NATURE|Vol 449|18 October 2007|doi:10.1038/nature06244

The Human Microbiome Project

Peter J. Turnbaugh, Ruth E. Ley, Micah Hamady, Claire M. Fraser-Liggett, Rob Knight & Jeffrey I. Gordon

A strategy to understand the microbial components of the human genetic and metabolic landscape and how they contribute to normal physiology and predisposition to disease.

Before the Human Genome Project was completed, some researchers predicted that ~100,000 genes would be found. So, many were surprised and perhaps humbled by the announcement that the human genome contains only ~20,000 protein-coding genes, not much different from the fruitfly genome. However, if the view of what constitutes a human is extended, then it is clear that 100,000 genes is probably an underestimate. The microorganisms that live inside and on humans (known as the microbiota) are estimated to outnumber human somatic and germ cells by a factor of ten. Together, the genomes of these microbial symbionts (collectively defined as the microbiome) provide traits that humans did not need to evolve on their own[1]. If humans are thought of as a composite of microbial and human cells, the human genetic landscape as an aggregate of the genes in the human genome and the microbiome, and human metabolic features as a blend of human and microbial traits, then the picture that emerges is one of a human 'supraorganism'.

To understand the range of human genetic and physiological diversity, the microbiome and the factors that influence the distribution and evolution of the constituent microorganisms must be characterized. This is one of the main goals of the Human Microbiome Project (HMP). The outcome might also provide perspective on contemporary human evolution: that is, on whether and how rapidly advancing technology, and the resultant transformation of human lifestyles and the biosphere, influences the 'micro-evolution' of humans and thereby health and predisposition to various diseases.

The HMP is a logical conceptual and experimental extension of the Human Genome Project. The HMP is not a single project. It is an interdisciplinary effort consisting of multiple projects, which are now being launched concurrently worldwide, including in the United States (as part of the next phase of the National Institutes of Health's Roadmap for Medical Research), Europe and Asia. The advent of highly parallel DNA sequencers and high-throughput mass spectrometers with remarkable mass accuracy and sensitivity is propelling microbiology into a new era, extending its focus from the properties of single organism types in isolation to the operations of whole communities. The new field of metagenomics involves the characterization of the genomes in these communities, as well as their corresponding messenger RNA, protein and metabolic products[2].

The HMP will address some of the most inspiring, vexing and fundamental scientific questions today. Importantly, it also has the potential to break down the artificial barriers between medical microbiology and environmental microbiology. It is hoped that the HMP will not only identify new ways to determine health and predisposition to diseases but also define the parameters needed to design, implement and monitor strategies for intentionally manipulating the human microbiota, to optimize its performance in the context of an individual's physiology. Examples of, and speculations about, the functional contributions of the microbiota are provided in Box 1.

In this article, we discuss the conceptual and experimental challenges that the HMP faces, as well as the rewards it might hold. We focus on the gut when providing examples, because this habitat harbours the largest collection of microorganisms.

Ecology and considerations of scale

Questions about the human microbiome are new only in terms of the system to which they apply. Similar questions have inspired and confounded ecologists working on macroscale ecosystems for decades. It is expected that the HMP will uncover whether the principles of ecology, gleaned from studies of the macroscopic world, apply to the microscopic world that humans harbour (see page 811). In particular, the following questions might be answered by the HMP. How stable and resilient is an individual's microbiota throughout one day and during his or her lifespan? How similar are the microbiomes between members of a family or members of a community, or across communities in different environments? Do all humans have an identifiable 'core' microbiome, and if so, how is it acquired and transmitted? What affects the genetic diversity of the microbiome (Fig. 1), and how does this diversity affect adaptation by the microorganisms and the host to markedly different lifestyles and to various physiological or pathophysiological states?

To address any question about the human microbiome, the microbiota needs to be sampled, and temporal and spatial scales need to be considered before undertaking this process. For example, microbial communities on human surfaces (that is, the skin and mucosal surfaces such as the gut) have a complex biogeography that can be defined at a range of distances: at the micrometre scale (the distribution of microorganisms on undigested food particles in the distal gut or across a mucosal barrier); at the centimetre scale (the distribution of communities around different teeth); and at the metre scale (the distribution of communities along the long axis of the gut).

Scale also has a further meaning. The core microbiome is whatever factors are common to the microbiomes of all or the vast majority of humans. At present, there are 6.7 billion humans on Earth. Because of various constraints, the human microbiome(s) will need to be characterized by comparing limited data types collected from a limited set of individuals. If human body habitats, such as the gut, are viewed as 'islands' in space and time, then island-biogeography theory, which was developed from studies of macroscale ecosystems[3], might be useful for understanding the observed microbial diversity. This theory states that community composition can depend strongly on the order in which species initially enter a community (a phenomenon known as multiple stable states[4]). The importance of the initial inoculating microbial community on the community composition at later stages is evident from animal studies. For example, in the mouse gut microbiota, the effects of maternal transmission (kinship) are apparent over several generations in animals of the same inbred strain[5]. Similarly, experiments in which the microbiota is transferred from one host to another, from conventionally raised mice or zebrafish to germ-free mice or zebrafish, demonstrate that the microbial community available to colonize the gut at the time of birth, together with the features of the gut habitat itself, conspire to select a microbiota[6]. To study the human microbiome, a few specific islands (humans) could be characterized in depth. Alternatively, the equivalent of a biogeography experiment could be carried out, in which general trends are inferred from a coarse-grained analysis of a larger number of humans, who are selected on the basis of demographic, geographical or epidemiological

factors. These strategies are complementary and, as discussed later, both will be needed to understand the human microbiome fully.

What do we know about the human microbiome?

Although the human microbiome is largely unexplored, recent studies have begun to reveal some tantalizing clues about its features.

Large variation in bacterial lineages between people

The decreasing cost and increasing speed of DNA sequencing, coupled with advances in the computational approaches used to analyse complex data sets[7–11], have prompted several research groups to embark on small-subunit (16S) ribosomal RNA gene-sequence-based surveys of bacterial communities that reside on or in the human body, including on the skin and in the mouth, oesophagus, stomach, colon and vagina[12–17] (see page 811). The 16S rRNA gene is found in all microorganisms and has enough sequence conservation for accurate alignment and enough variation for phylogenetic analyses. The largest reported data sets are for the gut, although the number of people sampled by using these culture-independent surveys is still limited. Most of the 10–100 trillion microorganisms in the human gastrointestinal tract live in the colon. More than 90% of all phylogenetic types (phylotypes) of colonic bacteria belong to just 2 of the 70 known divisions (phyla) in the domain Bacteria: the Firmicutes and the Bacteroidetes. For samples taken from the colon, the differences between individuals are greater than the differences between different sampling sites in one individual[15]. Moreover, faeces are representative of interindividual differences[5]. A recent study of 18,348 faecal 16S rRNA gene sequences collected from 14 unrelated adults over the course of a year showed large differences in microbial-community structure between individuals, and it established that community membership in each host was generally stable during this period[16]. How is such high interindividual diversity sustained? The observations about diversity in the human gut microbiota might fit with predictions of the neutral theory of community assembly, which states that most species share the same general niche (an ecological term that, in the case of microorganisms, refers to 'profession'), or the biggest niche, and therefore are likely to be functionally redundant[18]. Therefore, this theory predicts that highly variable communities (as defined by 16S rRNA gene lineages) will have high levels of functional redundancy between community members.

Ecosystem-level functions

Comparative metagenomics has uncovered functional attributes of the microbiome. The first reported application of metagenomic techniques to a human microbiome involved two unrelated, healthy adults. Compared with all previously sequenced microbial genomes and the human genome, metabolic reconstructions of the gut (faecal) microbiomes of these adults showed significant enrichment for genes involved in several metabolic pathways: the metabolism of xenobiotics (that is, foreign substances), glycans and amino acids; the production of methane; and the biosynthesis of vitamins and isoprenoids through the 2-methyl-D-erythritol 4-phosphate pathway[1].

The usefulness of comparative metagenomics is further underscored by a recent study, which showed that a host phenotype (obesity) can be correlated with the degree of representation of microbial genes involved in certain metabolic pathways[19]. Microbial-community DNA was isolated from the distal-gut contents of genetically obese animals (*ob/ob* mice, which have a mutation in the gene encoding leptin) and their lean littermates (+/+ or *ob*/+) and then sequenced. Predictions of microbial-community metabolism, based on community gene content, indicated that the obesity-associated gut microbiome has an increased capacity to harvest energy from the diet. Specifically, the *ob/ob* mouse microbiome was enriched for genes involved in importing and metabolizing otherwise indigestible dietary polysaccharides to short-chain fatty acids, which are absorbed by the host and stored as more complex lipids in adipose tissue. Biochemical analyses supported these predictions. Moreover, when adult germ-free wild-type mice were colonized with a gut microbiota from obese (*ob/ob*) or lean (+/+) mice, adiposity

Box 1 | Examples of functional contributions of the gut microbiota

Harvest of otherwise inaccessible nutrients and/or sources of energy from the diet, and synthesis of vitamins

The nutrient and/or energetic value of food is not absolute but is affected, in part, by the digestive capacity of an individual's microbiota[1,19,42–44]. This has implications for identifying individuals who are at risk of being malnourished or obese and treating them on the basis of a more personalized view of nutrition that considers their microbial ecology.

Metabolism of xenobiotics, and other metabolic phenotypes

The microbiota is a largely underexplored regulator of drug metabolism and bioavailability. Bioremediation-like functions of the microbiota, such as detoxifying ingested carcinogens, might affect a host's susceptibility to various neoplasms, both within and outside the gut. In addition, the metabolism of oxalate by the microbiota has been linked to a predisposition to the development of kidney stones[45]. Also, the modification of bile acids by microorganisms affects lipid metabolism in the host[44]. Ascribing metabolic phenotypes (also known as metabotypes) to the microbiota should extend our repertoire of personalized biomarkers of health and of disease susceptibility.

Renewal of gut epithelial cells

The renewal of gut epithelial cells is affected, in part, by interactions between the microbiota and immune cells. Effects could range from susceptibility to neoplasia[46] to the capacity for repairing a damaged mucosal barrier[47]. Germ-free mice renew gut epithelial cells at a slower rate than their colonized counterparts[47]. Comparing microbial communities that are physically associated with neoplasms and those with varying degrees of remoteness from the neoplasms might provide new mechanistic insights about cancer pathogenesis.

Development and activity of the immune system

The gut microbial community has an effect on both the innate immune system[48] and the adaptive immune system[49], and it contributes to immune disorders that are evident within and outside the gut. For example, in individuals with inflammatory bowel diseases, the immune response to the gut microbial community seems to be dysregulated: genome-wide association studies of patients with Crohn's disease have identified several human genes involved in both innate and adaptive immune responses[50]. In addition, susceptibility to colonization by enteropathogens is affected by the capacity of the microbiota to alter the expression of host genes encoding antimicrobial compounds[48,51]. Furthermore, the incidence of asthma is correlated with exposure to bacteria during childhood[52] and treatment with broad-spectrum antibiotics in early childhood[53].

Cardiac size

Germ-free animals have a smaller heart as a proportion of body weight than their colonized counterparts[54]. The mechanism underlying this phenotype has yet to be defined, but this finding emphasizes the importance of studying the extent to which human physiology is modulated by the microbiome.

Behaviour

Germ-free mice have greater locomotor activity than their colonized counterparts[43]. It will be interesting to study whether there are behavioural effects in humans. Has the microbiota evolved ways to benefit itself and its host by influencing human behaviour? Is altered production of neurologically active compounds (either directly, by the microbiota, or indirectly, by microbiota-mediated modulation of the expression of host genes that encode products normally involved in the biosynthesis and/or metabolism of these compounds) associated with any neurodevelopmental and/or psychiatric disorders?

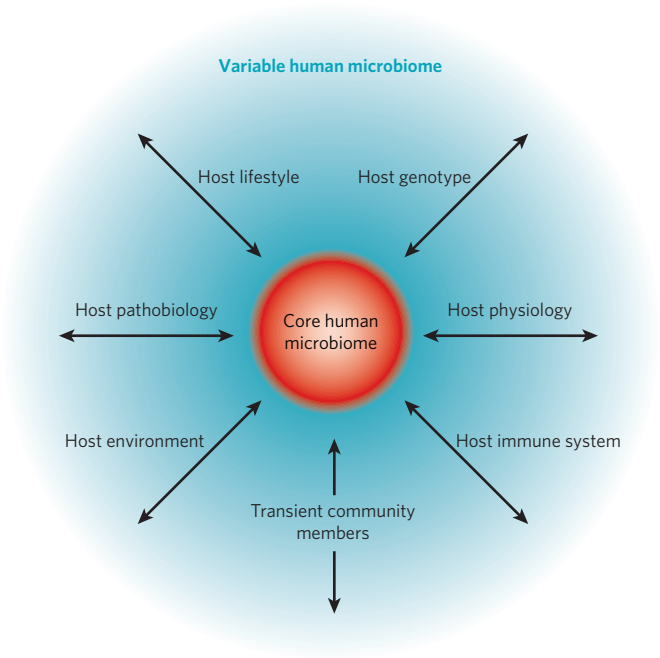

Figure 1 | The concept of a core human microbiome. The core human microbiome (red) is the set of genes present in a given habitat in all or the vast majority of humans. Habitat can be defined over a range of scales, from the entire body to a specific surface area, such as the gut or a region within the gut. The variable human microbiome (blue) is the set of genes present in a given habitat in a smaller subset of humans. This variation could result from a combination of factors such as host genotype, host physiological status (including the properties of the innate and adaptive immune systems), host pathobiology (disease status), host lifestyle (including diet), host environment (at home and/or work) and the presence of transient populations of microorganisms that cannot persistently colonize a habitat. The gradation in colour of the core indicates the possibility that, during human micro-evolution, new genes might be included in the core microbiome, whereas other genes might be excluded.

increased to a significantly greater degree in recipients of the microbiota from obese mice than in recipients of the microbiota from lean mice, supporting the conclusion that the obesity-associated gut microbiota has an increased (and transmissible) capacity to promote fat deposition[19]. This coupling of comparative metagenomics with germ-free animal models shows one way to proceed from *in silico* predictions to experimental tests of whole-community microbiome function.

Metagenomic data sets from different microbial ecosystems can also be compared, allowing the traits that are important to each to be uncovered[20]. An example of such an analysis is shown in Fig. 2. The human and mouse gut-microbiome data sets described in this section are compared with data sets obtained from three environmental communities: decaying whale carcasses located at the bottom of the ocean (known as whale falls), an agricultural-soil community and a survey of the Sargasso Sea[20,21]. DNA-sequencing reads were culled from each data set and matched to annotated genes in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database[22]. The gut microbiomes were found to cluster together and, compared with the environmental microbiomes, are enriched for predicted genes assigned to KEGG categories and pathways for carbohydrate and glycan metabolism (Fig. 2). Deeper sequencing of more human gut microbiomes will be required to determine whether these features are common traits of the human microbiome. (For further discussion of sampling issues, see the section Designing comparisons of microbial communities in humans.)

What will the HMP need for success?

Several factors will need to come together as this international effort is launched.

Sequencing more reference genomes

At present, metagenomic analyses of complex microbial communities are limited by the availability of suitable reference genomes, which are needed for confident assignment of the short sequences produced by the current generation of highly parallel DNA sequencers. These analyses are also constrained by a lack of knowledge about the niches of the organismal lineages that constitute these communities. An ongoing project to sequence the genomes of 100 cultured representatives of the phylogenetic diversity in the human gut microbiota[23] illustrates how reference genomes will help to interpret metagenomic studies. Capillary-sequencing reads from the human and mouse gut-microbiome data sets described earlier were matched to published microbial and eukaryotic genomes (KEGG database version 40 (ref. 22)) and 17 recently sequenced genomes of human gut bacteria (http://genome.wustl.edu/pub/) belonging to the divisions Bacteroidetes, Firmicutes and Actinobacteria (BLASTX best-BLAST-hit E value $< 10^{-5}$; http://www.ncbi.nlm.nih.gov/BLAST). These analyses showed that the quality of the sequence matches and the proportion of metagenomic read assignments increases with the inclusion of each additional gut bacterial genome.

The sequencing of more reference genomes, including genomes from multiple isolates of selected species-level phylotypes, should also help to answer questions about genetic variation within and between the major phylogenetic lineages in a given habitat, such as the gut. For example, a comparison of members of the Firmicutes and Bacteroidetes should provide insight into the extent of genetic redundancy and/or specialization between these two divisions. Given the extraordinary density of colonization in the distal gut (10^{11}–10^{12} organisms per ml of luminal content), these extra genomes would also provide an opportunity to determine more accurately the role of horizontal gene transfer in the evolution of gut microorganisms within and between hosts[24], as well as the extent to which the gene content of these microorganisms reflects their phylogenetic history.

To obtain reference-genome sequences, it will be crucial to develop new methods for retrieving microorganisms that cannot be cultured at present. Recently, several methods — fluorescence *in situ* hybridization with phylogenetic markers, flow cytometry, and whole-genome amplification and shotgun sequencing — have been used to obtain a partial genome assembly for a member of the candidate phylum TM7, providing a first look at a group of microorganisms with no culturable representatives[25]. In addition, methods such as the encapsulation of cells in gel microdroplets are aimed at enabling high-throughput culture of microorganisms in a simulated natural environment[26].

Linking short gene fragments to organisms

Because metagenomic data sets consist largely of unassembled sequence data, another major challenge is to link genes to organisms or at least to broader taxonomic classifications. Several approaches exist[27–29], but no tools have been developed for the automated analysis of large data sets containing mostly short sequence reads, without relying on phylogenetic marker genes. Thus, developing an accurate and scalable way to phylogenetically classify huge numbers of short sequence reads is essential.

The two general marker-independent approaches to phylogenetic assignment are to use Markov models based on the frequency of short nucleotide sequences (or 'words') in the reads and to use homology searching to place each sequence fragment in the context of a phylogenetic tree. Because of statistical sampling issues, the Markov-model-based approach is likely to be relatively insensitive, especially for short sequences and for sequences from heterogeneous genomes. The homology-search-based approach is probably more accurate and provides the additional advantage of placing each sequence in the context of a multiple alignment and a phylogenetic tree, which can then be used in further studies. However, sequences without identifiable homologues cannot be analysed in this way. A combination of these two general strategies is likely to be the best approach to understanding the functions associated with each metagenome.

There are three key issues when considering these approaches. First, it is important to understand how accurate the phylogenetic classification obtained by using each method can be, especially in the face of horizontal gene transfer. Second, it will be necessary to find better, faster and more scalable heuristics for generating huge phylogenetic trees that contain millions of sequences. Third, it is important to identify the best way to account for the effects of both the genome and the function of each encoded protein on the overall composition of each sequence. In particular, heterogeneous rates of evolution in different protein families pose substantial problems for search-based methods: considerable similarities at the primary-structure level might not persist over time, and the secondary and tertiary structures of the proteins are usually unknown, thus preventing the use of structure-based alignment techniques.

Designing comparisons of microbial communities in humans

Understandably, there will be great pressure at early stages of the HMP to focus on disease states. However, 'normal' states need to be defined before the effect of the microbiota on disease predisposition and pathogenesis can be evaluated, and this will require time, resources and discipline.

Several issues need to be considered when designing ways to generate an initial set of reference microbiomes from healthy individuals. What is the degree of genetic relatedness between those who are sampled: for example, should the initial focus be on monozygotic and dizygotic twins and their mothers? What is the place of the sampled individuals in the family structure? What age are they, and what are their demographics (for example, rural versus urban environment and lifestyle)? What are the ethical, legal and logistical barriers that need to be overcome to obtain, without exploitation, samples and metadata (that is, 'relevant' environmental and host parameters) from people with diverse cultural and socio-economic backgrounds? What types of comparison are needed: for example, should there be measurements of diversity within samples (α diversity); between samples (β diversity); between body habitats in a given individual; and/or between family members for a given habitat? And what protocols could or should be used for sampling surface-associated microbial communities? This last issue is a major unresolved technical problem. At present, there are no methods to retrieve sufficient quantities of microorganisms from various body surfaces, such as the skin and the vaginal mucosa, in a reproducible and representative manner, and sufficiently free of human cells, so that the microbiome can be sequenced. It is also unclear at what temporal and spatial scales this sampling should occur.

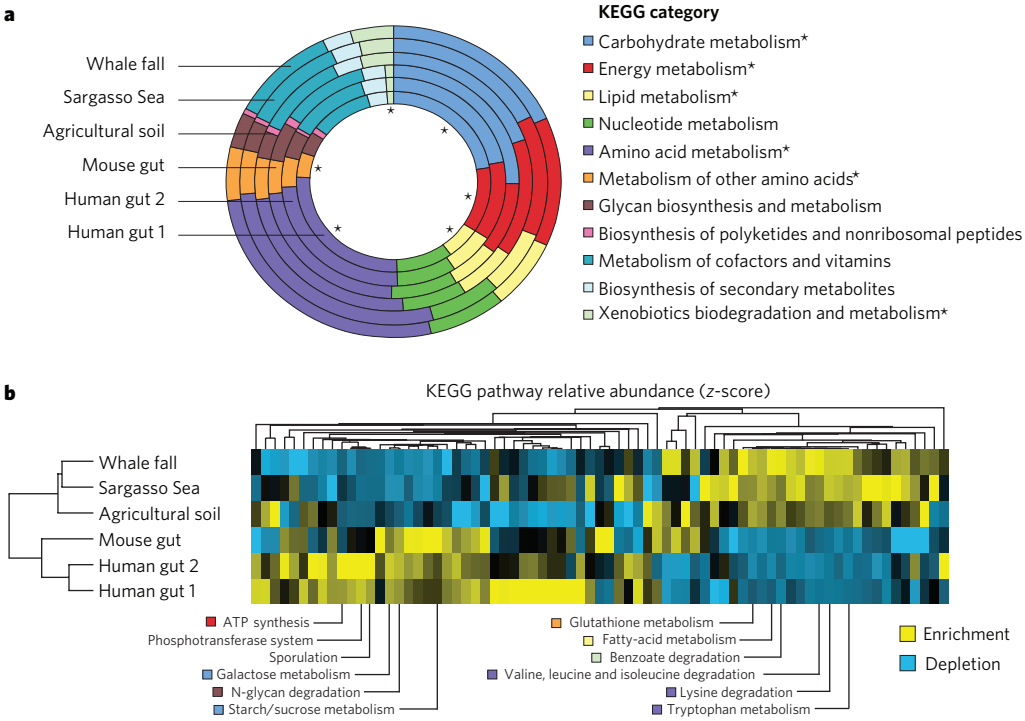

Figure 2 | Functional comparison of the gut microbiome with other sequenced microbiomes. a, Relative abundance of predicted genes, assigned to KEGG categories for metabolism. Several gut-microbiome data sets were analysed: a combined mouse gut data set ($n = 5$ animals)[19] and two human gut data sets[1]. Three 'environmental'-microbiome data sets were also analysed: a combined whale-fall data set ($n = 3$ samples, from three separate whale falls)[20], an agricultural soil data set[20] and a combined Sargasso Sea data set ($n = 7$ samples)[21]. Forward DNA-sequencing reads (from a capillary instrument) were culled from each data set and mapped onto reference microbial and eukaryotic genomes from the KEGG database (version 40; BLASTX best-BLAST-hit E value < 10^{-5})[22]. The best BLAST hit was used to assign each sequencing read to a KEGG orthologous group, which was then assigned to KEGG pathways and categories. The distribution of ~15,000 KEGG-category assignments across each of the six data sets was then used to construct two combined data sets of ~45,000 KEGG-category assignments each. Asterisks indicate categories that are significantly enriched or depleted in the combined gut data set compared with the combined environmental data set (χ^2 test, using the Bonferroni correction for multiple hypotheses, $P < 10^{-4}$). **b**, Hierarchical clustering based on the relative abundance of KEGG pathways. Metabolic pathways found at a relative abundance of more than 0.6% (that is, assignments to a given pathway divided by assignments to all pathways) in at least two microbiomes were selected. These relative-abundance values were transformed into z-scores[20], which are a measure of relative enrichment (yellow) and depletion (blue). The data were clustered according to microbiomes and metabolic pathways by using a euclidean distance metric (Cluster 3.0)[40]. The results were visualized by using Java Treeview[41]. The clustering of environmental data sets was consistent irrespective of the distance metrics used, including Pearson's correlation (centred or uncentred), Spearman's rank correlation, Kendall's tau and city-block distance. The 12 most discriminating KEGG pathways are shown (based on the ratio of the mean gut relative abundance to the mean environmental relative abundance). The KEGG category for each metabolic pathway is indicated by coloured squares. Pathway names without corresponding coloured squares include sporulation (which is involved in cell growth and death) and the phosphotransferase system (which is involved in membrane transport). The gut microbiome is enriched for proteins involved in sporulation (reflecting the high relative abundance of Firmicutes) and for pathways involved in importing and degrading polysaccharides and simple sugars.

As is the case for many ecological studies, we must choose between deep sampling of a small number of sites (individual people and body habitats) and broad sampling. Broad sampling would enable the general principles that control community structure and function to be uncovered. However, deep sampling of body habitats from a few individuals is needed to estimate the distribution of species and genes: these estimates, in turn, will allow modelling of the trade-offs between deeper sampling of fewer individuals and shallower sampling of more individuals. Unlike the situation with the International HapMap Project[30], which sought to describe common patterns of genetic variation in humans, there is no baseline expectation for the amount of diversity in different microbial communities, and the development of careful sampling models will be essential for optimizing the use of resources. Also, given the rapid development of new and more massively parallel sequencing technologies, systematic testing will be required to identify ways to maximize sequencing coverage affordably, while maintaining the ability to analyse and assemble genome fragments.

Ultimately, the goal is to associate differences in communities with differences in metabolic function and/or disease. Thus, another key challenge for the HMP is to define the concept of 'distance' between communities and to associate these distances with host biology and various metadata. UniFrac[11,31,32] and other phylogenetic techniques address this problem for 16S rRNA gene data sets and could be extended to the assessment of metagenomic data. With the distances defined, statistical techniques will need to be developed and refined so that multivariate data sets can be integrated into a unified framework, enabling the components of the microbiome that could affect human health and disease to be identified.

The HMP will also require researchers to move beyond comparative genomics to an integrated 'systems metagenomics' approach that accounts for microbial community structure (the microbiota), gene content (the microbiome), gene expression (the 'meta-transcriptome' and 'meta-proteome') and metabolism (the 'meta-metabolome'). Some progress has been made towards generating 'functional gene arrays', to determine the relative abundance of specific genes or transcripts in microbiomes[33–35]. More work is needed to improve the sensitivity of gene arrays and to apply this approach to complex communities such as the human microbiome. The construction and sequencing of complementary DNA libraries form an alternative approach, and these have already been used to examine microbial and eukaryotic mRNA from environmental samples[36,37]. However, high-throughput methods for eliminating highly abundant transcripts (for example, those from rRNA genes) are needed.

Box 2 | A proposal for staging the Human Microbiome Project

In this conceptualization, the HMP is portrayed as a three-tiered effort, with the first tier composed of three components (or pillars).

First tier: initial data acquisition and analysis

Pillar one: construct deep draft assemblies of reference genomes

- Select cultured representatives of microbial divisions in a given habitat by examining 'comprehensive' 16S-rRNA-gene-based surveys
- Create a publicly accessible database of human-associated 16S rRNA gene phylotypes (which could be referred to as the 'virtual microbial body') to facilitate selection by allowing comparisons within and between body habitats, within and between individuals, and between separate studies; and develop faster and better alignment algorithms for building phylogenetic trees
- Obtain phylotypes of interest from existing culture collections (both public and 'private'), with consent to deposit sequence data in the public domain
- Improve technology for culturing organisms that cannot be cultured at present
- Select a subset of 'species' for pan-genomic analysis (that is, the characterization of multiple isolates of a species-level phylotype), and develop better methods for detecting horizontal gene transfer
- Ensure data flow to, and data capture by, the Protein Structure Initiative (http://www.structuralgenomics.org)
- Deposit sequenced isolates, together with information about habitat of origin, conditions for growth and phenotypes, in a public culture repository that can maintain and distribute microorganisms

Pillar two: obtain reference microbiome data sets

- Focus on monozygotic and dizygotic twin pairs and their mothers
- Determine the advantages and disadvantages of different DNA-sequencing platforms
- Characterize, at a preliminary level, within-sample (α) diversity and between-sample (β) diversity
- Ensure the availability of user-friendly public databases in which biomedical and environmental metagenomic data sets are deposited, together with sample metadata
- Develop and optimize tools (distance metrics) for comparing 16S rRNA gene and community metagenomic data sets, and feed back to the pipeline in which cultured or retrieved representatives of different habitat-associated communities are selected and characterized
- Establish specimen and data archives with distribution capabilities
- Generate large-insert microbiome libraries for present and future functional metagenomic screens
- Coordinate with environmental metagenomics initiatives so that efforts to develop resources and tools are reinforced and shared

Pillar three: obtain shallower 16S rRNA gene and community metagenomic data sets from moderate number of samples

- Extend sampling of families (for example, to fathers, siblings and children of twins), expand the age range of individuals sampled, and explore demographic, socio-economic and cultural variables
- Establish a global sample-collection network, including countries in which social structures, technologies and lifestyles are undergoing rapid transformation
- Develop and optimize computational tools and metrics for comparing these diverse multivariate data sets
- Develop and optimize tools for analysing the transcriptome, proteome and metabolome, by using the same biological specimens used for sequencing community DNA, and develop and optimize tools for higher-throughput analyses
- Design and test experimental models for identifying the principles that control the assembly and robustness of microbial communities

Second tier: choice of individuals that represent different clusters, for additional deep sequencing

- Estimate sampling depth and number of individuals needed to characterize the 'full' human microbiome; the granularity of the characterization needs to match the data
- Search for relatives of human-associated microbial species and gene lineages in other mammalian microbial communities and in the environment, and sequence the genomes of these microorganisms (defining niches; feed back to the first tier)

Third tier: global human microbiome diversity project

- Sequence at a shallow level the microbiomes from a large (to be defined) sample of geographically, demographically and culturally diverse individuals
- Choose individuals with different clinical 'parameters', and carry out association studies and biomarker panning
- Sequence at a large scale reservoirs of microorganisms and genes (for example, soils and water sources), and associate this information with the fluxes of energy, materials, genes and microbial lineages into the human microbiome (with the help of microbial observatories and human observatories)
- Apply the knowledge gained (for example, towards developing diagnostic tests, therapies and strategies for improving the global food chain), and educate people (including the public, governments, and present and future researchers in the field)

Proteomic tools, including Elucidator (http://www.rosettabio.com/products/elucidator) and SEQUEST (http://fields.scripps.edu/sequest), are also available for analysing complex samples. And comprehensive microbial protein-sequence databases (for example, Protein Clusters; http://www.ncbi.nlm.nih.gov/sites/entrez?db=proteinclusters) are continually updated. In addition, custom databases can be created from metagenomic data sets and used to interpret mass-spectrometry data sets[38]. Given the limited knowledge of the biological transformations that human microbial communities support, meta-metabolomics is likely to be challenging. Tools and databases for metabolite identification still need to be developed, despite the existence of highly accurate instrumentation. (For example, Fourier-transform ion-cyclotron-resonance mass spectrometers have a mass accuracy of < 1–10 parts per million.) This situation should be helped by ambitious efforts that are underway to catalogue thousands of human-associated metabolites and to generate a searchable database[39]. Together, these complementary measurements will allow a far richer characterization of human microbial communities. They will also enable the variation that is typical of a healthy state to be defined, making it possible to search for deviations that are associated with disease.

Depositing and distributing data

Vast amounts of information will be generated by the HMP, as well as by metagenomic surveys of the environment, so new procedures and increased capabilities are required for depositing, storing and mining different data types. Important goals include the following: a minimum set of standards for annotation; a flexible, simple and open format for depositing metadata (taking a lesson from clinical studies because the relevant parameters are largely unknown); efficient analysis tools for the general user that are broadly applicable (including tools for meta-analyses of varied data types); and an adequate cyberinfrastructure to support the computing needs of the research community.

Using model systems

Although the HMP is human-focused, model organisms and other experimental systems are needed for aspects of the project that cannot be tested in humans: these will define how communities operate and interact with their hosts, characterize the determinants of community robustness and identify biomarkers of community composition and/or performance. Germ-free animals, both wild-type and genetically engineered, that have been colonized at various stages of their lives with simplified microbial communities composed of a few sequenced members, or with more complex consortia, should be useful because they provide the opportunity to constrain several variables, including host genotype, microbial diversity and environmental factors such as diet. *In vitro* models, including microfluidic-based techniques for single-cell sorting and measurements, should help to define the biological properties of microorganisms and the consequences of interactions between microorganisms.

A model for staging the HMP

On the basis of all of these considerations, one potential way of staging the HMP is outlined in Box 2. The search for data will be global in many senses. It embraces the planet and its (human) inhabitants. It requires individuals from the clinical, biological and physical-engineering sciences to participate, including those with expertise in disciplines ranging from mathematics to statistics, computer science, computational biology, microbiology, ecology, evolutionary biology, comparative genomics and genetics, environmental and chemical engineering, chemistry and biochemistry, human systems physiology, anthropology, sociology, ethics and law. It requires coordination between scientists, governments and funding agencies. And it is one element of a worldwide effort to document, understand and respond to the consequences of human activities — not only as they relate to human health but also as they relate to the sustainability of the biosphere. It is hoped that, just as microbial observatories have been set up to monitor changes in terrestrial and ocean ecosystems worldwide, an early outcome of the HMP will be the establishment of 'human observatories' to monitor the microbial ecology of humans in different settings.

Concluding remarks

Many outcomes of the HMP can be predicted: for example, new diagnostic biomarkers of health, a twenty-first century pharmacopoeia that includes members of the human microbiota and the chemical messengers they produce, and industrial applications based on enzymes that are produced by the human microbiota and can process particular substrates. One important outcome is anticipated to be a deeper understanding of the nutritional requirements of humans. This, in turn, could result in new recommendations for food production, distribution and consumption that are formulated based on knowledge of the microbiome. ■

Peter J. Turnbaugh, Ruth E. Ley and Jeffrey I. Gordon are at the Center for Genome Sciences, Washington University School of Medicine, St Louis, Missouri 63108, USA. Micah Hamady is at the Department of Computer Science, University of Colorado at Boulder, Boulder, Colorado 80309, USA. Claire M. Fraser-Liggett is at the Institute of Genome Sciences, University of Maryland School of Medicine, Baltimore, Maryland 21201, USA. Rob Knight is at the Department of Chemistry and Biochemistry, University of Colorado at Boulder, Boulder, Colorado 80309, USA.

1. Gill, S. R. *et al.* Metagenomic analysis of the human distal gut microbiome. *Science* **312,** 1355–1359 (2006).
2. The Committee on Metagenomics. *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet* (The National Academies Press, Washington DC, 2007).
3. MacArthur, R. H. & Wilson, E. O. *The Theory of Island Biogeography* (Princeton Univ. Press, Princeton, 1967).
4. Ambramsky, Z. & Rosenzweig, M. L. The productivity diversity relationship: Tilman's pattern reflected in rodent communities. *Nature* **309,** 150–151 (1984).
5. Ley, R. E. *et al.* Obesity alters gut microbial ecology. *Proc. Natl Acad. Sci. USA* **102,** 11070–11075 (2005).
6. Rawls, J. F., Mahowald, M. A., Ley, R. E. & Gordon, J. I. Reciprocal gut microbiota transplants from zebrafish and mice to germ-free recipients reveal host habitat selection. *Cell* **127,** 423–433 (2006).
7. Ludwig, W. *et al.* ARB: a software environment for sequence data. *Nucleic Acids Res.* **32,** 1363–1371 (2004).
8. Cole, J. R. *et al.* The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res.* **33,** D294–D296 (2005).
9. Schloss, P. D. & Handelsman, J. DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl. Environ. Microbiol.* **71,** 1501–1506 (2005).
10. DeSantis, T. Z. *et al.* NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. *Nucleic Acids Res.* **34,** W394–W399 (2006).
11. Lozupone, C., Hamady, M. & Knight, R. UniFrac — an online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics* **7,** 371 (2006).
12. Gao, Z. *et al.* Molecular analysis of human forearm superficial skin bacterial biota. *Proc. Natl Acad. Sci. USA* **104,** 2927–2932 (2007).
13. Pei, Z. *et al.* Bacterial biota in the human distal esophagus. *Proc. Natl Acad. Sci. USA* **101,** 4250–4255 (2004).
14. Bik, E. M. *et al.* Molecular analysis of the bacterial microbiota in the human stomach. *Proc. Natl Acad. Sci. USA* **103,** 732–737 (2006).
15. Eckburg, P. B. *et al.* Diversity of the human intestinal microbial flora. *Science* **308,** 1635–1638 (2005).
16. Ley, R. E., Turnbaugh, P. J., Klein, S. & Gordon, J. I. Microbial ecology: Human gut microbes associated with obesity. *Nature* **444,** 1022–1023 (2006).
17. Hyman, R. W. *et al.* Microbes on the human vaginal epithelium. *Proc. Natl Acad. Sci. USA* **102,** 7952–7957 (2005).
18. Hubbell, S. P. Neutral theory and the evolution of ecological equivalence. *Ecology* **87,** 1387–1398 (2006).
19. Turnbaugh, P. J. *et al.* An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444,** 1027–1031 (2006).
20. Tringe, S. G. *et al.* Comparative metagenomics of microbial communities. *Science* **308,** 554–557 (2005).
21. Venter, J. C. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304,** 66–74 (2004).
22. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32,** D277–D280 (2004).
23. Gordon, J. I. *et al.* Extending our view of self: the Human Gut Microbiome Initiative (HGMI). *National Human Genome Research Institute* <http://www.genome.gov/Pages/Research/Sequencing/SeqProposals/HGMISeq.pdf (2005).
24. Xu, J. *et al.* Evolution of symbiotic bacteria in the distal human intestine. *PLoS Biol.* **5,** e156 (2007).
25. Podar, M. *et al.* Targeted access to the genomes of low abundance organisms in complex microbial communities. *Appl. Environ. Microbiol.* **73,** 3205–3214 (2007).
26. Zengler, K. *et al.* Cultivating the uncultured. *Proc. Natl Acad. Sci. USA* **99,** 15681–15686 (2002).
27. Teeling, H. *et al.* TETRA: a web-service and stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics* **5,** 163 (2004).

28. McHardy, A. C. *et al.* Accurate phylogenetic classification of variable-length DNA fragments. *Nature Methods* **4,** 63–72 (2007).
29. von Mering, C. *et al.* Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science* **315,** 1126–1130 (2007).
30. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437,** 1299–1320 (2005).
31. Lozupone, C. & Knight, R. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* **71,** 8228–8235 (2005).
32. Lozupone, C. A., Hamady, M., Kelley, S. T. & Knight, R. Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. *Appl. Environ. Microbiol.* **73,** 1576–1585 (2007).
33. Wu, L. *et al.* Development and evaluation of functional gene arrays for detection of selected genes in the environment. *Appl. Environ. Microbiol.* **67,** 5780–5790 (2001).
34. Gentry, T. J. *et al.* Microarray application in microbial ecology research. *Microb. Ecol.* **52,** 159–175 (2006).
35. Gao, H. *et al.* Microarray-based analysis of microbial community RNAs by whole-community RNA amplification. *Appl. Environ. Microbiol.* **73,** 563–571 (2007).
36. Poretsky, R. S. *et al.* Analysis of microbial gene transcripts in environmental samples. *Appl. Environ. Microbiol.* **71,** 4121–4126 (2005).
37. Grant, S. *et al.* Identification of eukaryotic open reading frames in metagenomic cDNA libraries made from environmental samples. *Appl. Environ. Microbiol.* **72,** 135–143 (2006).
38. Ram, R. J. *et al.* Community proteomics of a natural microbial biofilm. *Science* **308,** 1915–1920 (2005).
39. Wishart, D. S. *et al.* HMDB: the human metabolome database. *Nucleic Acids Res.* **35,** D521–D526 (2007).
40. de Hoon, M. J., Imoto, S., Nolan, J. & Miyano, S. Open source clustering software. *Bioinformatics* **20,** 1453–1454 (2004).
41. Saldanha, A. J. Java Treeview — extensible visualization of microarray data. *Bioinformatics* **20,** 3246–3248 (2004).
42. Backhed, F. *et al.* The gut microbiota as an environmental factor that regulates fat storage. *Proc. Natl Acad. Sci. USA* **101,** 15718–15723 (2004).
43. 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).
44. Martin, F. J. *et al.* A top-down systems biology view of microbiome–mammalian metabolic interactions in a mouse model. *Mol. Syst. Biol.* **3,** doi:10.1038/msb4100153 (2007).
45. Sidhu, H., Allison, M. J., Chow, J. M., Clark, A. & Peck, A. B. Rapid reversal of hyperoxaluria in a rat model after probiotic administration of *Oxalobacter formigenes*. *J. Urol.* **166,** 1487–1491 (2001).
46. Chu, F. F. *et al.* Bacteria-induced intestinal cancer in mice with disrupted *Gpx1* and *Gpx2* genes. *Cancer Res.* **64,** 962–968 (2004).
47. Pull, S. L., Doherty, J. M., Mills, J. C., Gordon, J. I. & Stappenbeck, T. S. Activated macrophages are an adaptive element of the colonic epithelial progenitor niche necessary for regenerative responses to injury. *Proc. Natl Acad. Sci. USA* **102,** 99–104 (2005).
48. Hooper, L. V., Stappenbeck, T. S., Hong, C. V. & Gordon, J. I. Angiogenins: a new class of microbicidal proteins involved in innate immunity. *Nature Immunol.* **4,** 269–273 (2003).
49. Mazmanian, S. K., Liu, C. H., Tzianabos, A. O. & Kasper, D. L. An immunomodulatory molecule of symbiotic bacteria directs maturation of the host immune system. *Cell* **122,** 107–118 (2005).
50. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447,** 661–678 (2007).
51. Cash, H. L., Whitham, C. V., Behrendt, C. L. & Hooper, L. V. Symbiotic bacteria direct expression of an intestinal bactericidal lectin. *Science* **313,** 1126–1130 (2006).
52. Braun-Fahrlander, C. *et al.* Environmental exposure to endotoxin and its relation to asthma in school-age children. *N. Engl. J. Med.* **347,** 869–877 (2002).
53. Kozyrskyj, A. L., Ernst, P. & Becker, A. B. Increased risk of childhood asthma from antibiotic use in early life. *Chest* **131,** 1753–1759 (2007).
54. Wostmann, B. S., Bruckner-Kardoss, E. & Pleasants, J. R. Oxygen consumption and thyroid hormones in germfree mice fed glucose–amino acid liquid diet. *J. Nutr.* **112,** 552–559 (1982).

Acknowledgements We apologize that we could not cite many excellent studies because of space constraints.

Author Information Reprints and permissions information is available at npg.nature.com/reprints. Correspondence should be addressed to J.I.G. (jgordon@wustl.edu).