The Human Microbiome Project in 2011 and Beyond

Lita M. Proctor[1],*
[1]National Human Genome Research Institute, NIH, 5635 Fishers Lane, Bethesda, MD 20892-9305, USA
*Correspondence: lita.proctor@nih.gov

DOI 10.1016/j.chom.2011.10.001

The human microbiome comprises the genes and genomes of the microbiota that inhabit the body. We highlight Human Microbiome Project (HMP) resources, including 600 microbial reference genomes, 70 million 16S sequences, 700 metagenomes, and 60 million predicted genes from healthy adult microbiomes. Microbiome studies of specific diseases and future research directions are also discussed.

The NIH Human Microbiome Project: A Community Resource

Though other terms such as endogenous or commensal microbiota have been used to describe the resident microorganisms of the human body, these microbial communities are more than a collection of microbial cells to be counted. The human microbiome encompasses the full complement of microbial genes, gene products, and genomes of the microbiota (which include bacteria, archaea, eukaryotic viruses, bacteriophages, and eukaryotic microbes) that call the human body home and interact with the human host to prime immunity and to maintain host health. A revolution is occurring in our understanding of the basis of many common and complex diseases, infectious and otherwise, as the role of the human microbiome is incorporated into our thinking about health and disease. At 10–100 trillion cells, thousands of species—and at least 20 million unique microbial genes—the global microbiome contributes to the health and maintenance of the human superorganism. Interest in this system has been motivated throughout the past decade by simultaneous advances in sequencing technologies and in microbial ecology, by the recognition that the human genome is only part of our genetic composition, by an increased understanding that the human host and microbiota have coevolved. and that the microbiome is intimately involved in the development and maintenance of the immune system. To catalyze the field, in October 2007 the NIH formally launched the 5 year Human Microbiome Project (HMP) as a community resource program (http://commonfund.nih.gov/hmp), defined as a research project "specifically devised and implemented to create a set of data, reagents, or other material whose primary utility will be as a resource for the broad scientific community" (http://www.genome.gov/10506537). A marker paper that described the HMP and its data release policy serves as an outline of the HMP resources under development (Peterson et al., 2009).

Whereas other national and international research initiatives focus on the microbiome of a specific part of the body, the HMP is (1) surveying the microbiomes across the bodies of a cohort of healthy adults to produce a reference dataset of baseline microbiomes, (2) developing a catalog of microbial genome sequences of reference strains, and (3) evaluating the properties of microbiomes associated with specific gastrointestinal tract, urogenital, and skin diseases in a collection of Demonstration Projects. In addition, three programs in technology development, computational tools development, and in the ethical and legal implications of microbiome research were created to support the field. A Data Analysis and Coordination Center (DACC) was established to support the sequencing by the data processing and data analysis efforts of the 100+ member HMP Research Network Consortium and to serve as a portal to the data sets, the reference strain catalog, the computational tools, and the other resources developed for the larger research community (http://www.hmpdacc.org).

The NIH NCBI BioProject page, the public repository for the data, is an excellent source to learn about the data types produced in the program (http://www.ncbi.nlm.nih.gov/bioproject/43021). There are four projects listed under the HMP program based on the four data types produced: (1) targeted 16S ribosomal RNA gene sequences, used as a taxonomic marker, produced from the healthy adult cohort study; (2) whole-genome shotgun metagenomic sequences (metagenome: all of the gene sequences from one microbial community) produced from the healthy adult cohort study; (3) the reference strain microbial genome sequences; and (4) data sets produced in individual Demonstration Project activities. A conceptual diagram of the HMP (Figure 1) depicts how the six initiatives of the program interact through consortium activities. The consortium, which consists of the initiative research teams, members from the larger scientific community, and the NIH program staff, interact through over 20 Working Groups, biannual consortium meetings, and a DACC-managed shared electronic resource for consortial work. This work is described under the three broad categories of (1) sample collection, (2) data generation, and (3) data processing and analysis (Figure 1). Details of the initiatives, the consortium activities, and the resources produced thus far follow.

Progress So Far

Reference Strain Microbial Genome Sequence Catalog

The HMP has assembled a key reference data set of microbial genome sequences collected from the major body regions of the human microbiome, primarily bacterial, although it also includes archaea, viruses, bacteriophages, and eukaryotic microorganisms. The project's target catalog of 3000 microbial genome sequences is intended as a reference for the interpretation of the 16S ribosomal RNA gene sequences, as well as a scaffold for rapid assembly of metagenomic sequences determined from the microbial communities. A publication documenting the analysis of the first 178 microbial isolates was recently published (Nelson et al., 2010); just this subset of the catalog described over 550,000 predicted genes, 30,000 of which were novel.

As of this writing, almost 1900 microbial strains have been sequenced or are in progress for the HMP reference strain catalog (http://www.hmpdacc-resources.org/hmp_catalog). Approximately 600 of these are available in GenBank (http://www.ncbi.nlm.nih.gov/bioproject/28331), and cultures of the corresponding reference strains are available at the HMP Strain Repository in the ATCC/Biodefense and Emerging Infectious Diseases Research Repository (BEI) (http://www.beiresources.org).

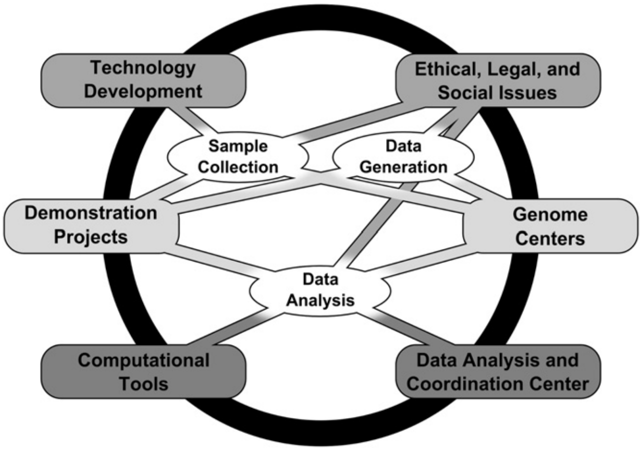

Figure 1. Conceptual Diagram of the HMP

The HMP is composed of six formal initiatives, shown around the circle; these include Technology Development; Ethical, Legal, and Social Issues; Genome Centers; the Data Analysis and Coordination Center; Computational Tools; and the Demonstration Projects. These initiatives interact through the activities of the 100+ member HMP Research Network Consortium, which also include members of the larger scientific community and NIH program staff. The consortium activities, shown in the three interior bubbles, include (1) sample collection, which includes the clinical protocols development and collection of microbiome specimens and nucleic acid sample preparation from the specimens in the healthy cohort study and in the Demonstration Projects; (2) data generation, which includes all of the sequencing activities for the healthy cohort, Demonstration Projects, and the reference strain microbial genomes, and (3) data analysis, which includes the extensive data processing, benchmarking, and quality-control steps needed to produce data for public release and for the analysis of microbiome sequence data by the consortium. The connecting lines graphically depict the major interactions between the initiatives.

Healthy Adult Cohort Study

The second major resource of the HMP is the largest study to date of the microbiomes of five major areas of the body of healthy adults (airway, skin, oral cavity, gastrointestinal tract, and vagina; see Figure 2). Several specific body sites were sampled within each major area (18 in total), and as the volunteers were clinically verified as being free of overt disease in all of the body sites, this study is known as the healthy adult cohort study.

Extensive exclusion criteria for selection of healthy volunteers were developed based on a combination of health history (particularly systemic disorders), use of antibiotics, probiotics, or immunomodulators, as well as physical examinations of each volunteer. Volunteers were not always initially free of disease in all body sites; a common example of this was with the oral cavity, where otherwise healthy volunteers had dental caries and required treatment before re-entering the study. Three hundred adult volunteers were enrolled at two clinical centers (Baylor College of Medicine, Houston, TX; Washington University, St. Louis, MO); these included equal numbers of 18- to 40-year-old men and women, 20% of whom identified themselves as a racial minority and 11% of whom identified themselves as Hispanic. Exclusion and inclusion criteria, clinical sampling procedures, and the corresponding clinical metadata can be found at the NCBI database of Genotypes and Phenotypes (dbGaP, http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000228). Of the 300 volunteers in this study, 279 were sampled twice and 100 were sampled a third time over approximately 22 months. Among the 18 total body sites sampled, the oral cavity had the largest number of sites (nine; see Figure 2), and all were directly sampled except for the gut tract, for which stool served as a proxy. Blood was collected for serum and for future whole-genome sequencing, and lymphocytes were harvested for cell lines; these specimens are being held at the NHGRI Coriell Repository for future distribution. The genome centers at four institutions (Baylor College of Medicine, Broad Institute, J. Craig Venter Institute, and Washington University at St. Louis) carried out the sequencing activities.

Of the >11,000 primary microbiome specimens collected for the full cohort, all have been sequenced for the 16S rRNA gene taxonomic marker. Metagenomic sequence data has additionally been generated from approximately 750 of the nucleic acid samples, comprising of 7 body sites from 100 subjects. About 50% of the full set of 16S data and 90% of the metagenomic data was targeted by the Consortium for a global analysis. The data sets for the global analysis comprise over 70 million 16S rRNA gene sequences, and after removing contaminating human sequence (on average ~50% of total metagenomic sequence) over 3.5 terabases (Tbp) of whole-genome shotgun metagenomic data. Although only about 60% of these metagenomic sequences could be aligned to a reference microbial genome sequence, annotation resulted in over 60 million predicted genes (i.e., open-reading frames). The human microbiome clearly contains a rich diversity of genetic information and function, much of it uncharacterized and often completely novel.

As of this writing, 126 publications in PubMed cite the HMP. The HMP Consortium is currently finalizing three major publications: the first, a description of the clinical protocol for microbiome specimen sampling; the second, a catalog of the HMP and its data products; and the third, a large-scale, global analysis of the healthy adult cohort study using the data sets described above. These results describe the range of normal microbial variation among healthy adults in a Western population. The microbiota differed among individuals when communities were analyzed at several taxonomic levels (genera, species, strains), or even when individual loci and genomic islands were considered. Differences in microbial membership were even greater between body sites than between individuals; as one example, even adjacent oral surfaces separated by only millimeters or less in distance within the same subject exhibited strikingly different community structures. However, even though

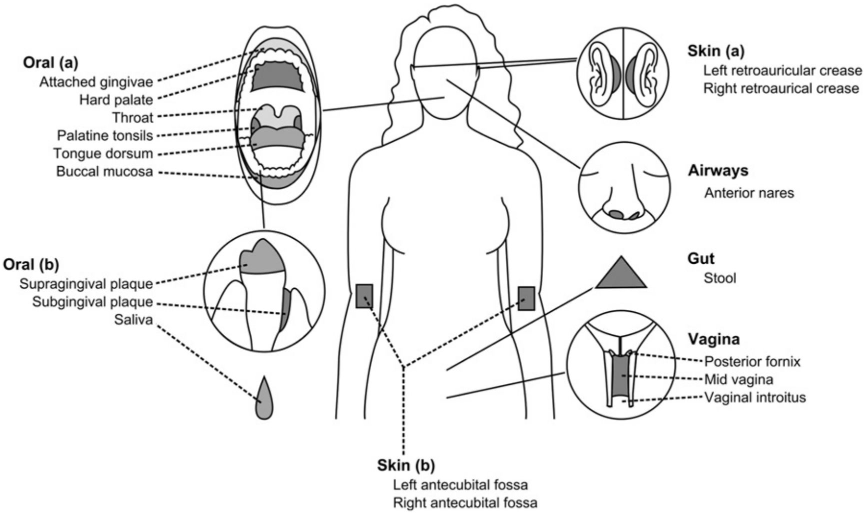

Figure 2. Schematic of the Body Sites Sampled for the HMP Healthy Adult Cohort Study
Three hundred individuals were sampled across a total of 18 body sites in 5 major body regions to collect microbiome specimens for sequence analysis. The oral cavity, skin, airway, and gastrointestinal tract regions were sampled in males, and the vagina was additionally sampled in females as the fifth major body region for the study. Eight distinct soft and hard surface sites were sampled in the oral cavity with saliva representing the ninth oral site, four sites were sampled on the skin, and three sites were sampled in the vagina. The airway was represented by a pooled sample of the anterior nares, and the distal gut tract region was represented by one sample of stool. Over 11,000 primary specimens for sequencing were collected in this study. (Figure adapted from Sitepainter visualization tool figure from Knight, Perrung, and Gonzalez, University of Colorado; tool available at http://www.hmpdacc/sp).

community structure varied greatly, the potential metabolic capabilities encoded in these communities' metagenomes were much more constant, both among body sites and between individuals. That is, although the microbiota in the healthy microbiome varied among individuals, the functions the microbiota are equipped to carry out remain remarkably stable within each body site. In addition to these findings, approximately 15–20 companion papers addressing specific questions about microbial prevalence, ecology, metabolism, and signaling functions, and the computational and analytical tools developed for the healthy cohort data are in review and will accompany the three main consortium publications.

Demonstration Projects

A third key resource from this activity is the Demonstration Projects, designed to evaluate microbiome characteristics in disease states with putative microbiome associations. Many complex diseases appear to have a microbiome component, and these projects were designed to characterize the microbiome in such cases in order to develop a reference data set of microbiome properties associated with specific disease and clinical phenotypes.

Eleven Demonstration Project studies have been launched to date, including six projects on microbiome-associated gastrointestinal diseases (Crohn's disease, ulcerative colitis, pediatric inflammatory bowel syndrome, neonatal necroticizing enterocolitis, and esophageal adenocarcinoma), three on urogenital conditions (changes associated with bacterial vaginosis, reproductive history, sexual history, and circumcision), and two on microbiome-associated skin diseases (eczema and psoriasis). The age groups across these studies range from birth to over 50 years old, and the size of some study cohorts approach 500 individuals. Almost all of the studies include 16S and shotgun metagenomic sequencing, and some also include functional data from the microbiome such as gene expression, microbial community proteomics, or metabolomics. Details of each project's purpose, experimental design and scope, data quality policies, anticipated analyses, and data release plans can be found in marker papers at Nature Precedings (http://precedings.nature.com/collections/human-microbiome-project).

Early results from some of these studies are showing that a characteristic microbiome community appears to be associated with each specific disease: three examples where this has been reported are neonatal necroticizing enterocolitis (NEC) (Wang et al., 2009), gastric esophageal reflux disease (GERD) (Yang et al., 2009), and pediatric irritable bowel syndrome (IBS) (Saulnier et al., 2011). These microbial signatures often include both taxonomic markers, such as altered overall community composition, and functional markers, such as differences in specific proteins identified from within the total protein content (i.e., metaproteome) of the disease-associated microbial community, providing a potential suite of markers for future development of diagnostic or prognostic applications.

In addition, some microbial biomarkers may precede the disease state, possibly allowing earlier detection and intervention. For example, GERD is characterized by a series of diseases, starting with reflux esophagitis, progressing to Barrett's esophagus in about 20% of cases, and, in rare cases, proceeding to the development of esophageal adenocarcinoma. The Pei/Nelson foregut microbiome esophageal adenocarcinoma study has found that those patients who go on to develop adenocarcinoma appear to have very similar foregut microbiomes to those patients with the intermediate stage of the disease (Barrett's esophagus), suggesting that microbiome composition in Barrett's esophagus may be one potential precursor marker for the cancer (Yang et al., 2009). In this case, it may be possible to develop diagnostic biomarkers for adenocarcinoma far before the cancer develops. The Demonstration Projects data sets and descriptions can be found at NCBI Bioprojects (http://www.ncbi.nlm.nih.gov/bioproject/46305).

Future Directions for Human Microbiome Research

We are at a pivotal point in the field of human microbiome research. The HMP has provided an extensive resource of datasets, computational tools, clinical methods, and scientific approaches to the study of the human microbiome. Here, we suggest a few key research areas to move the field forward.

We do not yet have a mechanistic understanding of the basic factors that regulate microbiome development during foundational events early in a person's life. We do know that the microbiome is acquired anew from the environment at birth (Dominguez-Bello et al., 2010) and that during early years, the maturing immune system (Round and Mazmanian, 2009), diet, and the assembling microbial community interact to establish the microbiome (Koenig et al., 2011). However, the roles of the source inoculum in the maturing microbiome are not yet clear—nor are those of the host immune system in regulating colonization by specific members of the microbiota, of the microbiome in regulating host tissue development, of the microbiome in resisting colonization by new microbes, or those of breastmilk and the infant diet on early microbial colonization of the gut and other body habitats.

Further, it appears that the microbiome retains much of its dynamic quality throughout life and is highly personalized (Costello et al., 2009), indicating that we may not yet understand what constitutes a healthy or, more generally, normal microbiome, particularly over the full lifetime of an individual (Claesson et al., 2011). Microbial transmission might occur environmentally, internally within and among body habitats, or epidemiologically through the interactions of human and other vertebrate hosts. We do not yet understand the significance of interactions between early microbiome events and microbiome function and change throughout life. For example, some studies have suggested that a disturbed microbiome at infancy, e.g., through antibiotic use, may predispose one to allergies later in life (Bisgaard et al., 2011); other disorders (e.g., Crohn's disease, asthma, type 1 diabetes, multiple sclerosis, celiac disease, and others) have also been associated with a disturbed, altered, or impoverished microbiome in infancy.

In addition, host genetics, culture, and ancestry remain largely unexplored areas of interaction with the human microbiome. To date, with some exceptions (e.g., De-Filippo et al., 2010), most microbiome studies have not included significant populations of non-European ancestry to capture the breadth of factors that may contribute to microbiome assembly or stability. Further, no major microbiome study has included host genetics; it is imperative we begin to consent volunteers as broadly as possible in our efforts to include all of the factors that may contribute to the microbiome. Results from genome-wide association studies over the past decade demonstrate it will be crucial to leverage multiple large populations with well-understood structure and prospectively determined phenotypes in order to derive robust genetic associations with quantitative microbiome traits. Given that variation in the microbiome appears to be far greater than human genetic variability, repeated studies in each target population will be needed to identify keystone microbiome signatures against a complex and contextually dependent background.

Though the microbiome of each region of the body is unique to and important for the health of the host, the gastrointestinal microbiome may arguably be considered the "cardinal microbiome," as it is the community that most directly interacts with the host immune system (Round and Mazmanian, 2009), as well as contributing to food digestion and energy supply for host cell metabolism. These functions also include regulation of the host and of other microbiomes through signaling molecules and metabolites that circulate throughout the body, although the extent to which such functions might also be performed by local microbiomes is not yet clear. There has been considerable interest in understanding whether the human gut microbiome can be categorized into predominant types, or "enterotypes"; patterns of variability are reproducible across human populations (Arumugam et al., 2011), although this variability appears to be associated with long-term diet (De Filippo et al., 2010) but not short-term diet (Wu et al., 2011). A concerted effort to study the relationship between diet and the microbiome in human populations would be an important foundational effort, as would an investigation of the systemic role of the gut community and how it interacts with the host tissues and with other microbial communities across the body. As the gut microbiome in particular appears to be amenable to manipulation (Manichanh et al., 2010), an ecological understanding based on these studies may hold the potential for disease treatment and, perhaps most importantly, prevention through microbiome therapeutics.

Just as advances in sequencing technologies paved the way for the characterization of microbiome composition, new technologies are now needed to study microbiome function and its interactions with the host. These new resources should include technology development for high-throughout methodologies such as metatranscriptomics, metaproteomics, and metabolomics as well as new model systems for the study of microbiome function. Opportunities for collaboration in the development of some of these new resources may now be on the horizon. For example, the NIH will soon be initiating a new program in metabolomics to include technology development (http://commonfund.nih.gov/Metabolomics). This activity would provide an ideal opportunity to collaborate in the development of methodologies for microbiome metabolomics, as we will specifically need to move beyond composition to an understanding of microbiome function and its interaction with the host if the microbiome is to be fully integrated into the study of human health and disease.

Furthermore, cohort studies could serve as one platform from which numerous investigations could address microbiome development, variability of the microbiome across populations, temporal changes, and functional properties in response to diet or disease. With proper consent and privacy safeguards in place, the genome sequences of the cohort members themselves would provide invaluable information for integration with the microbiome assays. Initial opportunities in this area are already becoming available; for example, a collaboration between the NIH, the CDC, and the EPA is conducting the National Children's Study (http://www.nationalchildrensstudy.gov), which will follow 100,000 children from birth to 21 years. This cohort study is designed to examine the effect of the environment on the health and development of children. Young Lives, a British-led international study of childhood poverty (http://www.younglives.org.uk) is following 12,000 children in four developing countries—Ethiopia, India, Peru, and Vietnam. These and other cohort studies could provide ideal frameworks from which to analyze the microbiome from birth in diverse populations.

It is perhaps useful to recall here that the microbiome is not inherited but acquired anew each generation. This suggests that the reservoir for the commensal and beneficial microbes (as well as pathogens) that contribute to the microbiome may be found not only in other hosts but also in the environment. Appreciation for the continuum that exists between the host and the environment is growing (http://www.onehealthinitiative.com) and should serve as a guiding principle in future studies of the microbiome. In fact, opportunities may already be on the horizon for initiatives that bridge the host with the environment. For example, joint agency activities such as the newly formed ‘‘USDA-NSF-NIH Research Coordination Working Group,’’ which is focusing on possible programs in the areas of obesity, nutrition, microbiome, and plant genomics, could place these initiatives in the appropriate environmental framework.

Finally, for these studies to benefit the broadest community, these activities will require a flexible and user-friendly infrastructure that links diverse aspects of the microbiome including microbial composition and function, and host phenotype and genotype. All of these must be associated with appropriate, ready-to-use computational and analytical tools that are accessible to a broad spectrum of microbiological, ecological, and bioinformatic expertise. In fact, it will be the routine access and use of this network of data and tools that will move this field into the clinical realm. Diverse populations should be included in all of these studies in order to circumscribe and relate the fundamental properties of the microbiome with other features of the human hosts themselves. High-throughput methodologies to measure microbiome function, including interactions among the microbes, among microbial communities, and between microbe and host—in conjunction with large cohort studies and all supported by a well-designed infrastructure—will establish the needed resources and data for future research and application of the microbiome in health and in disease.

ACKNOWLEDGMENTS

Thanks are due to Chris Wellington and Shaila Chhibba, NHGRI/NIH, who developed Figures 1 and 2. I am also grateful to Dirk Gevers, Broad Institute; Curtis Huttenhower, Harvard University School of Public Health; Rob Knight, University of Colorado-Boulder; and Pamela McInnes, National Institute for Dental and Cranial Research, for making valuable contributions to this commentary.

REFERENCES

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D.R., Fernandes, G.R., Tap, J., Bruls, T., Batto, J.M., et al. MetaHIT Consortium (2011). Nature *473*, 174–180.

Bisgaard, H., Li, N., Bonnelykke, K., Chawes, B.L., Skov, T., Paludan-Muller, G., Stokholm, J., Smith, B., and Krogfelt, K.A. (2011). J. Allergy Clin. Immunol. *128*, 646–652.

Claesson, M.J., Cusack, S., O’Sullivan, O., Greene-Diniz, R., de Weerd, H., Flannery, E., Marchesi, J.R., Falush, D., Dinan, T., Fitzgerald, G., et al. (2011). Proc. Natl. Acad. Sci. USA *108* (*Suppl 1*), 4586–4591.

Costello, E.K., Lauber, C.L., Hamady, M., Fierer, N., Gordon, J.I., and Knight, R. (2009). Science *326*, 1694–1697.

De Filippo, C., Cavalieri, D., Di Paola, M., Ramazzotti, M., Poullet, J.B., Massart, S., Collini, S., Pieraccini, G., and Lionetti, P. (2010). Proc. Natl. Acad. Sci. USA *107*, 14691–14696.

Dominguez-Bello, M.G., Costello, E.K., Contreras, M., Magris, M., Hidalgo, G., Fierer, N., and Knight, R. (2010). Proc. Natl. Acad. Sci. USA *107*, 11971–11975.

Koenig, J.E., Spor, A., Scalfone, N., Fricker, A.D., Stombaugh, J., Knight, R., Angenent, L.T., and Ley, R.E. (2011). Proc. Natl. Acad. Sci. USA *108* (*Suppl 1*), 4578–4585.

Manichanh, C., Reeder, J., Gibert, P., Varela, E., Llopis, M., Antolin, M., Guigo, R., Knight, R., and Guarner, F. (2010). Genome Res. *20*, 1411–1419.

Nelson, K.E., Weinstock, G.M., Highlander, S.K., Worley, K.C., Creasy, H.H., Wortman, J.R., Rusch, D.B., Mitreva, M., Sodergren, E., Chinwalla, A.T., et al. Human Microbiome Jumpstart Reference Strains Consortium (2010). Science *328*, 994–999.

Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., Schloss, J.A., Bonazzi, V., McEwen, J.E., Wetterstrand, K.A., Deal, C., et al. NIH HMP Working Group (2009). Genome Res. *19*, 2317–2323.

Round, J.L., and Mazmanian, S.K. (2009). Nat. Rev. Immunol. *9*, 313–323.

Saulnier, D.M., Riehle, K., Mistretta, T.A., Diaz, M.A., Mandal, D., Raza, S., Weidler, E.M., Qin, X., Coarfa, C., Milosavljevic, A., et al. (2011). Gastroenterology. Published online July 8, 2011. 10.1053/j.gastro.2011.06.072.

Wang, Y., Hoenig, J.D., Malin, K.J., Qamar, S., Petrof, E.O., Sun, J., Antonopoulos, D.A., Chang, E.B., and Claud, E.C. (2009). ISME J. *3*, 944–954.

Wu, G.D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.-Y., Keilbaugh, S.A., Bewtra, M., Knights, D., Walters, W.A., Knight, R., et al. (2011). Science *334*, 105–108.

Yang, L., Lu, X., Nossa, C.W., Francois, F., Peek, R.M., and Pei, Z. (2009). Gastroenterology *137*, 588–597.