LETTERS

A core gut microbiome in obese and lean twins

Peter J. Turnbaugh[1], Micah Hamady[3], Tanya Yatsunenko[1], Brandi L. Cantarel[5], Alexis Duncan[2], Ruth E. Ley[1], Mitchell L. Sogin[6], William J. Jones[7], Bruce A. Roe[8], Jason P. Affourtit[9], Michael Egholm[9], Bernard Henrissat[5], Andrew C. Heath[2], Rob Knight[4] & Jeffrey I. Gordon[1]

The human distal gut harbours a vast ensemble of microbes (the microbiota) that provide important metabolic capabilities, including the ability to extract energy from otherwise indigestible dietary polysaccharides[1–6]. Studies of a few unrelated, healthy adults have revealed substantial diversity in their gut communities, as measured by sequencing 16S rRNA genes[6–8], yet how this diversity relates to function and to the rest of the genes in the collective genomes of the microbiota (the gut microbiome) remains obscure. Studies of lean and obese mice suggest that the gut microbiota affects energy balance by influencing the efficiency of calorie harvest from the diet, and how this harvested energy is used and stored[3–5]. Here we characterize the faecal microbial communities of adult female monozygotic and dizygotic twin pairs concordant for leanness or obesity, and their mothers, to address how host genotype, environmental exposure and host adiposity influence the gut microbiome. Analysis of 154 individuals yielded 9,920 near full-length and 1,937,461 partial bacterial 16S rRNA sequences, plus 2.14 gigabases from their microbiomes. The results reveal that the human gut microbiome is shared among family members, but that each person's gut microbial community varies in the specific bacterial lineages present, with a comparable degree of co-variation between adult monozygotic and dizygotic twin pairs. However, there was a wide array of shared microbial genes among sampled individuals, comprising an extensive, identifiable 'core microbiome' at the gene, rather than at the organismal lineage, level. Obesity is associated with phylum-level changes in the microbiota, reduced bacterial diversity and altered representation of bacterial genes and metabolic pathways. These results demonstrate that a diversity of organismal assemblages can nonetheless yield a core microbiome at a functional level, and that deviations from this core are associated with different physiological states (obese compared with lean).

We characterized gut microbial communities in 31 monozygotic twin pairs, 23 dizygotic twin pairs and, where available, their mothers ($n = 46$) (Supplementary Tables 1–5). Monozygotic and dizygotic co-twins and parent–offspring pairs provided an attractive model for assessing the impact of genotype and shared early environmental exposures on the gut microbiome. Moreover, genetically 'identical'[9] monozygotic twin pairs gain weight in response to overfeeding in a more reproducible way than unrelated individuals[10] and are more concordant for body mass index (BMI) than dizygotic twin pairs[11].

Twin pairs who had been enrolled in the Missouri Adolescent Female Twin Study (MOAFTS[12]) were recruited for this study (mean period of enrolment in MOAFTS, 11.7 ± 1.2 years; range, 4.4–13.0 years). Twins were 21–32 years old, of European or African ancestry, and were generally concordant for obesity (BMI $\geq 30\,kg\,m^{-2}$) or leanness (BMI = 18.5–24.9 $kg\,m^{-2}$) (one twin pair was lean/overweight (overweight defined as BMI ≥ 25 and < 30) and six pairs were overweight/obese). They had not taken antibiotics for at least 5.49 ± 0.09 months. Each participant completed a detailed medical, lifestyle and dietary questionnaire: study enrolees were broadly representative of the overall Missouri population for BMI, parity, education and marital status (see Supplementary Results). Although all were born in Missouri, they currently live throughout the USA: 29% live in the same house, but some live more than 800 km apart. Because faecal samples are readily attainable and representative of interpersonal differences in gut microbial ecology[7], they were collected from each individual and frozen immediately. The collection procedure was repeated again with an average interval between sampling of 57 ± 4 days.

To characterize the bacterial lineages present in the faecal microbiotas of these 154 individuals, we performed 16S rRNA sequencing, targeting the full-length gene with an ABI 3730xl capillary sequencer. Additionally, we performed multiplex pyrosequencing with a 454 FLX instrument to survey the gene's V2 variable region[13] and its V6 hypervariable region[14] (Supplementary Tables 1–3).

Complementary phylogenetic and taxon-based methods were used to compare 16S rRNA sequences among faecal communities (see Methods). No matter which region of the gene was examined, individuals from the same family (a twin and her co-twin, or twins and their mother) had a more similar bacterial community structure than unrelated individuals (Fig. 1a and Supplementary Fig. 1a, b), and shared significantly more species-level phylotypes (16S rRNA sequences with ≥97% identity comprise each phylotype) ($G = 55.2$, $P < 10^{-12}$ (V2); $G = 12.3$, $P < 0.001$ (V6); $G = 11.3$, $P < 0.001$ (full-length)). No significant correlation was seen between the degree of physical separation of family members' current homes and the degree of similarity between their microbial communities (defined by UniFrac[15]). The observed familial similarity was not due to an indirect effect of the physiological states of obesity versus leanness; similar results were observed after stratifying twin pairs and their mothers by BMI category (concordant lean or concordant obese individuals; Supplementary Fig. 2). Surprisingly, there was no significant difference in the degree of similarity in the gut microbiotas of adult monozygotic compared with dizygotic twin pairs (Fig. 1a). However, we could not assess whether monozygotic and dizygotic twin pairs had different degrees of similarities at earlier stages of their lives.

Multiplex pyrosequencing of V2 and V6 amplicons allowed higher levels of coverage compared with what was feasible using Sanger sequencing, reaching on average 3,984 ± 232 (V2) and 24,786 ± 1,403 (V6) sequences per sample. To control for differences

[1]Center for Genome Sciences. [2]Department of Psychiatry, Washington University School of Medicine, St Louis, Missouri 63108, USA. [3]Department of Computer Science. [4]Department of Chemistry and Biochemistry, University of Colorado, Boulder, Colorado 80309, USA. [5]CNRS, UMR6098, Marseille, France. [6]Josephine Bay Paul Center, Marine Biological Laboratory, Woods Hole, Massachusetts 02543, USA. [7]Environmental Genomics Core Facility, University of South Carolina, Columbia, South Carolina 29208, USA. [8]Department of Chemistry and Biochemistry and the Advanced Center for Genome Technology, University of Oklahoma, Norman, Oklahoma 73019, USA. [9]454 Life Sciences, Branford, Connecticut 06405, USA.

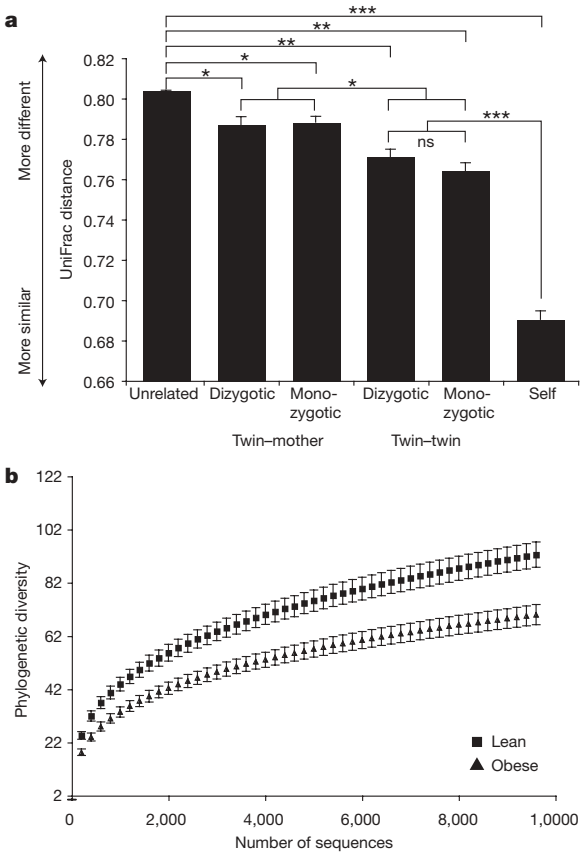

Figure 1 | 16S rRNA gene surveys reveal familial similarity and reduced diversity of the gut microbiota in obese individuals. a, Average unweighted UniFrac distance (a measure of differences in bacterial community structure) between individuals over time (self), twin pairs, twins and their mother, and unrelated individuals (1,000 sequences per V2 data set; Student's *t*-test with Monte Carlo; $*P < 10^{-5}$; $**P < 10^{-14}$; $***P < 10^{-41}$; mean ± s.e.m.). **b**, Phylogenetic diversity curves for the microbiota of lean and obese individuals (based on 1–10,000 sequences per V6 data set; mean ± 95% confidence intervals shown).

in coverage, all analyses were performed on an equal number of randomly selected sequences (200 full-length, 1,000 V2 and 10,000 V6). At this level of coverage, there was little overlap between the sampled faecal communities. Moreover, the number of 16S rRNA gene sequences belonging to each phylotype varied greatly between faecal microbiotas (Supplementary Tables 6–8).

Because this apparent lack of overlap could reflect the level of coverage (Supplementary Tables 1–3), we subsequently searched all hosts for bacterial phylotypes present at high abundance using a sampling model based on a combination of standard Poisson and binomial sampling statistics. The analysis allowed us to conclude that no phylotype was present at more than about 0.5% abundance in all of the samples in this study (see Supplementary Results). Finally, we sub-sampled our data set by randomly selecting 50–3,000 sequences per sample; again, no phylotypes were detectable in all individuals sampled within this range of coverage (Supplementary Fig. 3).

Samples taken from the same individual at the initial collection point and 57 ± 4 days later were consistent with respect to the specific phylotypes found (Supplementary Figs 4 and 5), but showed variations in relative abundance of the major gut bacterial phyla (Supplementary Fig. 6). There was no significant association between UniFrac distance and the time between sample collections. Overall, faecal samples from the same individual were much more similar to one another than samples from family members or unrelated individuals (Fig. 1a), demonstrating that short-term temporal changes in community structure within an individual are minor compared with inter-personal differences.

Analysis of 16S rRNA data sets produced by the three PCR-based methods, plus shotgun sequencing of community DNA (see below), revealed a lower proportion of Bacteroidetes and a higher proportion of Actinobacteria in obese compared with lean individuals of both ancestries (Supplementary Table 9). Combining the individual *P* values across these independent analyses using Fisher's method disclosed significantly fewer Bacteroidetes ($P = 0.003$), more Actinobacteria ($P = 0.002$) but no significant difference in Firmicutes ($P = 0.09$). These findings agree with previous work showing comparable differences in both taxa in mice[2] and a progressive increase in the representation of Bacteroidetes when 12 unrelated, obese humans lost weight after being placed on one of two reduced-calorie diets[6].

Across all methods, obesity was associated with a significant decrease in the level of diversity (Fig. 1b and Supplementary Fig. 1c–f). This reduced diversity suggests an analogy: the obese gut microbiota is not like a rainforest or reef, which are adapted to high energy flux and are highly diverse; rather, it may be more like a fertilizer runoff where a reduced-diversity microbial community blooms with abnormal energy input[16].

We subsequently characterized the microbial lineage and gene content of the faecal microbiomes of 18 individuals representing six of the families (three lean and three obese European ancestry monozygotic twin pairs and their mothers) through shotgun pyrosequencing (Supplementary Tables 4 and 5) and BLASTX comparisons against several databases (KEGG[17] (version 44) and STRING[18]) plus a custom database of 44 reference human gut microbial genomes (Supplementary Figs 7–10 and Supplementary Results). Our analysis parameters were validated using control data sets comprising randomly fragmented microbial genes with annotations in the KEGG database[17] (Supplementary Fig. 11 and Supplementary Methods). We also tested how technical advances that produce longer reads might improve these assignments by sequencing faecal community samples from one twin pair using Titanium pyrosequencing methods (average read length of 341 ± 134 nucleotides (s.d.) versus 208 ± 68 nucleotides for the standard FLX method). Supplementary Fig. 12 shows that the frequency and quality of sequence assignments is improved as read length increases from 200 to 350 nucleotides.

The 18 microbiomes were searched to identify sequences matching domains from experimentally validated carbohydrate-active enzymes (CAZymes). Sequences matching 156 total CAZy families were found within at least one human gut microbiome, including 77 glycoside hydrolase, 21 carbohydrate-binding module, 35 glycosyltransferase, 12 polysaccharide lyase and 11 carbohydrate-esterase families (Supplementary Table 10). On average, 2.62 ± 0.13% of the sequences in the gut microbiome could be assigned to CAZymes (a total of 217,615 sequences), a percentage that is greater than the most abundant KEGG pathway ('Transporters'; 1.20 ± 0.06% of the filtered sequences generated from each sample) and indicative of the abundant and diverse set of microbial genes directed towards accessing a wide range of polysaccharides.

Category-based clustering of the functions from each microbiome was performed using principal components analysis (PCA) and hierarchical clustering[19]. Two distinct clusters of gut microbiomes were identified based on metabolic profile, corresponding to samples with an increased abundance of Firmicutes and Actinobacteria, and samples with a high abundance of Bacteroidetes (Fig. 2a). A linear regression of the first principal component (PC1, explaining 20% of the functional variance) and the relative abundance of the Bacteroidetes showed a highly significant correlation ($R^2 = 0.96$, $P < 10^{-12}$; Fig. 2b). Functional profiles stabilized within each individual's microbiome after 20,000 sequences had been accumulated (Supplementary Fig. 13). Family members had more similar profiles than unrelated individuals (Fig. 2c), suggesting that shared bacterial community structure ('who's there' based on 16S rRNA analyses) also translates into shared community-wide relative abundance of metabolic pathways. Accordingly, a direct comparison of functional

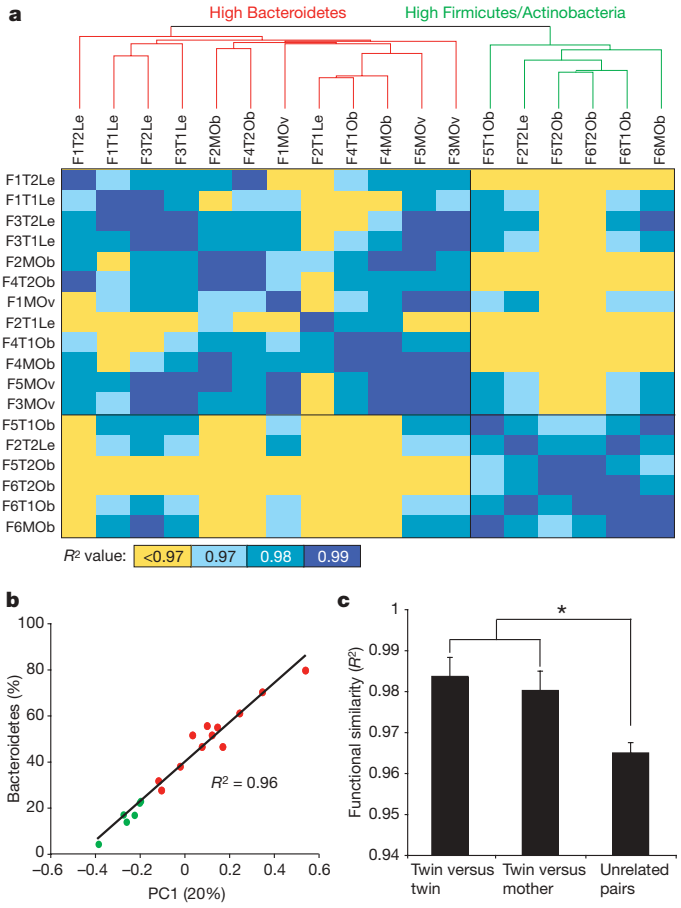

Figure 2 | Metabolic-pathway-based clustering and analysis of the human gut microbiome of monozygotic twins. a, Clustering of functional profiles based on the relative abundance of KEGG metabolic pathways. All pairwise comparisons were made of the profiles by calculating each R^2 value. Sample identifier nomenclature: family number, twin number or mother, and BMI category (Le, lean; Ov, overweight; Ob, obese; for example, F1T1Le stands for family 1, twin 1, lean). **b**, The relative abundance of Bacteroidetes as a function of the first principal component derived from an analysis of KEGG metabolic profiles. **c**, Comparisons of functional similarity between twin pairs, between twins and their mother, and between unrelated individuals. Asterisk indicates significant differences (Student's *t*-test with Monte Carlo; $P < 0.01$; mean ± s.e.m.).

and taxonomic similarity (see Supplementary Methods) disclosed a significant association: individuals with similar taxonomic profiles also share similar metabolic profiles ($P < 0.001$; Mantel test).

Functional clustering of phylum-wide sequence bins representing microbiome reads assigned to 23 human gut Firmicutes and 14 Bacteroidetes reference genomes showed discrete clustering by phylum (Supplementary Figs 14a and 15). Bootstrap analyses of the relative abundance of metabolic pathways in the microbiome-derived Firmicutes and Bacteroidetes sequence bins disclosed 26 pathways with a significantly different relative abundance (Supplementary Fig. 14a). The Bacteroidetes bins were enriched for several carbohydrate metabolism pathways, whereas the Firmicutes bins were enriched for transport systems. This finding is consistent with our CAZyme analysis, which revealed a significantly higher relative abundance of glycoside hydrolases, carbohydrate-binding modules, glycosyltransferases, polysaccharide lyases and carbohydrate esterases in the Bacteroidetes sequence bins (Supplementary Fig. 14b).

One of the major goals of the International Human Microbiome Project(s) is to determine whether there is an identifiable 'core microbiome' of shared organisms, genes or functional capabilities found in a given body habitat of all or the vast majority of humans[1]. Although all of the 18 gut microbiomes surveyed showed a high level of β-diversity with respect to the relative abundance of bacterial phyla (Fig. 3a), analysis of the relative abundance of broad functional categories of genes and metabolic pathways (KEGG) revealed a generally consistent pattern regardless of the sample surveyed (Fig. 3b and Supplementary Table 11): the pattern is also consistent with results we obtained from a meta-analysis of previously published gut microbiome data sets from nine adults[20,21] (Supplementary Fig. 16). This consistency is not simply due to the broad level of these annotations, as a similar analysis of Bacteroidetes and Firmicutes reference genomes revealed substantial variation in the relative abundance of each category (see Supplementary Fig. 17). Furthermore, pairwise comparisons of metabolic profiles obtained from the 18 microbiomes in this study revealed an average value of R^2 of 0.97 ± 0.002 (Fig. 2a), indicating a high level of functional similarity.

Overall functional diversity was compared using the Shannon index[22], a measurement that combines diversity (the number of different metabolic pathways) and evenness (the relative abundance of each pathway). The human gut microbiomes surveyed had a stable and high Shannon index value (4.63 ± 0.01), close to the maximum possible level of functional diversity (5.54; see Supplementary Methods). Despite the presence of a small number of abundant metabolic pathways (listed in Supplementary Table 11), the overall functional profile of each gut microbiome is quite even (Shannon evenness of 0.84 ± 0.001 on a scale of 0–1), demonstrating that most metabolic pathways are found at a similar level of abundance. Interestingly, the level of functional diversity in each microbiome was significantly linked to the relative abundance of the Bacteroidetes ($R^2 = 0.81$, $P < 10^{-6}$); microbiomes enriched for Firmicutes/Actinobacteria had a lower level of functional diversity. This observation is consistent with an analysis of simulated metagenomic reads generated from each of 36 Bacteroidetes and Firmicutes genomes (Supplementary Fig. 18): on average, the Bacteroidetes genomes have a significantly higher level of both functional diversity and evenness (Mann–Whitney *U*-test, $P < 0.01$).

At a finer level, 26–53% of 'enzyme'-level functional groups (KEGG/CAZy/STRING) were shared across all 18 microbiomes, whereas 8–22% of the groups were unique to a single microbiome (Supplementary Fig. 19a–c). The 'core' functional groups present in all microbiomes were also highly abundant, representing 93–98% of the total sequences. Given the higher relative abundance of these 'core' groups, more than 95% were found after 26.11 ± 2.02 megabases of sequence were collected from a given microbiome, whereas the 'variable' groups continued to increase substantially with each additional megabase of sequence. Of course, any estimate of the total size of the core microbiome will depend on sequencing effort, especially for

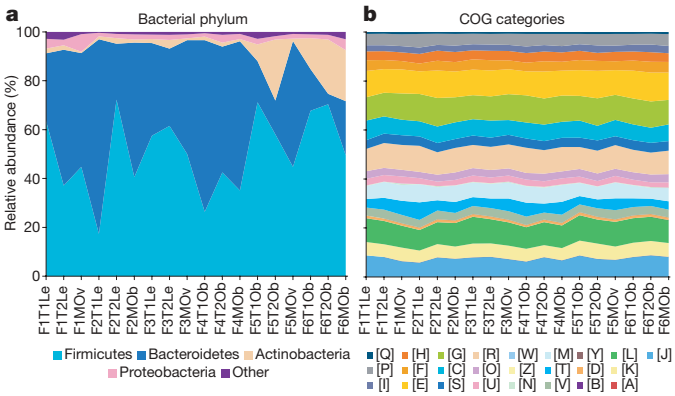

Figure 3 | Comparison of taxonomic and functional variations in the human gut microbiome. a, Relative abundance of major phyla across 18 faecal microbiomes from monozygotic twins and their mothers, based on BLASTX comparisons of microbiomes and the National Center for Biotechnology Information non-redundant database. **b**, Relative abundance of categories of genes across each sampled gut microbiome (letters correspond to categories in the COG database).

functional groups found at a low abundance. On average, our survey achieved more than 450,000 sequences per faecal sample, which, assuming an even distribution, would allow us to sample groups found at a relative abundance of 10^{-4}. To estimate the total size of the core microbiome based on the 18 individuals, we randomly subsampled each microbiome in 1,000 sequence intervals (Supplementary Fig. 19d). Based on this analysis, the core microbiome is approaching a total of 2,142 total orthologous groups (one site binding (hyperbola) curve fit, $R^2 = 0.9966$), indicating that we identified 93% of functional groups (defined by STRING) found within the core microbiome of the 18 individuals surveyed. Of these core groups, 71% (CAZy), 64% (KEGG) and 56% (STRING) were also found in the nine previously published, but much lower coverage, data sets generated by capillary sequencing of adult faecal DNA[20,21] (average of 78,413 ± 2,044 bidirectional reads per sample; see Supplementary Methods).

Metabolic reconstructions of the 'core' microbiome revealed significant enrichment for several expected functional categories, including those involved in transcription and translation (Fig. 4). Metabolic profile-based clustering indicated that the representation of 'core' functional groups was highly consistent across samples (Supplementary Fig. 20), and included several pathways that are likely important for life in the gut, such as those for carbohydrate and amino-acid metabolism (for example, fructose/mannose metabolism, amino-sugar metabolism and N-glycan degradation). Variably represented pathways and categories include cell motility (only a subset of Firmicutes produce flagella), secretion systems and membrane transport (for example, phosphotransferase systems involved in the import of nutrients, including sugars; Fig. 4 and Supplementary Fig. 20).

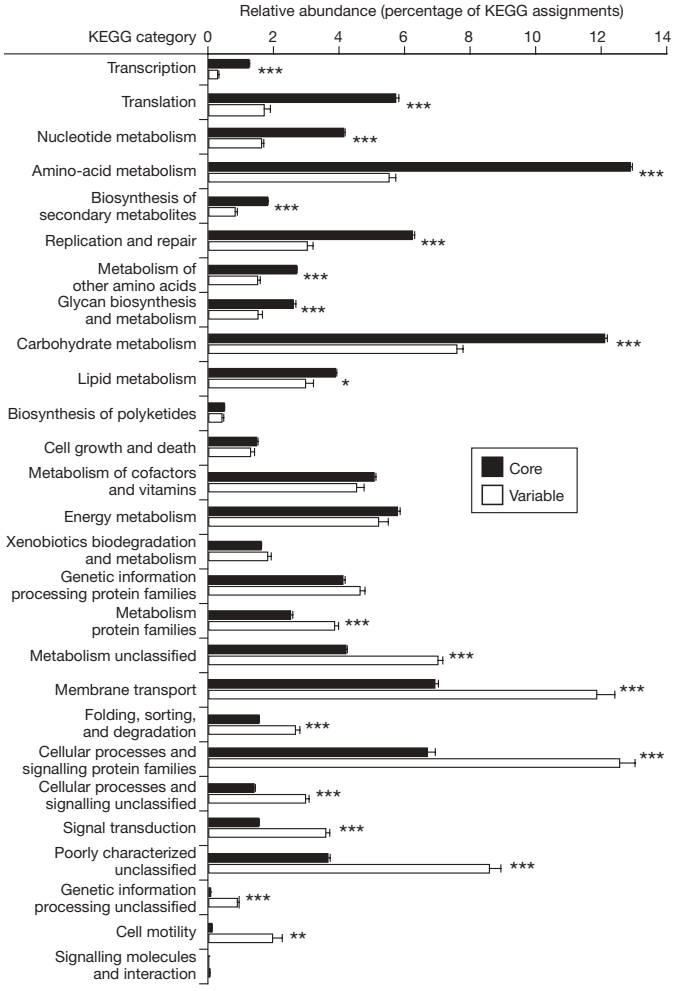

Figure 4 | KEGG categories enriched or depleted in the core versus variable components of the gut microbiome. Sequences from each of the 18 faecal microbiomes were binned into the 'core' or 'variable' microbiome based on the co-occurrence of KEGG orthologous groups (core groups were found in all 18 microbiomes whereas variable groups were present in fewer (<18) microbiomes; see Supplementary Fig. 19a). Asterisks indicate significant differences (Student's *t*-test, $*P < 0.05$, $**P < 0.001$, $***P < 10^{-5}$; mean ± s.e.m.).

The distribution of CAZy glycoside hydrolase and glycosyltransferase families was compared between each pair of microbiomes (see Supplementary Table 10 for CAZy families with a relative abundance greater than 1%). This analysis revealed that all individuals had a similar profile of glycosyltransferases ($R^2 = 0.96 \pm 0.003$), whereas the profiles of glycoside hydrolases were significantly more variable, even between family members ($R^2 = 0.80 \pm 0.01$; $P < 10^{-30}$, paired Student's *t*-test). This suggests that the number and spectrum of glycoside hydrolases is affected by 'external' factors such as diet more than the glycosyltransferases.

To identify metabolic pathways associated with obesity, only non-core associated (variable) functional groups were included in a comparison of the gut microbiomes of lean versus obese twin pairs. A bootstrap analysis[23] was used to identify metabolic pathways that were enriched or depleted in the variable obese gut microbiome. For example, similar to a mouse model of diet-induced obesity[4], the obese human gut microbiome was enriched for phosphotransferase systems involved in microbial processing of carbohydrates (Supplementary Table 12). All gut microbiome sequences were compared with the custom database of 44 human gut genomes: an odds ratio analysis revealed 383 genes that were significantly different between the obese and lean gut microbiome (q value < 0.05; 273 enriched and 110 depleted in the obese microbiome; Supplementary Tables 13 and 14). By contrast, only 49 genes were consistently enriched or depleted between all twin pairs (see Supplementary Methods).

These obesity-associated genes were representative of the taxonomic differences described above: 75% of the obesity-enriched genes were from Actinobacteria (compared with 0% of lean-enriched genes; the other 25% are from Firmicutes) whereas 42% of the lean-enriched genes were from Bacteroidetes (compared with 0% of the obesity-enriched genes). Their functional annotation indicated that many are involved in carbohydrate, lipid and amino-acid metabolism (Supplementary Tables 13 and 14). Together, they comprise an initial set of microbial biomarkers of the obese gut microbiome.

Our finding that the gut microbial community structures of adult monozygotic twin pairs had a degree of similarity that was comparable to that of dizygotic twin pairs, and only slightly more similar than that of their mothers, is consistent with an earlier fingerprinting study of adult twins[24], and with a recent microarray-based analysis, which revealed that gut community assembly during the first year of life followed a more similar pattern in a pair of dizygotic twins than 12 unrelated infants[25]. Intriguingly, another fingerprinting study of monozygotic and dizygotic twins in childhood showed a slightly reduced similarity profile in dizygotic twins[26]. Thus, comprehensive time-course studies, comparing monozygotic and dizygotic twin pairs from birth through adulthood, as well as intergenerational analyses of their families' microbiotas, will be key to determining the relative contributions of host genotype and environmental exposures to (gut) microbial ecology.

The hypothesis that there is a core human gut microbiome, definable by a set of abundant microbial organismal lineages that we all share, may be incorrect: by adulthood, no single bacterial phylotype was detectable at an abundant frequency in the guts of all 154 sampled humans. Instead, it appears that a core gut microbiome exists at the level of shared genes, including an important component involved in various metabolic functions. This conservation suggests a high degree of redundancy in the gut microbiome and supports an ecological view of each individual as an 'island' inhabited by unique

collections of microbial phylotypes: as in actual islands, different species assemblages converge on shared core functions provided by distinctive components. Our findings raise the question of how core functionality is assembled in this body habitat. Understanding the underlying principles should provide insights about microbial adaptation to, and mutualistic community assembly within, a wide range of environments.

METHODS SUMMARY

Faecal samples were collected from each individual. Community DNA was prepared and used for pyrosequencing (454 Life Sciences), as well as for PCR and sequencing of bacterial 16S rRNA genes. Shotgun reads were mapped to reference genomes using the National Center for Biotechnology Information 'non-redundant' database, KEGG[17], STRING[18], CAZy (http://www.cazy.org/) and a 44-member human-gut microbial genome database. Metabolic reconstructions were performed based on CAZy, KEGG and STRING annotations. The relative abundance of KEGG metabolic pathways is referred to as a 'metabolic profile'.

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

Received 29 June; accepted 14 October 2008.
Published online 30 November 2008.

1. Turnbaugh, P. J. *et al.* The human microbiome project. *Nature* **449,** 804–810 (2007).
2. Ley, R. E. *et al.* Obesity alters gut microbial ecology. *Proc. Natl Acad. Sci. USA* **102,** 11070–11075 (2005).
3. Turnbaugh, P. J. *et al.* An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444,** 1027–1031 (2006).
4. Turnbaugh, P. J., Bäckhed, 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).
5. Bäckhed, F. *et al.* The gut microbiota as an environmental factor that regulates fat storage. *Proc. Natl Acad. Sci. USA* **101,** 15718–15723 (2004).
6. Ley, R. E., Turnbaugh, P. J., Klein, S. & Gordon, J. I. Human gut microbes associated with obesity. *Nature* **444,** 1022–1023 (2006).
7. Eckburg, P. B. *et al.* Diversity of the human intestinal microbial flora. *Science* **308,** 1635–1638 (2005).
8. Frank, D. N. *et al.* Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc. Natl Acad. Sci. USA* **104,** 13780–13785 (2007).
9. Bruder, C. E. *et al.* Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. *Am. J. Hum. Genet.* **82,** 763–771 (2008).
10. Bouchard, C. *et al.* The response to long-term overfeeding in identical twins. *N. Engl. J. Med.* **322,** 1477–1482 (1990).
11. 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).
12. Heath, A. C. *et al.* Ascertainment of a mid-western US female adolescent twin cohort for alcohol studies: assessment of sample representativeness using birth record data. *Twin Res.* **5,** 107–112 (2002).
13. Hamady, M., Walker, J. J., Harris, J. K., Gold, N. J. & Knight, R. Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nature Methods* **5,** 235–237 (2008).
14. Sogin, M. L. *et al.* Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc. Natl Acad. Sci. USA* **103,** 12115–12120 (2006).
15. Lozupone, C., Hamady, M. & Knight, R. UniFrac-an online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics* **7,** 371 (2006).
16. Watson, S. B., McCauley, E. & Downing, J. A. Patterns in phytoplankton taxonomic composition across temperate lakes of differing nutrient status. *Limnol. Oceanogr.* **42,** 487–495 (1997).
17. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32,** D277–D280 (2004).
18. von Mering, C. *et al.* STRING 7-recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.* **35,** D358–D362 (2007).
19. de Hoon, M. J., Imoto, S., Nolan, J. & Miyano, S. Open source clustering software. *Bioinformatics* **20,** 1453–1454 (2004).
20. Gill, S. R. *et al.* Metagenomic analysis of the human distal gut microbiome. *Science* **312,** 1355–1359 (2006).
21. Kurokawa, K. *et al.* Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res.* **14,** 169–181 (2007).
22. Dinsdale, E. A. *et al.* Functional metagenomic profiling of nine biomes. *Nature* **452,** 629–632 (2008).
23. Rodriguez-Brito, B., Rohwer, F. & Edwards, R. A. An application of statistics to comparative metagenomics. *BMC Bioinformatics* **7,** 162 (2006).
24. Zoetandal, E. G., Akkermans, A. D. L., Akkermans-van Vliet, W. M., de Visser, J. A. & de Vos, W. M. The host genotype affects the bacterial community in the human gastrointestinal tract. *Microb. Ecol. Health Dis.* **13,** 129–134 (2001).
25. Palmer, C., Bik, E. M., Digiulio, D. B., Relman, D. A. & Brown, P. O. Development of the human infant intenstinal microbiota. *PLoS Biol.* **5,** e177 (2007).
26. Stewart, J. A., Chadwick, V. S. & Murray, A. Investigations into the influence of host genetics on the predominant eubacteria in the faecal microflora of children. *J. Med. Microbiol.* **54,** 1239–1242 (2005).

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

Acknowledgements We thank: S. Wagoner and J. Manchester for technical support; S. Marion and D. Hopper for recruitment of participants and sample collection; A. Goodman, B. Muegge, and M. Mahowald for suggestions; S. Huse (Marine Biological Laboratory), F. Niazi and S. Attiya (454 Life Sciences), C. Markovic, L. Fulton, B. Fulton, E. Mardis and R. Wilson (Washington University Genome Sequencing Center) and S. Macmil, G. Wiley, C. Qu, and P. Wang (University of Oklahoma) for their assistance with sequencing; and P. M. Coutinho (Université de Provence, France) for help with the CAZy analysis. Deep draft assemblies of reference gut genomes were generated as part of a National Human Genome Research Institute (NHGRI)-sponsored human gut microbiome initiative (http://genome.wustl.edu/pub/organism/Microbes/Human_Gut_Microbiome/). This work was supported in part by the National Institutes of Health (DK78669/ES012742/AA09022/HD049024), the National Science Foundation (OCE0430724), the W.M. Keck Foundation, and the Crohn's and Colitis Foundation of America.

Author Contributions P.J.T., A.C.H., R.K. and J.I.G. designed the experiments. P.J.T., T.Y., A.D., R.E.L., M.L.S., W.J.J., B.A.R., J.P.A. and M.E. generated the data. P.J.T., M.H., M.L.S., B.L.C., A.D., B.H., A.C.H., R.K. and J.I.G. analysed the data. P.J.T., A.C.H., R.K. and J.I.G. wrote the manuscript with input from the other members of the team.

Author Information This Whole Genome Shotgun project is deposited in DDBJ/EMBL/GenBank under accession number 32089. 454 pyrosequencing reads are deposited in the NCBI Short Read Archive. Nearly full-length 16S rRNA gene sequences are deposited in GenBank under accession numbers FJ362604–FJ372382. Annotated sequences are also available in MG-RAST (http://metagenomics.nmpdr.org/). 454-generated 16S rRNA sequences with sample identifiers are also available at http://gordonlab.wustl.edu/SuppData.html. Correspondence and requests for materials should be addressed to J.I.G. (jgordon@wustl.edu).

METHODS

Community DNA preparation. Faecal samples were frozen immediately after they were produced. De-identified samples were stored at −80 °C before processing. Ten to twenty grams of each sample was pulverized in liquid nitrogen with a mortar and pestle. An aliquot (approximately 500 mg) of each sample was then suspended, while frozen, in a solution containing 500 µl of extraction buffer (200 mM Tris (pH 8.0), 200 mM NaCl, 20 mM EDTA), 210 µl of 20% SDS, 500 µl of a mixture of phenol:chloroform:isoamyl alcohol (25:24:1, pH 7.9), and 500 µl of a slurry of 0.1-mm diameter zirconia/silica beads (BioSpec Products). Microbial cells were subsequently lysed by mechanical disruption with a bead beater (BioSpec Products) set on high for 2 min at room temperature, followed by extraction with phenol:chloroform:isoamyl alcohol, and precipitation with isopropanol. DNA obtained from three separate aliquots of each faecal sample were pooled (≥200 µg DNA) and used for pyrosequencing (see below).

16S rRNA gene-sequence-based surveys. Complementary phylogenetic- and taxon-based methods were used to compare 16S rRNA sequences among faecal communities. Phylogenetic clustering with UniFrac[15] is based on the principle that communities can be compared in terms of their shared evolutionary history, as measured by the degree to which they share branch length on a phylogenetic tree. We complemented this approach with taxon-based methods[27], which disregard some of the information contained in the phylogenetic tree of the taxa in question, but have the advantage that specific taxa unique to, or shared among, groups of samples can be identified (for example, those from lean or obese individuals). Before both types of analysis, we grouped 16S rRNA gene sequences into operational taxonomic units (OTUs/phylotypes) using both cd-hit[28] and the furthest-neighbour-like algorithm, with a sequence identity threshold of 97%, which is commonly used to define 'species'-level phylotypes. Taxonomy was assigned using the best-BLAST-hit against Greengenes[29] (E value cutoff of 10^{-10}, minimum 88% coverage, 88% identity) and the Hugenholtz taxonomy (downloaded from http://greengenes.lbl.gov/Download/Sequence_Data/Greengenes_format/ on 12 May 2008, excluding sequences annotated as chimaeric).

Selection of operational taxonomic units. 16S rRNA gene-derived pyrosequencing data were pre-processed to remove sequences with low-quality scores, sequences with ambiguous characters or sequences outside the length bounds (V6 < 50 nucleotides, V2 < 200 nucleotides), and binned according to sample-specific barcode (see, for example, ref. 13). Similar sequences were identified using Megablast[30] and cd-hit, with the following parameters: E value 10^{-10} (Megablast only); minimum coverage 99%; minimum pairwise identity 97%. Candidate OTUs were identified as sets of sequences connected to each other at this level using a maximum of 4,000 hits per sequence. Each candidate OTU was considered valid if the average density of connection was above threshold; otherwise, it was broken up into smaller connected components[27].

Tree building and UniFrac clustering for PCA analysis. A relaxed neighbour-joining tree was built from one representative sequence per OTU using Clearcut[31], employing the Kimura correction (the PH Lane mask was applied to V2 and full-length data), but otherwise with default comparisons. Unweighted UniFrac[15] was run using the resulting tree. PCA was performed on the resulting matrix of distances between each pair of samples. To determine if the UniFrac distances were on average significantly different for pairs of samples (that is, between twin pairs, between twins and their mother, or between unrelated individuals), we performed a t-test on the UniFrac distance matrix, and generated a P value for the t-statistic by permutation of the rows and columns as in the Mantel test, regenerating the t-statistic for 1,000 random samples, and using the distribution to obtain an empirical P value.

Rarefaction and phylogenetic diversity measurements. To determine which individuals had the most diverse communities of gut bacteria, rarefaction plots and phylogenetic diversity measurements, as described by Faith[32], were made for each sample. Phylogenetic diversity is the total amount of branch length in a phylogenetic tree constructed from the combined 16S rRNA data sets, leading to the sequences in a given sample. To account for differences in sampling effort between individuals, and to estimate how far we were from sampling the diversity of each individual completely, we plotted the accumulation of phylogenetic diversity (branch length) with sampling effort, in a manner analogous to rarefaction curves. We generated the phylogenetic diversity rarefaction curve for each individual by applying custom python code (http://bmf2.colorado.edu/unifrac/about.psp) to the Arb parsimony insertion tree[27].

Pyrosequencing of total community DNA. Shotgun sequencing runs were performed on the 454 FLX pyrosequencer from total faecal community DNA. Two samples were also analysed in a single run using Titanium extra-long-read pyrosequencing technology (see Supplementary Tables 4 and 5). Sequencing reads with degenerate bases ('Ns') were removed along with all duplicate sequences, as sequences of identical length and content are a common artefact of the pyrosequencing methodology. Finally, human sequences were removed by identifying sequences homologous to the *Homo sapiens* reference genome (BLASTN $E < 10^{-5}$, %identity > 75, score > 50).

CAZyme analysis. Metagenomic sequence reads were searched against a library of modules derived from all entries in the carbohydrate-active enzymes (CAZy) database (www.cazy.org using FASTY[33], $E < 10^{-6}$). This library consists of approximately 180,000 previously annotated modules (catalytic modules, carbohydrate-binding modules and other non-catalytic modules or domains of unknown function) derived from about 80,000 protein sequences. The number of sequencing reads matching each CAZy family was divided by the number of total sequences assigned to CAZymes and multiplied by 100 to calculate a relative abundance. An R^2 value was calculated for each pair of CAZy profiles. We then compared the distribution of glycoside hydrolase similarity scores with the distribution of glycosyltransferase similarity scores.

Statistical analyses. Xipe[23] (version 2.4) was used for bootstrap analyses of pathway enrichment and depletion, using the parameters sample size = 10,000 and confidence level = 0.95. Linear regressions were performed in Excel (version 11.0, Microsoft). Mann–Whitney and Student's t-tests were used to identify statistically significant differences between two groups (Prism version 4.0, GraphPad; Excel version 11.0, Microsoft). The Bonferroni correction was used to correct for multiple hypotheses. The Mantel test was used to compare distance matrices: the matrix of each pairwise comparison of the abundance of each reference genome, and the abundance of each metabolic pathway, were compared (Mantel program in Python using PyCogent[34]; 10,000 replicates). Data are represented as mean ± s.e.m. unless otherwise indicated.

Microbiome sequences were compared against the custom database of 44 gut genomes (BLASTX $E < 10^{-5}$, bitscore > 50, and %identity > 50). A gene-by-sample matrix was then screened to identify genes 'commonly-enriched' in either the obese or lean gut microbiome (defined by an odds ratio greater than 2 or less than 0.5 when comparing the pooled obese twin microbiomes with the pooled lean twin microbiomes, and when comparing each individual obese twin microbiome with the aggregate lean twin microbiome, or vice versa). The statistical significance of enriched or depleted genes was then calculated using a modified t-test (q value < 0.05; calculated with code supplied by M. Pop and J.R. White, University of Maryland). We also searched for genes that were consistently enriched or depleted in all six monozygotic twin pairs. A gene-by-sample matrix was generated based on BLASTX comparisons of each microbiome with our custom 44-genome database, to calculate an odds ratio based on the frequency of each gene in each twin versus the respective co-twin. The analysis revealed only 49 genes (odds ratio > 2 or < 0.5): they represent a variety of taxonomic groups, including Firmicutes, Bacteroidetes and Actinobacteria, and did not show any clear functional trends.

27. Ley, R. E. *et al.* Evolution of mammals and their gut microbes. *Science* **320,** 1647–1651 (2008).
28. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22,** 1658–1659 (2006).
29. DeSantis, T. Z. *et al.* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* **72,** 5069–5072 (2006).
30. Zhang, Z., Schwartz, S., Wagner, L. & Miller, W. A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.* **7,** 203–214 (2000).
31. Sheneman, L., Evans, J. & Foster, J. A. Clearcut: a fast implementation of relaxed neighbor joining. *Bioinformatics* **22,** 2823–2824 (2006).
32. Faith, D. P. Conservation evaluation and phylogenetic diversity. *Biol. Conserv.* **61,** 1–10 (1992).
33. Pearson, W. R., Wood, T., Zhang, Z. & Miller, W. Comparison of DNA sequences with protein sequences. *Genomics* **46,** 24–36 (1997).
34. Knight, R. *et al.* PyCogent: a toolkit for making sense from sequence. *Genome Biol.* **8,** R171 (2007).