/sites/default/files/presentations/27Pons.pdf (2010).
27. Jiang, D., Huang, J. & Zhang, Y. The cross-validated AUC for MCP-logistic regression with high-dimensional data. *Stat. Methods Med. Res.* http://dx.doi.org/10.1177/0962280211428385 (28 November 2011).
28. Shannon, C. E. A mathematical theory of communication. *Bell Sys. Tech. J.* **27,** 379–423 (1995), 623–656 (1948).
29. Silverman, B. W. *Density Estimation for Statistics and Data Analysis* (Chapman and Hall, 1986).
30. R Development Core Team. *R: A Language and Environment for Statistical Computing* http://www.R-project.org (R Foundation for Statistical Computing, 2011).