1

THE NIH HUMAN MICROBIOME PROJECT

LITA M. PROCTOR, SHAILA CHHIBBA, JEAN McEWEN,
JANE PETERSON, and CHRIS WELLINGTON

NHGRI/NIH, Bethesda, Maryland

CARL BAKER

NIAMS/NIH, Bethesda, Maryland

MARIA GIOVANNI

NIAID/NIH, Bethesda, Maryland

PAMELA McINNES and R. DWAYNE LUNSFORD

NIDCR/NIH, Bethesda, Maryland

1.1. INTRODUCTION

The human microbiome is the full complement of microbial species and their genes and genomes that inhabit the human body. The National Institutes of Health (NIH) Human Microbiome Project (HMP) is a community resource project designed to promote the study of complex microbial communities involved in human health and

The Human Microbiota: How Microbial Communities Affect Health and Disease,
First Edition. Edited by David N. Fredricks.

disease. The HMP has increased the appreciation for the features of the human microbiome that all people share as well as the features that are highly personalized. Host genetics, the environment, diet, the immune system, and many other factors all interact with the human microbiota to regulate the composition and function of the microbiome. As a scientific resource, the HMP has publically deposited to date or made available over 800 reference microbial genome sequences, hundreds of microbial isolates from the human microbiome, over 3 terabases (Tbp) of metagenomic microbial sequence, over 70 million 16S rRNA reads, close to 700 microbiome metagenome assemblies, over 5 million unique predicted genes, and a comprehensive bodywide survey of the human microbiome in hundreds of individuals from a healthy adult cohort. A number of demonstration projects are contributing a wealth of knowledge about the association of the microbiome with specific gut, skin, and urogenital diseases. Other key resources include the development of new computational tools, technologies, and scientific approaches to investigate the microbiome, and studies of the ethical, legal, and social implications of human microbiome research. This chapter captures the historical context of the HMP and other international research endeavors in the human microbiome, highlights the multiple initiatives of the HMP program and the products from this activity, and closes with some suggestions for future research needs in this emerging field.

1.2. GENESIS OF HUMAN MICROBIOME RESEARCH AND THE HUMAN MICROBIOME PROJECT (HMP)

It sometimes seems that research on the human microbiome blossomed overnight. However, the conceptual and technological foundations for the study of the human microbiome began to emerge before the 1990s and can be found within many disciplines. Microbial ecologists who studied microorganisms and microbial communities in the environment recognized early on that most microorganisms in nature were not culturable and so developed alternate approaches to the study of microbial communities. An early and broadly adopted approach for investigating microorganisms in the environment, based on the three-domain system for biological classification [1], was the use of the 16S ribosomal RNA gene as a taxonomic marker for interrogating microbial diversity in nature [2]. With the growth of non-culture-based, molecular techniques in the 1980s and 1990s for study of environmental microorganisms and communities, some medical microbiologists turned these tools to the human body and found far greater microbial diversity than expected, even in well-studied sites such as the oral cavity [3–5].

In the infectious disease field, recognition was growing that many diseases could not satisfy Koch's postulates as the pathogenesis of many of these diseases appeared to involve multiple microorganisms. The term *polymicrobial diseases* was coined to describe those diseases with multiple infectious agents [6]. We now recognize that many of these formerly classified polymicrobial diseases, such as abscesses, AIDS-related opportunistic infections, conjunctivitis, gastroenteritis, hepatitis, multiple sclerosis, otitis media, periodontal diseases, respiratory diseases, and genital infections, are associated with multiple microbial factors, that is, with the entire microbiome. In an essay on the history of microbiology and infectious disease, Lederberg [7], who coined the term *microbiome*, called for "a more ecologically informed metaphor" to understand the relationship between humans and microbes.

The field of immunology was also undergoing its own revolution with the recognition that the innate and adaptive immune systems not only evolved to eliminate specific pathogens but are also intimately involved in shaping the composition of the commensal intestinal microbiota [8–10]. Recognition was also growing in this field that the microbiota is involved in regulating gut development and function [11,12].

Another key catalyst for discussions about the inclusion of the microbiome in the study of human health and disease was the publication of the first drafts of the human genome sequence. Relman and Falkow [13] noted on this occasion that a "second human genome project" should be undertaken to produce a comprehensive inventory of microbial genes and genomes associated with the human body. Lead by Davies [14], they renewed a call for considering the role of the human-associated microorganisms in development and in health and disease. Also, by 2005 or so, as sequencing costs began to drop, sequencing technology offered the opportunity to consider extensive surveys of the microbial communities associated with the human. Early human studies focusing on the most complex of human microbiomes, the digestive tract [15,16], demonstrated the tremendous complexity as well as the functional potential of the human microbiome.

The time appeared right to undertake a comprehensive study of the human microbiome—the full complement of microbial species and their genes and genomes that inhabit the body. A meeting, organized by the French National Institute for Agricultural Research (INRA), of European, North American and Asian scientists and government agency and private-sector representatives was convened in Paris in 2005 to discuss how to approach such a comprehensive study. This 2-day meeting covered a broad range of topics, including sequencing all of the bacteria in the human microbiome, the impact of the human microbiome on the study of health, and the possible structure of a human digestive tract microbiome program. Recommendations from this first international meeting included the formation of an International Human Microbiome Consortium and an agreement to release data rapidly, share data standards, and develop reference datasets (`http://www.human-microbiome.org/fileadmin/user_upload/Paris-recommendations.pdf`). Around this same time, the National Academy of Sciences published a report on metagenomics [17] (`http://books.nap.edu/catalog.php?record_id=11902`), which highlighted this new discipline with its focus on the combination of genomics, bioinformatics, and systems biology to study microbial communities in nature; this report also informed the scientific community of the potential of this new discipline. The Paris meeting was followed by several other international meetings in 2007 and 2008.

These discussions led to the formation of the European Commission's call for studies on human metagenomics. The NIH also invited community comment during this incubation period. A number of white papers identified specific needs for the field that included a reference microbial genome sequence catalog, animal models for microbiome studies, benchmarking studies for the analysis of 16*S* rRNA and microbiome metagenome sequencing, computational tools for the field, and considerations of the ethical aspects of human microbiome research. Pilot projects to develop protocols for sequencing the human microbiome were begun by the NIH National Human Genome Research Institute (NHGRI) in mid-2007. The

NIH Common Fund–supported Human Microbiome Project (HMP) was formally launched in late 2007 with the intent to produce a number of major community resources: a reference catalog of microbial genome sequences, a large cohort study to survey microbiomes across the human body in healthy adults, a suite of demonstration projects to examine correlations of changes in the microbiome with disease, and the computational tools to analyzing microbiome metagenomic sequence data (http://commonfund.nih.gov/hmp/). Funding of the Metagenomics of the Human Intestinal Tract (MetaHIT) program began in 2008, which included scientific partnerships across eight European countries (http://www.metahit.eu/). Other large-scale efforts in human microbiome research emerged in close order around the world and include, among others, the NIH HIV Lung Microbiome Project, the Gambian Gut Microbiome Project, the INRA French/China program MicroObes, the Canadian Human Microbiome Initiative, the Australian Jumpstart Human Microbiome Project, and the Korean Twin Cohort Microbiome Diversity project.

1.3. GUIDING PRINCIPLES, STRUCTURE, AND INITIATIVES OF THE HMP PROGRAM

1.3.1. HMP Guiding Principles and Creation of a Community Resource Project

The Human Microbiome Project was envisioned as a community resource program. A community resource program is 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). It was recognized that the metagenomic and associated metadata from human microbiome research are unique research resources. In order to establish and serve as a community resource, the guiding principles for the HMP included rapid data release into public databases. These follow the guiding principles that were created for the Human Genome Project and have been used for all large genome projects at NIH (https://commonfund.nih.gov/hmp/datareleaseguidelines.aspx).

At the same time, it was expected that users of the prepublication data would acknowledge the scientific contribution of the HMP data producers by following normal standards of scientific etiquette and fair use of unpublished data. These standards were outlined in the 2003 Fort Lauderdale agreement (http://www.genome.gov/10506537) and further elaborated in the 2009 Toronto meeting agreement (Toronto International Data Release Workshop Authors, [18]; doi: 10.1038/461168a). An HMP Research Network Consortium was established to enhance collaborative activities and to support large-scale analyses of the HMP data, the products of which would contribute to the overall community resource. A consortium agreement, signed by all members outlined the request to acknowledge the data producers’ contributions. New consortium members, nominated by existing consortium members, are asked to agree to the consortium statement. A marker paper that described the HMP and its data release policy was published (NIH HMP Working Group, [19]; doi 10:1101/gr096651.109) and serves as an outline of the large-scale analyses that the HMP Consortium is undertaking.

In addition, a data use agreement was drafted to provide guidance for users of the prepublication data from the larger community. The data use agreement, posted on the DACC website (http://hmpdacc.org/resources/data_browser.php), reiterated the Fort Lauderdale and Toronto meeting guidelines and also provided guidance on how publications that use HMP data should acknowledge and cite the HMP Consortium and the NIH as a source of the data. Finally, an agreement was made that all reagents, such as the reference microbial strains to be sequenced, should be deposited in appropriate repositories.

For the healthy cohort study, it was recognized that whole-genome shotgun sequencing (WGS) of nucleic acid extracts would capture various amounts of the human subject genome sequence, depending on the amount of human tissue collected during the microbiome sampling procedure. It was decided that the human genome sequence would not be made publically available but that the research community, with appropriate authorization, should have access to human subject data for research on the human microbiome. The NIH National Center for Biotechnology Information (NCBI) database of genotypes and phenotypes (dbGaP: http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/about.html) was adopted at the public database for the HMP clinical metadata and sequence data (http://www.ncbi.nlm.nih.gov/gap?term=Human%20Microbiome%20Project). The dbGaP has two levels of access—open and controlled—in order to regulate the distribution of the sequence and health information of the study volunteers. Open access contains publically accessible data. Controlled access requires approval by a NIH Data Access Committee (DAC) for legitimate microbiome research purposes.

The WGS sequence data were computationally filtered to remove the human subject sequence before these data were deposited in the open access portion of the sequence read archive (SRA) in dbGaP. The criteria and procedure for removing human sequence is described later in this chapter. Clinical patient metadata were deposited in the controlled access portion of dbGaP. The procedures for requesting access to the controlled data can be found at the following website: https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?login=&page=login. The HMP-targeted 16*S* ribosomal RNA gene sequence data were deposited in the open access SRA in dbGaP as there is no human sequence associated with these data.

Whereas other national and international programs focused on the microbiome of a specific body site, the HMP decided to survey the microbiomes of multiple body sites in healthy adults to produce baseline data for healthy microbiomes, develop a catalog of microbial genome sequences of microbiome reference strains, and evaluate the associations of microbial communities with specific diseases. A Data Analysis and Coordination Center (DACC) was created to manage the data from the sequencing activities, process the sequence data to consortium agreed-on standards for further analysis, coordinate the data analysis activities in the consortium, and serve as a portal for the scientific community to access the datasets, tools, and other resources generated by the program. In addition, initiatives in technology development; computational tools; and the ethical, legal, and societal implications of microbiome research were created to support the field. There are three sources of information about the HMP program. The NIH Common Fund website provides an overview of the main initiatives in the program (https://commonfund.nih.gov/hmp/). The NCBI Bioprojects pages describe the data types produced in the program (http://www.ncbi.nlm.nih.gov/bioproject/43021). There are four projects listed by NCBI under the HMP umbrella based on the four data types produced: (1) the 16S rRNA gene and (2) whole-genome shotgun metagenome datasets produced from the healthy adult cohort study, (3) the reference strain microbial genome

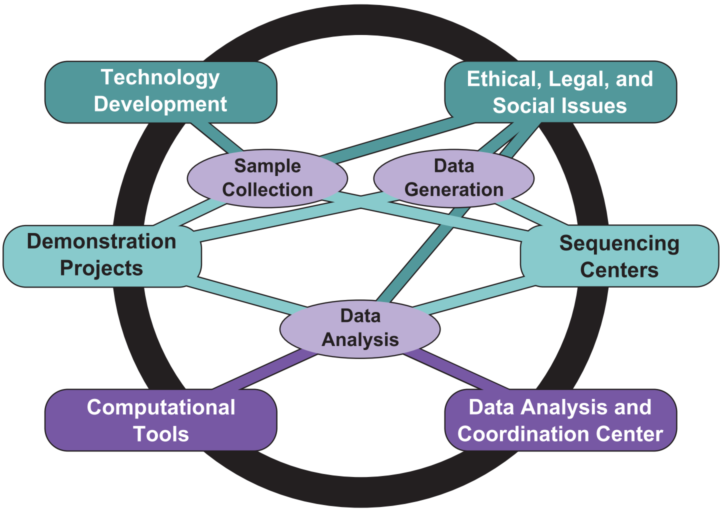

Figure 1.1. Conceptual diagram of the NIH Human Microbiome Project. The HMP program is comprises of six formal Initiatives, shown around the circle and include technology development, ethical, legal, and social issues; sequencing centers, the data analysis and coordination center; computational tools; and the demonstration projects. These initiatives interact through the activities of the ≥200-member HMP research network consortium, which also includes 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 extract sample preparation from the specimens in the healthy cohort study and in the demonstration projects; (2) data generation, which includes 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.

sequence dataset, and (4) the datasets produced in the individual demonstration project activities. Finally, the DACC provides an extensive web resource that describes the datasets produced by the program, the derivative datasets developed by the HMP Working Groups, the suite of computational tools developed for the analyses, and other contextual information about the HMP (www.hmpdacc.org). A conceptual diagram of the initiatives within the HMP program and their interrelationships and how the initiative research teams and the research consortium interacts provides another view of this program (Figure 1.1). Using this figure, the HMP program is described below.

1.3.2. HMP Large-Scale Sequencing Centers

In order to establish scientific approaches and protocols for the Human Microbiome Project and to be able to sequence very large numbers of HMP samples, the first initiative in the HMP included the support of four large-scale sequencing centers:

Baylor College of Medicine (http://www.hgsc.bcm.tmc.edu/), the Broad Institute (http://www.broadinstitute.org/), the J. Craig Venter Institute (http://www.jcvi.org/), and Washington University at St. Louis (http://genome.wustl.edu/). These sequencing centers are responsible for (1) developing the protocols, (2) sequencing microbiome samples from a baseline adult population of healthy human subjects and reference strain microbial genomes, (3) analyzing the microbiome sequence data, (4) providing computational approaches, and (5) contributing to the analysis of the healthy subject microbiome data. These centers are also responsible for supporting the sequencing activities in several of the demonstration projects (discussed in further detail below). Further, the sequencing center project investigators provided oversight for data production objectives and goals.

1.3.3. Data Coordination and Analysis

Data Analysis and Coordination Center (DACC)

The Data Analysis and Coordination Center (DACC) was established in order to facilitate data deposition and to coordinate processing and analysis of the very large datasets produced by the HMP (www.hmpdacc.org). In order to support HMP activities, the DACC established a human microbiome database and developed a comprehensive analysis pipeline. The DACC plays a major role in the establishment, coordination, and support of an HMP Research Network Consortium, which was made up of members of the microbiome community interested in participating in the analysis of the large HMP dataset as well as the various workgroups, which focus on specific tasks. The DACC hosts an electronic collaboration site where data analyses, workgroup discussions, and publication drafts can be shared within the consortium. The DACC supports extensive community outreach and training activities. For example, the DACC website includes the project catalog of the reference genome sequences, a browser that includes links to the datasets from the benchmarking activities, the healthy cohort study, the demonstration projects, and many of the bioinformatics and computational tools that are used in the project.

HMP Workgroups

It was recognized that large-scale analyses of these new and complex datasets, particularly of the healthy adult cohort data (discussed below) would add value to the resources emerging from the program. This would require the efforts of a large group of scientists. Thus, the Data Analysis Working Group (DAWG) was formed and consisted of a combination of HMP grantees as well as individuals in the scientific community with specific expertise in the analysis of metagenomic data, all who joined the Research Network Consortium. During the 2 years of active data processing, analysis, and interpretation of the healthy cohort dataset, this group met weekly on conference calls, held biannual research network consortium meetings; held a virtual jamboree, which was a 1-day online meeting to discuss the healthy cohort data analyses with experts in the microbiology and diseases of the body sites; and exchanged computational tools, analyses and draft manuscripts through a consortium-managed electronic resource.

At one time or another, there were over 200 members of the 20 workgroups tackling specific tasks; several of these workgroups also work together toward larger goals or provide oversight and guidance toward major program objectives. For example, the Strains Working Group works with the Annotation and the Finishing Working Groups to coordinate the selection, sequencing, and annotation of the reference strains for the project. The Data Generation and Processing Working Groups works with the Data Release Working Group to agree on common processed datasets for downstream analysis. As the consortium is working together to analyze these datasets for major publications and for companion papers, each member of the consortium agrees to guiding principles on data use and HMP consortium acknowledgment in publications.

1.3.4. Reference Strain Microbial Genome Sequences

The HMP sought to create a public reference dataset of microbial (primarily from bacteria but also from some archaea, viruses, bacteriophages, and eukaryotic microbes) genome sequences of microorganisms collected from the major body sites. The goal was to create a catalog of genome sequences from 3000 bacterial strains and as many viral/phage and eukaryotic microbial strains as possible. The microbial genome sequence dataset is intended to provide a reference for the interpretation of 16S rRNA sequences and to serve as scaffolding for assemblies of the metagenomic sequences derived from microbiome samples. As an extension of this public resource, cultures of sequenced strains that were donated from personal laboratory collections were deposited at the HMP Repository with the NIAID Biodefense and Emerging Infections Research Resource Repository (BEI: `http://www.beiresources.org/`). Approximately 100 of these cultures that are expected to be in high demand will be in a "shelf-ready" state and will be available immediately to the scientific community. Another several hundred cultures are archived and can be prepared once requests for specific cultures are received by BEI.

At project inception, guidelines for inclusion of strains in the microbial reference genome dataset were established and focused on aspects of each nominated organism including (1) its phylogeny and uniqueness, (2) its established clinical significance, (3) its abundance or dominance in a body site, (4) whether identical species were found in different body sites, and (5) whether there was an opportunity to explore pangenomes (pangenome, the core genome containing genes present in all strains of a microbial species plus other genes present in one or more strains of the species) (`http://www.hmpdacc.org/doc/sops/reference_genomes/strains/StrainSelection.pdf`). Microbiologists and clinicians with body-site-specific expertise were consulted to identify and provide, when possible, strains for sequencing based on these guidelines. In addition, the HMP has continued to solicit feedback and strain nominations from the global community and hosts a web portal for this purpose (`http://www.hmpdacc.org/outreach/feedback.php`). All nominations are discussed and decided on by the Strains Working Group, representing all sequencing centers, DACC and NIH.

Microbiome strains were contributed by investigators in the field from their personal laboratory collections or were identified from public culture collections, including the American Type Culture Collection (ATCC), the German Collection of Microorganisms and Cell Cultures (DSMZ), the UK National Collection of Type Cultures (NCTC), the Belgian Co-ordinated Collections of Microorganisms (BCCM) as well as the Culture Collection from University of Goteborg (CCUG)

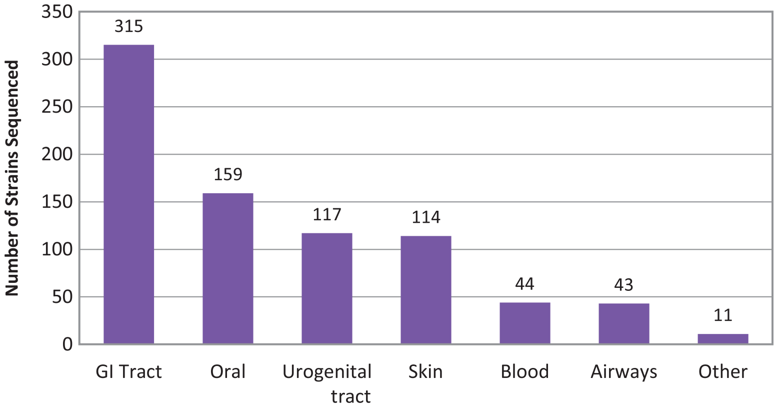

Figure 1.2. Distribution of HMP reference sequence bacterial strains by major body site. Note that additional body sites (blood) outside of the typical HMP major body sites served as sources of the isolates. "Other" refers to isolates collected from other, miscellaneous body sites. (Data and figure courtesy of Drs. Heather Huot-Creasy, DACC and Ashlee Earl, Broad Institute. Additional details are available at http://www.hmpdacc.org/refernce_genomes/statistics_specific.php.)

and the Biological Resource Center of Institut Pasteur (CIP). Workgroups of different body sites experts were convened to identify the sources of strains to be sequenced. These microbial strains came from a wide variety of body sites, with GI tract samples contributing about a third of the strains and oral, skin, and urogenital samples contributing approximately equal numbers of strains. The airway, blood, and additional body site samples make up the remaining sources for these strains (Figure 1.2). A publication documenting the analysis of the first 178 microbial isolates was published (viz., the Human Microbiome Jumpstart Reference Strains Consortium [20]). This analysis described 550,000 predicted genes, 30,000 of which are novel.

As of this writing, over 1300 strains have been sequenced (~800) or targeted for sequencing (~500) by the four sequencing centers (http://www.hmpdacc-resources.org/hmp_catalog/main.cgi?section=HmpSummary&page=showSummary). This list comprises primarily bacterial strains, although some bacteriophages, eukaryotic microbes, and methanogenic archaea have been included. The sequences are available in GenBank. The Strains Working Group made a decision to finish the completed sequences to various levels; approximately 30 are finished genome sequences, and most are at the high quality draft level of finishing [21].

Because only a fraction (current estimate ~60%) of the human-associated microbes are in culture and available for sequencing, a technology development initiative aimed at isolating uncultivable microorganisms was created. This program included support for innovative cultivation techniques to isolate new strains from the body sites and the application of single cell genomics methodologies to reach this project goal.

In order to guide this effort, the Strains Working Group has conducted an analysis of the healthy cohort 16S data to develop a priority list of the top 100 most desirable bacterial strains to target for sequencing. The approach used to identify new or novel taxa that have not yet been sequenced was to select 16*S* sequences for all of the body sites that had less than 90% identity to already sequenced strains and were found in at least 30% of all samples from a particular body site. Then, using the 16*S* data, the body sites were identified that contained most of the strains that had not yet been sequenced; this analysis resulted in a little over 100 targeted strains. This analysis showed that 73 of the 100 desired strains were located in the oral cavity and 30 were located in the gut; the remainder were evenly distributed across the other three major body sites. These data are being used to guide the technology development teams in their sample sorting efforts and in their searches for novel strains. In addition, collaborations between the demonstration project teams and the technology development teams are endeavoring out to identify tissue types and samples that could serve as material for isolating new strains for cultivation or cells for further analysis.

1.3.5. Healthy Adult Cohort Study of Multiple Microbiomes

The third initiative of the HMP represents the largest cohort study to date of the microbiomes of the multiple body habitats of healthy adults. There have been differences in the terms used to describe the microbiome body habitats sampled for this study. In this chapter, we will consistently refer to the oral, skin, nares, gut, and vagina areas as the major body *sites*. Specific areas within each major body site will be called body *subsites*. As these volunteers were clinically evaluated and determined to be healthy, this study is typically called the *healthy adult cohort study*, and the goal of the study was to collect and analyze minimally disturbed microbiomes. The study can be broken down into three components: the clinical phase, the sequencing phase, and the data analysis phase.

Clinical Phase

Experts in clinical research and ethical issues advised on the inclusion and exclusion criteria and on the consent forms developed for the study. Extensive exclusion criteria for the selection of healthy volunteers were developed and were based on a combination of health history (particularly systemic disorders such as hypertension, cancer, autoimmune disorders), use of antibiotics, probiotics or immunomodulators, and body mass index, as well as physical examination of the volunteers such as presence of skin lesions and oral and dental health status. It was common to find that these apparently healthy volunteers were not always "healthy" in all body sites. An example of this dichotomy was with the oral cavity, where otherwise healthy volunteers had dental caries that resulted because they were not eligible for enrollment until the dental disease was treated and the mouth determined to be healthy. Women were required to have a history of regular menstrual cycles.

The subjects were informed that their microbiome samples and microbiome sequence data would be coded to anonymize study participants, that controlled access databases would be used to store the clinical metadata and human genome sequence data, and that permission to use these data for microbiome research purposes would be regulated by the NIH Data Access Committee (DAC) to ensure that the data were being used properly. The volunteers consented to allow researchers to use their human sequence data for microbiome research but were assured that their identities would not be revealed to the researchers or to the public. The volunteers were also assured that all reasonable effort would be expended to separate their human sequence from the microbial sequence data before the microbial data were deposited in open access databases, which is open to all users of the database and does not require a DAC review.

A comprehensive clinical protocol was developed to ensure that minimally disturbed microbiomes were sampled. All of the body sites were directly sampled except for the digestive tract, in which stool served as a proxy for all distal gut regions. Saliva was collected from each subject at each visit. Blood and serum were also collected from each subject at the first visit, DNA was extracted from one aliquot of the blood for future whole-genome sequencing, and lymphocytes were harvested from a second aliquot and stored at –80°C for future preparation of cell lines. The human subject genome sequences, the bulk DNA, and the cell lines will be made into additional community resources. The blood, DNA extracts and serum is stored at the NHGRI Sample Repository for Human Genetic Research (Coriell Institute for Medical Research, Camden, NJ). The two clinical laboratories (Baylor College of Medicine and Washington University in St. Louis) extracted the DNA from the body site samples using the same commercial kit and standard operational procedures and distributed the DNA to the four institutions (Broad Institute, Baylor College of Medicine, J. Craig Venter Institute, and Washington University in St. Louis) carrying out the sequencing activities. The MoBio Powersoil DNA extraction kit (www.mobio.com) was selected after pilot studies to test different commercial extraction kits.

In this study, 300 adult volunteers were selected from a total of approximately 550 screened individuals. Approximately 20% self-identified as a racial minority and about 11% self-identified as Hispanic. The total pool of volunteers was split between two clinical sites: one in the southwestern United States (Houston, TX) and the other in the midwestern United States (St. Louis, MO). An equal number of adult men and women in the 18–40 year-old range were recruited for the study. The body mass index (BMI) range for the volunteers was 18–35. The mean blood pressure of the volunteers was 120/70, and the vast majority did not smoke. In addition, the majority of the volunteers self-reported as generally meat eaters and that they had been breastfed during infancy.

Enrollment and sampling of the volunteers commenced in December 2008 and were completed in October 2010. Of the 300 study participants, 279 were sampled twice and 100 were sampled a third time; the interval between the first and third samplings averaged approximately 10 months. A number of subsites within each body site were sampled, so there were 18 total subsites in five major body sites (oral, skin, nares, gut, and vagina for women) sampled; the oral body site had the largest number of subsites sampled (9) (Figure 1.3). Deposition of the full clinical metadata set in dbGaP was completed in February 2011, approximately 4 months after the last sampling was completed (http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000228.v3.p1). These metadata were released in editions because the clinical teams conducted continuous in-house analysis of the metadata to verify that there were no "identifiable" traits or combinations of traits in the metadata that could reveal a specific clinical subject. A manual of procedures detailing the clinical sampling protocol and criteria for sampling can be found at the dbGaP website (http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000228.v2.p1).

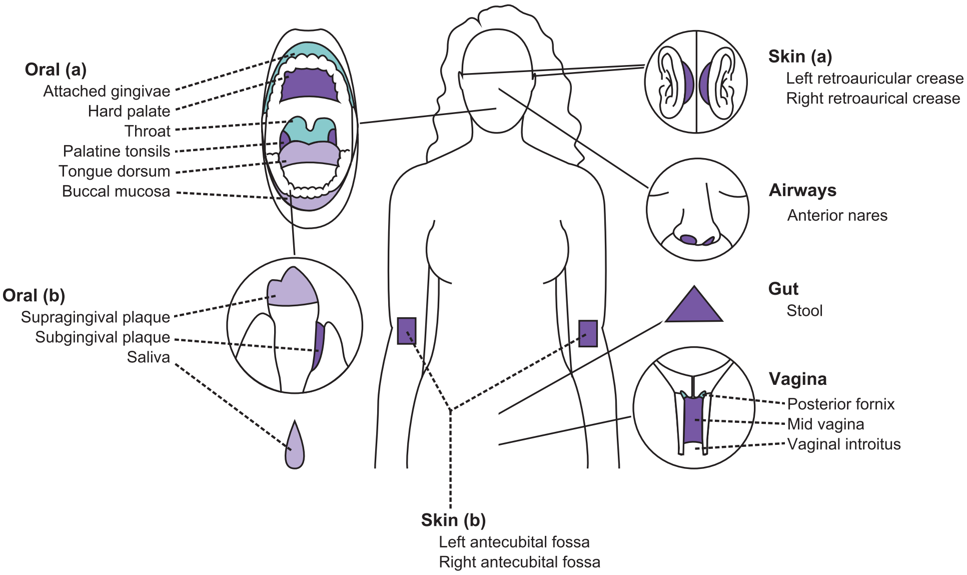

Figure 1.3. Schematic of the body sites sampled for the HMP healthy adult cohort study. Three hundred individuals were sampled across a total of 18 body subsites in five major body sites to collect tissue or body fluids for nucleic acid extraction and subsequent sequence analysis. The oral cavity, skin, airway, and gut sites were sampled in males, and the vagina was additionally sampled in females as the fifth major body site for the study. Eight distinct soft and hard surface subsites were sampled in the oral cavity with saliva representing the ninth oral subsite, four subsites were sampled on the skin, and three subsites 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. (This figure was adapted from the Sitepainter visualization tool figure, courtesy of R. Knight, M. Perrung, and A. Gonzalez, University of Colorado. Tool available at `www.hmpdacc/sp`.)

Sequencing Phase

As a part of the pilot project for this initiative, the four sequencing centers undertook a series of benchmarking exercises to determine appropriate protocols for sequencing the healthy human microbiome DNA and to compare consistency of results across the sequencing facilities. The group developed a mock microbiome community of a 22 bacterial species assemblage as a test specimen to evaluate DNA extraction, primer selection for library construction, and sequencing protocols. On the basis of these data, the group decided that primers for the variable region V3–V5 of the 16*S* rRNA gene would be used for the targeted 16*S* sequencing of all of the samples and, as needed, the V1–V2, V1–V3, and or V6–V9 regions would be targeted to amplify specific bacterial groups that do not amplify well with the V3–V5 primers. A manuscript describing the benchmarking exercise is in review.

As might be expected, DNA yield varied greatly across the body site samples (Table 1.1). As an example, stool yielded the greatest amount of total DNA (~9.5–21.0 ng/μL) whereas skin samples yielded the lowest, at 0.001 ng/μL. There were over 12,000 unique primary samples collected from the 300 subjects. Primary samples included samples collected in order to sequence the 16*S* rRNA gene or the metagenome of the microbiota as well as urine, blood, and saliva; 11,000 of those samples

TABLE 1.1 Range in DNA Yield (ng/μL) of Samples Collected from the Five Major Body Sites in the HMP Healthy Adult Cohort Study[a]

Body Site	DNA Yield (ng/μL)
GI tract (1)	9.49–21.08
Oral (9)	0.16–4.72
Nares (1)	1.05–2.10
Skin (4)	0.001–0.156
Vagina (3)	4.02–8.57

[a] Values in parentheses indicate the number of subsites sampled within each body site. Skin is reported to three places because overall yield was lower than that for other body site samples. Single swab (nares, vagina, skin, soft oral subsites) and curette (hard oral sites) samples and single stool subsamples (50–800 μL) were directly extracted using the MoBio PowerSoil kit and DNA extract eluted in 10 μL. DNA concentrations measured by fluorometric assay by the Baylor College of Medicine and Washington University clinical labs. DNA concentrations for each body site derived from three replicate extracts.

Source: Data and table courtesy of Dr. Joe Petrosino, Baylor College of Medicine.

were used for nucleic acid extraction. A majority of the samples were analyzed by targeted sequencing of 16S clone libraries with the Roche 454 sequencing technology. In addition, a fraction of the samples were analyzed by metagenomic whole genome shotgun sequencing using both the 454 and the Illumina GAII technologies.

The targeted 16S sequences and WGS sequences were deposited at NCBI databases by the participating sequencing centers. The 16S sequences were deposited in the open access sequence read archive (SRA) of dbGaP. The metagenomic sequences as well as the clinical metadata were deposited in the controlled access portion of dbGaP since they included information about the human subjects. Clinical metadata collected from these volunteers included elements such as gender, age, BMI, vital signs, vaginal pH, medical history, and other key information about the subjects. Since these WGS sequences contained human subject sequence, NCBI developed a computational tool, Bestmatch Tagger (BMTagger), to computationally filter the human sequence from the total sequence. The algorithm discriminates between human reads and microbial reads by comparing consecutive sequences of 18mer-length nucleotides found in the total sequence with those found in the human genome sequence and then includes an alignment procedure that finds all matches for any missing alignments. The human genome reference sequence used was the Genome Reference Consortium's most current refinement of the human genome sequence (GRCh36, `http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/index.html`) (S. Sherry, K. Rotmistrovsky, R. Agarwala, and NCBI, personal communication, 08/01/11). The filtered WGS sequence was deposited in the open access SRA as microbiome metagenomic sequence data.

Data Processing and Analysis Phase

In preparation for the data analysis phase, a group of scientists from the microbiome community, the sequencing centers, and the DACC as well as NIH staff were brought together to form a HMP Data Analysis Working Group (DAWG). As there was continuous sequence data production, the DAWG declared a data freeze on May 1, 2010 on a subset of the 16*S* rRNA sequence data and on July 1, 2010 on a subset of the WGS metagenomic sequence data in order to define a common, master dataset for the follow-on global analysis activities to be undertaken by the research consortium.

Of the >11,000 primary samples collected for the full study, the May 1 freeze targeted 16*S* rRNA data and included 5300 samples from 18 body subsites of 5 major body sites from 242 subjects (113 females, 129 males), and the July 1 freeze WGS data included 736 samples from 16 body subsites of 5 major body sites from 102 subjects. No third visit samples had been sequenced by the data freeze, but of the 242 subjects, a subset of 131 (~54%) included samples from two visits generally spaced by 6 months and up to a year between visits. These datasets included a total of ~74 million 16*S* rRNA reads. Once the contaminating human sequence was removed (which represented on average ~60% of the total sequence), a total of 3.5 terabases (Tbp) of metagenomic WGS sequence was generated for subsequent analysis.

For each major body site, the typical sequence generated from a sample ranged between 10.8 and 12.8 gigabases (Gbp) (average 12 Gbp). However, the ratio of microbial sequence reads to total sequence reads (i.e., the percent of human DNA sequence and sequence from other contaminating DNA) varied greatly across the body sites (Figure 1.4). The largest fraction of microbial reads to total reads was found in the gut samples (stool, ~98%). Nares, skin, and vaginal samples yielded about 10–25% microbial sequence reads to total reads.

Two kinds of metagenome assemblies were produced from the processed whole-genome shotgun data. The processed metagenome sequences were assembled using SOAPdenovo. Hybrid metagenome assemblies from processed Illumina and Roche 454 sequence reads were also produced using Newbler. These two kinds of metagenome assemblies were prepared in order to support different types of analyses. For example, the de novo assemblies were used for comparisons against the reference microbial genome sequences to determine microbiome community composition, and the hybrid assemblies were used for the reconstruction of metabolic modules and pathways inferred from the whole-genome shotgun data.

The DAWG and its various workgroups developed processed datasets in 2010–2011 that the DAWG agreed would serve as the common, master processed datasets for downstream data analyses. These finalized datasets include (1) 16*S* data that had been quality-controlled and processed to remove errors at agreed-on stringency levels, (2) metagenomic data mapped to a global list of microbial reference genome sequences from both the HMP sequencing efforts and microbiome reference strain data available in GenBank, (3) metagenomic assemblies produced either de novo or as hybrid assemblies, and (4) other such data products for use by the DAWG (Table 1.2). The approximate sizes of each data type are also shown.

The results from the global analysis of the healthy cohort study describe the range of normal microbial variation among healthy adults in a Western population. The microbial composition differed among individuals when these communities were analyzed at several taxonomic levels (genera, species, strains). Further, previous observations about community structure seem to be true for all of the major body sites examined in this study: the microbial communities grouped by body site and not by individual. In addition, there was great variability in microbial

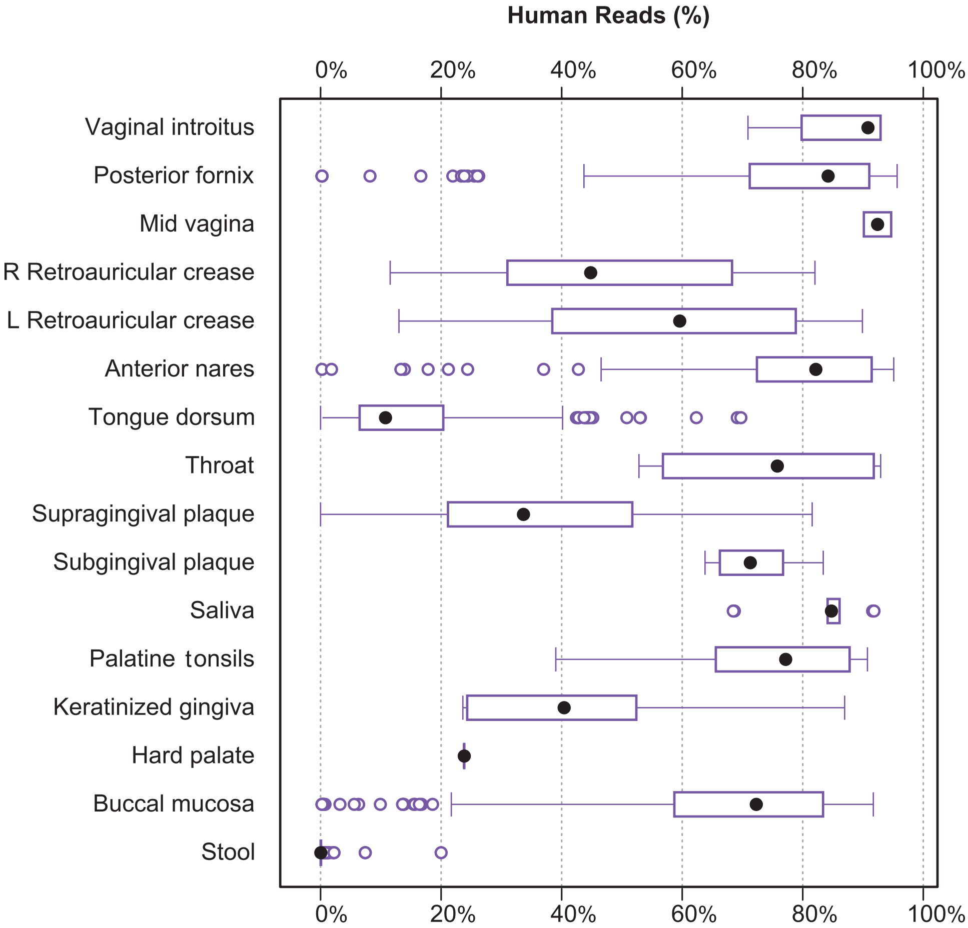

Figure 1.4. Percent human sequence reads in total sequences of whole-genome shotgun reads from HMP healthy cohort microbiome nucleic acid extracts. Boxplots represent the range in percent of human reads per body site (*x* axis) with black dot representing the mean. Body sites are listed on *y* axis. Note that the majority of samples had significant human contamination, at levels of ≥60% of total sequence. (Analysis and graph courtesy of Drs. Dirk Gevers and Katherine Huang of the Broad Institute.)

composition between subsites within a body site. As one example, even adjacent surfaces of the oral cavity separated by only millimeters or in closer proximity within the same subject exhibited strikingly different community structures.

Even though community structure varied greatly between body sites, the potential metabolic capabilities encoded in these metagenomes were much more constant, both among body sites and between individuals. Over 5 million unique genes were cataloged from the healthy cohort analysis. However, although the microbial community composition in the healthy microbiome varied among individuals, the predicted core functions that the microbiota are equipped to carry out remain remarkably stable within each body site, particularly for major metabolic pathways.

TABLE 1.2. Finalized Datasets[a] Used by HMP DAWG for Analysis of Healthy Cohort Data (May 1, 2010 and July 1, 2010 Data Freezes)

Name	Description	Approixmate Size
16S high quality (V1–3, V3–5, V6–9)	Aggressively filtered and trimmed reads (low error rates, short reads) (454)	1.5-terabyte data
16S low quality (V1–3, V3–5, V6–9)	Less aggressively filtered and trimmed reads (higher error rates, longer reads) (454)	1.5-terabyte data
WGS mappings to reference genomes	Alignments between WGS data (both reads and ORFs) and reference genomes	2-terabyte data
MetaHIT mappings to reference genomes	Alignments between metaHIT WGS data and reference genomes	0.1-terabyte data
WGS pretty good assemblies (PGAs)	De novo assemblies from the WGS data	0.3-terabyte data
WGS hybrid assemblies	Mixed Illumina and 454 WGS metagenomic data, resulting in long, high-quality assemblies	0.1-terabyte data
WGS read annotations	Gene predictions and functional assignments from assemblies	3.6-terabyte data
Orthologous gene family abundances	Relative abundances of KOs from read-level blastx results	741 samples, 13,328 KO families
Functional/metabolic pathway coverage	Presence/absence of KEGG modules and pathways	741 samples, 246 small modules, 290 large pathways
Functional/metabolic pathway abundance	Relative abundances of KEGG modules and pathways	741 samples, 246 small modules, 290 large pathways

[a] These datasets are available on the HMP DACC website: www.hmpdacc.org.

These results also suggest that a careful examination of specialized metabolic functions, such as vitamin, toxin, or antimicrobial production or the production of signaling molecules or novel metabolites, will be key to deciphering the signature characteristics of each microbiome of the body.

Although major metabolic pathways appear to be common across all microbiomes, in fact we still know little about most of the predicted genes or proteins in the human microbiome. In analysis of the healthy cohort data, a large fraction (43%) of the metagenome sequence from the five major body sites could not be aligned to the reference genome sequences and the majority of the annotated genes (80–90% or over 4 million genes) and predicted proteins (75–85%) could not be assigned a function. Clearly, a next key step is to characterize the functional properties of the microbiome at both the strain and total community levels.

Further, most (although by no means all) communities are colonized predominantly by one specific group of bacteria. Most signature groups, in turn, consist of predominantly one specific microbial taxon, with subtypes present in lower abundance. This likely reflects niche specialization within these communities. Further, localized environmental factors such as vaginal pH were important in some communities. A very interesting future question will be what the "most important" factors are influencing lifelong microbiome composition, whether they are genetics, diet, birth environment, geography, or combinations of these factors.

1.3.4. Demonstration Projects of Microbiome–Disease Associations

The fourth resource of the HMP included a group of projects that were designed to determine whether correlations between microbiome community composition and specific diseases can be detected. It was recognized at the inception of the initiative that studies could not yet be conducted to determine whether there are causal relationships between specific diseases and changes in the microbiome. There was, however, sufficient evidence for a number of diseases that appeared to include a role for microbial communities in the disease processes. The "demonstration projects" program has this question as its goal in a number of different putative microbiome-associated diseases. The demonstration projects began with a 1-year pilot phase during which 15 projects recruited subjects and tested sampling protocols. Following an administrative review, 11 projects from the initial pool of 15 were funded to continue their work for 3 additional years.

Of these 11 studies, six projects study the microbiome associated with gut diseases, three study the microbiome and urogenital conditions or diseases, and two study the microbiome and skin diseases (Table 1.3). Depending on the study, the age groups recruited ranged from birth to over 50 years old, and the number of subjects recruited ranged from 19 to 489. Most are case–control studies. Almost all of the studies included targeted 16*S* rRNA gene sequencing, and some included WGS metagenome sequencing of the microbiomes inhabiting unaffected body sites and the diseased tissue of interest. Some of these studies also included the analysis of functional markers of the microbiome such as gene expression or gene products of the microbial communities or metabolomic studies of the microbiome.

These projects are a diverse set of carefully controlled case studies with large cohort sizes that support the correlation of microbiome changes with development of specific diseases. These studies will contribute valuable datasets for further study as they include detailed clinical metadata such as the disease phenotype along with phylogenetic and total community analysis of the microbiomes from controls and disease-associated tissues. Many of these studies also include microbial genome sequences from reference strains isolated from the diseased tissue of interest. The data are rapidly released into the public domain. Many of these studies also include characterization of the microbiomes prior to disease development, in response to the presence of disease or, in some cases, in response to standard-of-care interventions and so include additional dimensions of analysis to the study of the associations of microbial communities with specific diseases.

Early results from some of these demonstration project studies are beginning to suggest that a characteristic microbiome community appears to be associated with the specific disease under study. For example, neonatal enterocolitis, esophageal adenocarcinoma, ulcerative colitis, Crohn's disease, and eczema all appear to have a characteristic microbial community associated with the disease state, which is different from the microbial composition of control tissues. Further, the microbial signatures associated with the some of these disease states include both structural markers, such as the community composition, as well as functional markers, such as

TABLE 1.3. Summary of HMP Demonstration Projects[a]

Principal Investigator(s)	Short Title	Number of Subjects	Age and Other Criteria	Study Type	Sequence Data Type, Technology	WHO ICD General Disease	WHO ICD Specific Disease
Phillip Tarr, Washington Univ. School of Medicine, St. Louis, MO	The Neonatal Microbiome and NEC	489	<1500 g at birth, other criteria	Case–control	Metagenome, Roche 454 GS FLX titanium	Certain conditions originating in perinatal period	Necrotizing enterocolitis
Gary Wu, James Lewis, Frederic D. Bushman, Univ. Pennsylvania School of Medicine, Philadelphia	Diet, Genetic Factors, and the Gut Microbiome in Crohn's Disease	128	4 studies: FSM (>18 yo); CAFE(18–40 yo); COMBO (2–50 yo); PLEASE (<22 yo); other criteria	Cross-sectional, controlled trial, longitudinal cohort	16S rRNA and metagenome, Roche 454 GS FLX titanium	Diseases of digestive system	Crohn's disease
James Versalovic, Baylor College of Medicine, Texas Children's Hospital, Houston	The Human Gut Microbiome and Recurrent Abdominal Pain in Children	44	7–12 yo; other criteria	Case–control	16S rRNA, Roche 454 GS FLX titanium	Diseases of digestive system	Irritable bowel syndrome

Vincent Young, Univ. Michigan , Ann Arbor; Eugene Change, Univ. Chicago; Folker Meyer, Argonne National Lab, Argonne, IL; Tom Schmidt and James Tiedje, Michigan State Univ., East Lansing, MI; Mitchell Sogin, Marine Bio Lab, Woods Hole, MA	Ulcerative Colitis Human Microbiome Project	23	>18 yo, other criteria	Longitudinal	16S rRNA and metagenome, Roche 454 GS FLX titanium	Diseases of digestive system	Ulcerative colitis
Claire Fraser-Liggett, Univ. Maryland School of Medicine, Baltimore	Human Gut Microbiome in Crohn's Disease	19	Five twin pairs, other criteria	Twin	Metagenome, Roche 454 GS FLX titanium	Diseases of digestive system	Crohn's disease
Zhiheng Pei, New York Univ. Langone Medical Center, New York; Karen Nelson, J. Craig Venter Institute, Rockville, MD	Foregut Microbiome in Development of Esophageal Adenocarcinoma	42	>50 yo, other criteria	Case–control	Metagenome, Roche 454 GS FLX titanium	Neoplasms	Malignant neoplasm of esophagus
J. Dennis Fortenberry, Indiana Univ. School of Medicine, Indianapolis	Urethral Microbiome of Adolescent Males	55	14–17 yo, other criteria	Longitudinal, observational, cohort	Metagenome, Roche 454 GS FLX titanium	Certain infectious and parasitic diseases	Infections with a predominantly sexual mode of transmission

(*Continued*)

TABLE 1.3. (*Continued*)

Principal Investigator(s)	Short Title	Number of Subjects	Age and Other Criteria	Study Type	Sequence Data Type, Technology	WHO ICD General Disease	WHO ICD Specific Disease
Gregory Buck, Virginia Commonwealth Univ., Richmond	The Vaginal Microbiome: Disease, Genetics, and the Environment	460	>18 yo, other criteria	Twin, clinical cohort	16S rRNA, Roche 454 GS FLX titanium	Diseases of genitourinary system	Bacterial vaginosis
Jacques Ravel, Univ. Maryland School of Medicine, Baltimore; Larry Forney, Univ. Idaho	The Microbial Ecology of Bacterial Vaginosis	200	>18 yo, reproductive age, other criteria	Longitudinal, prospective	16S rRNA, Roche 454 FLX titanium	Diseases of genitourinary system	Bacterial vaginosis
Martin Blaser, New York Univ., Langone Medical Center, New York	Cutaneous Microbiome in Psoriasis	200	Not specified	Longitudinal, case–control	16S rRNA and metagenome, Roche 454 GS titanium, RNAseq metatranscriptomics	Diseases of skin and subcutaneous tissue	Psoriasis vulgaris
Julie Segre, National Human Genome Research Institute, NIH, Bethesda, MD; Heidi Kong, National Cancer Institute, NIH, Bethesda, MD	Skin Microbiome, Atopic Dermatitis, and Immunodeficiency	33	3–40 yo, other criteria	Longitudinal, case–control	Whole-genome genotyping and ABI de novo sequencing	Diseases of skin and subcutaneous tissue	Atopic dermatitis

[a] This tabulation includes project investigator names and affiliations (PIs), short titles, number of subjects, age range (yo = years old) and main inclusion criteria, study type, sequence data type, and sequencing technology used for the project. Please refer to the Nature Preceding marker papers for more detail on each study. Categorization of the general and specific disease(s) under study are indicated according to the WHO International Code of Diseases 2010 (ICD10) classification system. The WHO ICD10 general and specific disease categories are linked to the ICD website.

the metaproteome of the disease-associated microbial community, providing a suite of markers for possible diagnostic or prognostic applications.

In addition, it appears in some cases that microbial markers may be emerging that appear to precede the disease state. For example, gastric esophageal reflux disease (GERD) is characterized by a series of diseases, starting with reflux esophagitis, continuing to Barrett's esophagus in about 20% of cases, and in rare cases of Barrett's esophagus, proceeding to the development of esophageal adenocarcinoma. In the Pei/Nelson esophageal adenocarcinoma demonstration project study, it was found that the microbiome of the intermediate stage of the disease (Barrett's esophagus) appears to be very similar to the microbiome of those patients who go on to develop adenocarcinoma, suggesting that the microbiome in Barrett's esophagus is a potential precursor state to the cancer. In this case, it may be possible to develop diagnostic biomarkers for adenocarcinoma far before the cancer develops. A summary (Table 1.3) includes links to the Bioprojects pages at NCBI that describes the projects and leads the user to the data. This table also lists the general and specific disease(s) under study in each project as they are categorized according to the WHO International Disease Classification 2010 (ICD10) system (`http://www.who.int/classifications/icd/en/`). Summaries of the hypotheses, aims, and data types to be produced are documented as marker papers in Nature Proceedings (`http://precedings.nature.com/search?query=human+microbiome+project`). Highlights of the early results from these projects are provided below and are categorized by body site.

Gut Diseases and the Microbiome

Three of the gut disease projects include the microbiomes of younger populations, such as neonates [Tarr, necrotizing enterocolitis (NEC)] or children [Versalovic, irritable bowel syndrome (IBS)]. Wu, Lewis and Bushman, in their multifaceted project on Crohn's disease (CD), also included a study of the effects of an elemental diet on pediatric patients with inflammatory bowel disease (IBD). Four of the gut projects focus on gut diseases in adults [Fraser–Liggett, CD; three of the four studies by Wu et al. involve CD in adults; Young, ulcerative colitis (UC)]. The fourth digestive tract project was described earlier and is somewhat unusual in that it is the one study of the association of the microbiome with a cancer (Pei/Nelson, esophageal adenocarcinoma).

The gut disease studies with young patients are showing some promising early results. The Tarr study on NEC found that antibiotic treatment, the standard of care for premature infants, decreased gut microbial diversity and that this was associated with the development of NEC. Further, they found that key host immune system markers increased before the appearance of NEC, although the specificity and timing of these host signals need further study. In the Versalovic study, there appear to be IBS-specific microbial signatures in those children with IBS and further, assemblages of specific gut microbial taxa that may distinguish between the occurrence of subtypes of pediatric IBS. In the pediatric study of the Wu et al. project, elemental dietary interventions appeared to change gut microbiome composition within 24 h of intervention, suggesting that elemental diet therapy can have a major impact on the composition of the gut microbiome in pediatric patients with CD and possibly the disease itself.

Early results from the studies of adult microbiomes and gut diseases are promising as well. The Fraser–Liggett CD study compared twins with either ileal CD (iCD) or colonic CD (cCD). This was a multifaceted study, and many aspects of the microbiome (microbial composition, gene content, and gene products) were correlated with patient clinical metadata. Early results suggest that, although the picture is not yet clear for cCD, a combination of specific microbial assemblages and their genes and products appear to correlate with iCD and are consistent with the increased inflammation seen in the iCD gut. These markers may lead to diagnostic tools for assessing the development of iCD.

The Young et al. study is also a multifaceted study of the association of a gut disease, UC, with the microbiome and includes an interesting experimental model component. IBD consists of iCD, cCD, and UC, a disease of the colon. For some UC patients, the colon must be removed (colectomy) and a pseudorectum ("pouch") is formed from a segment of their small intestines. In over 50% of these patients, the pseudorectum may itself become inflamed, a condition known as "pouchitis." The Young et al. study follows patients who have undergone the pseudorectum surgery and has found that the microbiome composition of patients with pouchitis shifts to a microbial community more similar to the colon microbiome composition of UC patients, even though the pouchitis occurs in a structure formed from the small intestine, not the colon. This pouchitis study appears to be a good experimental model for UC and provides insights for isolating the role of the gut microbiome in the cause or contribution to the development of UC.

Urogenital Diseases and the Microbiome

The microbiome–urogenital disease association studies include bacterial vaginosis and sexually transmitted infections (STIs) and the vaginal microbiome (Buck) and included both longitudinal studies and twin studies. The Fortenberry study included the relationship of circumcision, sexual history, and STIs with the penile microbiome of adolescent males. Early results from the Buck study suggest that there may be a genetic component to the composition of the vaginal microbiome. Further, the vaginal microbiome composition appears to respond to the hormonal cycle as microbial diversity appears to be lowest at midmenstrual cycle or, in other words, during ovulation.

The Fortenberry study included monthly sampling of both the urethra and the coronal sulcus of the penis of adolescent males to characterize changes in the microbiome over time, in response to sexual activity and between circumcised and uncircumcised males. Early results show that, although there are differences in the penile microbiome between, for example, circumcised and uncircumcised males, the microbial composition appeared to be fairly stable over time. Further, the urethral microbiome composition differed between those males with and without STIs. These results may be applicable to the treatment of sexually transmitted diseases.

Skin Diseases and the Microbiome

One microbiome-associated skin disease project is a study of atopic dermatitis in children (eczema, Segre). Eczema is characterized by periodic exacerbations (known as *flares*) that result in highly inflamed skin. Until recently, eczema has been studied as a single pathogen disease. Segre's research team has examined the role of the skin microbiome in modulating the extent and duration of the disease and includes both pediatric and adult patients. In a longitudinal study of pediatric patients with eczema, the Segre lab found that total skin microbial diversity was reduced during flares with a concomitant increase in *Staphylococcus aureus*. Whether this dominant organism is a consequence or cause of the shifts in microbial diversity and of the disease is currently under study.

Because of the nature of the work, the demonstration projects are more akin to individual investigator projects, so no formal HMP analysis workgroup was formed. However, an informal workgroup has come together to discuss strategies for the analysis of each project's sequence data, to participate in tutorials created for the projects on the various computational tools being developed for human microbiome studies and to discuss common data standards in order to make the results from each study comparable. Publications from the demonstration projects are described in Section 1.4.

1.3.7. Technology Development

Two additional HMP initiatives were designed to provide resources for the HMP effort and for the field in general, technology development to isolate novel microorganisms and computational tools development. Because early estimates indicated that a large fraction (~40%) of the microorganisms associated with the human microbiome were not yet in culture, it was recognized that there was a critical need for new approaches that could isolate or enrich for new and novel microorganisms from the microbiome. In order to address this need, the technology development program supported 10 projects that are working on a wide variety of methodologies to enrich for and isolate specific populations of cells for downstream applications (Table 1.4). In many cases, these projects were intended as an investment in the long-term development of new technologies that could be applied in a 5–10-year time horizon. A few details on each project are provided below.

Five of the projects focus on enrichment and isolation of specific populations of cells by a variety of flow cytometric or microfluidic approaches (Han/Bradbury, Podar, Singh, Worthen, Relman, Chang). The Relman project includes the use of optical tweezers for isolating specific cells and *in situ* gene expression measurements of individual cells. The Chang lab is developing sorting and enriching techniques for specific populations from the colonic mucosa-associated microbial communities. Two projects are applying novel cultivation methodologies, one to isolate microaerophilic bacteria, those cells that grow best under low oxygen tension (Young/Schmidt); and a second to sort and encapsulate single cells into a gel matrix for microcolony cultivation (Doktycz). One project is developing a pipeline to process cells from flow sorting to single-cell genome sequence analysis (Zhang/Lo). The Marzaili project is focusing on a DNA purification methodology to enrich for low-abundance microbial sequences from a mixed assemblage.

The intended products from these activities include pure cultures, DNA isolated from a culture, whole-genome amplification products from single cells, or enrichments of specific strains within a mixture of cells. Some of these investigators whose methodologies had sufficiently matured will also collaborate with some of the

TABLE 1.4. HMP Technology Development Projects Listed by Project Title and Investigator

Project Titles	Investigator
FACS-MABE: a method for sorting and enriching the as-yet uncultured bacterial species from the human distal gut	Emma Allen-Vercoe (Univ. Guelph)
Species-by-species dissection of microbiomes using phage display and flow sorting	Andrew Bradbury (LANL)
Isolation, selection, and polony amplification of single cells in a gel matrix	Ronald Davis (Stanford)
Functional sorting of microbial cells from complex microbiota	Mitchel Doktycz (UT-Battelle, ORNL)
Novel cultivation methods for the domestication of vaginal bacteria	David Fredricks (Fred Hutchinson Cancer Research Center)
Confining single cells to enhance and target cultivation of human microbiome	Rustem Ismagilov (Univ. Chicago)
Culturing uncultivatable gut microorganisms	Kim Lewis (Northeastern Univ.)
Metagenomic dissestion of the gut microbiota	Xiaoxia Lin (Univ. Mich)
Tools for human microbiome studies	John Nelson (GE Global Research)
Targeted genomic characterization of uncultured bacteria from human microbiota	Mircea Podar (UT-Battelle, ORNL)
Optimization of a microfluidic device for single bacterial cell genomics	David Relman (Stanford)
Cultivation and characterization of microaerobes from the human microbiome	Thomas Schmidt (MSU)
FISH “N” Chips: A microfluidic processor for isolating and analyzing microbes	Anup Singh (Sandia National Laboratories)
Multidimensional separation of bacteria	Scott Worthen (CHOP)
An integrated lab-on-chip system for genome sequencing of single microbial cells	Kun Zhang (UCSD)

demonstration project research teams to apply their methodologies to enrich for and isolate specific cells from control and diseased tissue microbiome samples.

1.3.8. Computational Tools

The sixth Initiative of the HMP includes the development of computational and bioinformatic tools to support and advance human microbiome sequence data analysis, particularly of metagenomic sequence data. The computational tools program supports 10 projects for a wide variety of tools for this purpose (Table 1.5). The common theme across all of these projects was the recognition that next-generation sequencing technologies are able to produce orders of magnitude more sequence than the traditional Sanger chemistry sequence technologies. In fact, terabase-range datasets of metagenomes from complex microbial communities are increasing in frequency, yet computational tools are not routinely available that can accommodate these massive datasets.

It is perhaps valuable to take a moment to note the fundamental differences between sequence analysis of a metagenome and sequence analysis of the genome

TABLE 1.5. HMP Computational Tools Projects Listed by Project Title and Investigator

Project Title or Description	Investigator
Algorithmically Tuned Protein Families, Rule-Base and Characterized Proteins	Daniel Haft (J. Craig Venter Institute)
Novel Computational Tools for Studying the Human Microbiome	David Fredricks (Fred Hutchinson Cancer Research Center)
Functional activity and interorganismal interactions in the human microbiome	Curtis Huttenhower (Harvard Univ.)
New Tools for Understanding the Composition and Dynamics of Microbial Communities	Robin Knight (Univ. Colorado-Boulder)
Novel Methods for Effective Analysis Assembly and Comparison of HMP Sequences	Weizhong Li (UC San Diego)
High Performance Validation and Classification of Metagenomic Ribosomal-RNA Sequences	Dan Franks (Univ. Denver)
Assembly and analysis software for exploring the human microbiome	Mihai Pop (Univ. Maryland)
Identifying population-level variation in cross-sectional and longitudinal HMP studies	Patrick Schloss (Univ. Michigan)
Exploiting Microbiome Sequences for Improved Models of Protein-DNA Interactions	Gary Stormo (Washington Univ. – St. Louis)
Fragment assembly and metabolic/species diversity analysis for HMP data	Yuzhen Ye (Indiana Univ.)

of a single microorganism. The sequence reads derived from the DNA of a single microbial species can be assembled, aligned, and annotated because (1) one is trying to reconstruct the genome of a single microorganism and all of the sequence come from a single organism, (2) there are databases with the genome sequences of similar or related microorganisms that can be used for comparison and as a guide to reconstructing the new genome, and (3) most microbial genomes are closed, circular structures, which increases the probability that a full genome sequence can be completed.

A robust microbial genome sequence database is needed in order to identify organisms and their close relatives in WGS metagenome datasets. To address this need, the HMP set a goal to add the sequence of at least 3000 new bacterial genomes to the current database. As of this writing, 800 microbial genome sequences had been completed and deposited in NCBI while another 500 microbial genome sequences are in progress (`http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi, accessed 02/13/12`).

The four traits of massive size, fragmentary nature of the data, data complexity, and rapidly changing sequence technologies for metagenomic data have demanded the development of a wide range of new and novel analytical tools that can operate efficiently, dependably handle large datasets, and operate at sufficient speeds for routine data analysis of metagenomic data. The computational tools program projects can generally be grouped into two categories: (1) those that focus on the diversity of the microbial community and (2) those that focus on the metabolic potential of the microbial community. Some projects are contributing to the development of tools for the upstream aspects of metagenomic sequence analysis. Franks and Eddy are developing tools to improve the quality of the primary sequence data and to guide 16S rRNA gene sequence alignments using this gene's primary and secondary structure characteristics. The individual sequence reads need to be systematically pieced together, but assembling metagenomic data is very different from assembling a genome sequence as all of the sequences in the community have not been sampled or species information about the community is incomplete so that few reference genomes are available for piecing together a metagenome sequence. Pop is developing metagenomic assembly validation/quality control tools and alignment tools to address these issues. Li is creating a metagenomic meta-assembler from existing assembly tools in order to select the best features of each tool that will increase speed, performance, and capability of an assembly tool. See Tables 1.5 and 1.7.

Several investigators are developing a suite or pipeline of tools that can process sequence data and produce analytical results at different endpoints of completion. Some pipelines address the annotation component of metagenomic analysis; for example, Ye has developed a suite of tools that can take the metagenomic sequence data from the assembled state through to annotation and metabolic pathway reconstruction. In order to accommodate the current inability to name all of the species in a microbial community, Schloss has developed a pipeline of tools that employs both a taxon-free approach (i.e., does not require the identification of the microorganism) and a phylogenetic approach (i.e., evolutionary relatedness of microorganisms) to characterize microbial community diversity and visualization tools to describe this diversity. Fredricks is developing novel methods for refining placement of microbial sequences in phylogenetic trees. Knight has produced a suite of tools that include particularly strong visualization tools for conducting time series analyses of microbial composition. See Tables 1.5 and 1.7.

Three projects have focused on the functional properties of the microbiome. Huttenhower is developing tools for analysis of the metabolic potential in microbiomes from metagenomic data and to identify the interspecies regulatory networks in the microbiome. Stormo is developing tools to analyse the regulatory properties of the microbiome, particularly the DNA-protein interactions that play a role in regulating the transcription of genes across the microbiome. Haft is developing tools for the analysis of protein families predicted from the metagenome data. Publications and links to tools that are available on the grantee's website, the DACC or on `www.sourceforge.net` are listed in Section 1.4. See also Tables 1.5 and 1.7.

1.3.9. Ethical, Legal, and Societal Implications of Microbiome Research

A unique feature of the HMP among the many human microbiome programs around the world is the inclusion of a program in the ethical, legal, and social implications (ELSI) of human microbiome research. ELSI studies have become a legacy of the Human Genome Project and of the extramural research program at the National Human Genome Research Institute (NHGRI), which is congressionally mandated to allocate 5% of its research budget to the support of studies in this area. A number of interesting ethical issues have arisen in the course of HMP. Other issues could potentially arise and are worthy of study. Studies funded under HMP include issues related to the equitable selection of research participants, identifiability of individuals through microbiome profiles, informed consent to participate

TABLE 1.6. The HMP Ethical, Legal, and Societal Implications (ELSI) Projects Listed by Project Investigator and Title or Description

Investigator	Institution	Title or Description
Mildred Cho and Pamela Sankar	Stanford Univ.	Toward a Framework for Policy Analysis of Microbiome Research
Paul Spicer	Univ. Oklahoma, Norman	Indigenous Communities and Human Microbiome Research
Diane Hoffmann	Univ. Maryland, Baltimore	Federal Regulation of Probiotics: An Analysis of Existing Regulatory Framework
Amy Lynn Mcguire	Baylor College of Medicine	Ethical, Legal, and Social Dimensions of Human Microbiome Research
Rosamond Rhodes	Mount Sinai School of Medicine	Human Microbiome Research and the Social Fabric
Ruth Farrell and Richard Sharp	Cleveland Clinic	Patient perceptions of bioengineered probiotics and clinical metagenomics

in human microbiome research, data sharing practices and protection of privacy, invasiveness of sampling protocols, and the return of research results and incidental findings to the research participants. Further, the research results may have a significant impact on the nature and direction of clinical medicine and on potential products, such as probiotics, that could be developed on the basis of the research findings, and all of these raise broader societal implications.

The HMP supports six investigative teams to conduct research on a broad range of ELSI topics (Table 1.6). The Cho–Sankar team is conducting an analysis of risk/benefit concepts in microbiome research and of how experimental design may reflect perceptions of ethical issues in human microbiome research. Spicer is conducting an analysis of the potential impact of human microbiome research results on the concepts of social and ancestral identity in indigenous peoples. Hoffmann is conducting an analysis of federal regulation of probiotics, in order to assess whether the current regulatory framework ensures probiotic safety and the accuracy of health-related claims. McGuire is studying the perceptions and attitudes of current research participants in the "healthy cohort" study of the HMP, as well as of HMP researchers, regarding a wide range of ethical issues in human microbiome research, with the goal of making recommendations for guidelines for the management of ethical issues in future microbiome research. Rhodes is developing guidelines for educating the lay and scientific public about ethical issues with respect to human subject research, biobanking, public health, and commercialization of products for treating the human microbiome. The Sharp–Farell team is conducting an analysis of patient perceptions of probiotics for the treatment of medical conditions.

1.4. PRODUCTS FROM THE HUMAN MICROBIOME PROJECT

There are a number of products from the research conducted within the HMP. Several have already been outlined in previous sections such as the reference strains sequencing activities and cultures at BEI. Research efforts are now focusing on expanding the collection to include eukaryotic microbes, eukaryotic viruses, and bacteriophages, and outreach activities are underway to communicate with those groups that specialize in these microbial taxa to contribute strains to the HMP sequencing efforts.

1.4.1. Derivative Datasets from the Healthy Adult Cohort Study

Sequence data prepared by the participating sequencing centers were deposited in the sequence read archive (SRA) as the data were being produced. As noted earlier, the DACC, DAWG, and their workgroups processed the data according to agreed-on parameters and also developed a number of derivative datasets of the 16*S* and metagenomic WGS sequence data. The group carried out these data processing steps in order to create master, common sets of data for downstream analyses. In 2011, the DAWG also decided that two specific derivative datasets from the suite of datasets should also be released to provide a community resource for microbiome researchers. This set included deconvoluted, trimmed 16*S* data that included 74 million 16*S* reads from over 6000 of the healthy adult cohort study samples; the 16*S* rRNA variable region V3–V5 was sequenced for all of these, while the V1–V3 and V6–V9 regions were also sequenced for a subset of these 6000 samples. De novo metagenomic assemblies from an initial group of 690 of these samples were also included in this release. In addition, the DACC is preparing a gene index of all proteins predicted from these assemblies, which will also be released as a community resource. In keeping with rapid release of data to the public, these data derivative releases were posted on the DACC website (`http://www.hmpdacc.org/doc/PGA_16SData_post.pdf`), distributed to Newswire (`http://www.newswire.com`) and posted on ASM's MicrobeWorld (`http://www.microbeworld.org/index.php?option=com_jlibrary&view=article&id=6682`).

1.4.2. Computational Tools for Human Microbiome Research

Many computational and bioinformatic tools have been developed by the HMP Computational Tools grantees and were either refined under HMP support or were already developed and adapted to HMP data types. The list of tools is provided (Table 1.7), along with links to their lab websites, the DACC or `www.sourceforge.net`. Other tools were developed by the sequencing centers and other members of the HMP Research Network Consortium, or existing tools were adapted for use in human microbiome analysis. These are included on the DACC website filed under "Get Tools."

1.4.3. Publications from the HMP

As of this writing 194 publications cite the NIH Human Microbiome Project for support. Online supplemental material for this chapter provided by the publisher includes the HMP PubMed publications list (see Table 1.8).

TABLE 1.7. Computational Tools Developed or Modified for HMP

Yuzhen Ye, Indiana Univ.: fragment assembly and metabolic/species diversity analysis for HMP (Ye lab website: http://omics.informatics.indiana.edu/hmp/index.php)

FragGeneScan: a tool for fragmental gene prediction in short reads (http://omics.informatics.indiana.edu/FragGeneScan/)

AbundanceBin: abundance-based tool for binning metagenomic sequences (http://omics.informatics.indiana.edu/AbundanceBin/)

MinPath: a parsimony approach for biological pathway reconstructions using protein family predictions (http://omics.informatics.indiana.edu/MinPath/)

AbundantOTU (and AbundantOTU+): a tool for fast and accurate identification and quantification of abundant species from pyrosequences of 16S rRNA by using consensus alignment (http://omics.informatics.indiana.edu/AbundantOTU/)

PHYLOSHOP: a tool for extracting ribosomal RNA fragments from WGS and simple analysis of ribosomal RNAs (http://omics.informatics.indiana.edu/mg/phyloshop/)

RAPSearch: a fast tool for protein similarity search (http://omics.informatics.indiana.edu/mg/RAPSearch/)

RAPSearch2: an even faster RAPSearch that supports multithreading (http://omics.informatics.indiana.edu/mg/RAPSearch2/)

SWIFT: a fast protein similarity search tool that utilizes a reduced amino acid alphabet and suffix arrays to detect seeds of flexible length; in development

PathRecruit: an online resource for computing and visualizing both the functional diversity and species diversity for metagenomic samples, in development

Mihai Pop, Univ. Maryland, College Park: assembly and analysis software for exploring the human microbiome (Pop lab software website: http://www.cbcb.umd.edu/~mpop/Software.shtml)

Minimus-SR: short-read version of our assembler Minimus (available open-source at amos.sourceforge.net in short-read assembly package http://sourceforge.net/apps/mediawiki/amos/index.php?title=Minimus)

Bambus 2: extensions to AMOS package for analysis of assembly graphs (http://sourceforge.net/apps/mediawiki/amos/index.php?title=Bambus2)

Crossbow: cloud-computing-enabled sequence aligner and variant caller (http://bowtie-bio.sourceforge.net/crossbow/index.shtml)

Metastats: statistical package for comparing metagenomic samples (http://metastats.cbcb.umd.edu/)

Metapath: statistical package for comparing metagenomic samples at the pathway level (http://cbcb.umd.edu/~boliu/metapath/)

Phymm: statistical binning for metagenomic data (http://cbcb.umd.edu/software/phymm/)

Contrail: cloud-enabled assembler (http://sourceforge.net/apps/mediawiki/contrail-bio/index.php?title=Contrail)

DNAclust: fast and accurate clustering of DNA sequences (http://dnaclust.sourceforge.net/)

Rob Knight, University of Colorado, Boulder: New tools for understanding the composition and dynamics of microbial communities

QIIME (Quantitative Insights Into Microbial Ecology): software package for comparison and analysis of microbial communities, primarily based on high-throughput amplicon sequencing data generated on a variety of platforms, but also supporting analysis of other types of data (such as shotgun metagenomic data) (http://qiime.sourceforge.net/)

SitePainter: allows users to visualize the different HMP body sites based on gradients of colors to represent available datasets (http://www.hmpdacc.org/sp/)

Daniel Haft, J. Craig Venter Institute: algorithmically-tuned protein families, also rule_base and characterized proteins

TIGRFAMs—resource consisting of curated multiple sequence alignments, hidden Markov models (HMMs) for protein sequence classification, and associated information designed to support automated annotation of (mostly prokaryotic) proteins (http://www.jcvi.org/cgi-bin/tigrfams/index.cgi)

(*Continued*)

TABLE 1.7. (*Continued*)

Partial phylogenetic profiling (PPP) software and comparative genomics database: (ftp://ftp.icvi.org/pub/data/ppp)

CHAR (Database of Experimentally Characterized Proteins): new annotation rules created and included in distributions of the JCVI-produced tool AutoAnnotate (http://www.jcvi.org/cms/research/projects/annotation-service/)

TIGRFAMs: added collections of HMM-based protein family definitions for automated annotation pipelines (ftp://ftp.icvi.Qrg/pub/data/TIGRFAMs/)

CRISPR—Cas system classification (http://www.nature.com/nrmicro/journal/v9/n6/full/nrmicro2577.html)

Gary Stormo, Washington University at St Louis: exploiting microbiome sequences for improved models of protein-DNA interactions (Stormo lab page: http://ural.wustl.edu/resources.html#Software)

Daniel Frank, University of Colorado, Denver: high performance validation and classification of metagenomic ribosomal-RNA sequences (Software that is at least in the beta testing stage is provided with no restrictions at: http://www.phyloware.com/Phyloware/Home.html; XplorSeq—Mac OSX software for sequence analysis: http://www.phyloware.com/Phyloware/XplorSeq.html)

Patrick Schloss, University of Michigan: identifying population-level variation in cross-sectional and longitudinal HMP studies. Mothur: single resource that incorporates functionality of several tools to analyze microbial ecology data (http://www.mothur.org/)

Weizhong Li, University of California, San Diego: novel methods for effective analysis assembly and comparison of HMP sequences

Meta-assembler for 454 reads (http://camera.calit2.net/)

FR-HIT: A new fragment recruitment method called FR-HIT has been implemented. FR-HIT has similar sensitivity as BLASTN but is about 2 orders of magnitude faster in recruiting raw reads. FR-HIT is slower than some mapping programs, but it can recruit several times more reads (http://weizhong-lab.ucsd.edu/frhit/). Source code (http://code.google.com/p/frhit/)

Meta-RNA(H3): The rRNA prediction method Meta-RNA was improved using Hmmer3 (http://weizhong-lab.ucsd.edu/meta_rna/)

WebMGA: A collection of web servers, WebMGA, has been created. In addition to the new Meta-rRNA and cd-hit-454, WebMGA includes ~20 other commonly used tools such as ORF calling, sequence clustering, quality control of raw reads, removal of contaminations and functional annotation (http://weizhong-lab.ucsd.edu/metagenomic-analysis/)

Curtis Huttenhower, Harvard University School of Public Health: functional activity and inter-organismal interactions in the human microbiome (http://huttenhower.org/galaxy/)

The LEfSe algorithm has been developed to discover and explain microbial and functional biomarkers in the human microbiota and other microbiomes. Its accuracy in comparison to existing methods using both synthetic data and published metagenomic functional gene family catalogs has been validated. The method is freely available online and has already, before publication, received nearly 200 unique nonrobot visitors.

HUMAnN, an end-to end system for reconstructing gene families and functional and metabolic pathways from metagenomic (or metratranscriptomic) data, has been developed. HUMAnN has been validated as a significant improvement over state-of-the-art using a synthetic metagenomes and was used for metabolic reconstruction of 649 samples (>2.5Tbp sequence) from 7 body sites on 102 individuals as part of the HMP (http://huttenhower.sph.harvard.edu/humann)

David Fredricks, Fred Hutchinson Cancer Research Center: novel computational tools for studying the human microbiome

Bioconductor software package for processing of high throughput sequence reads (http://bioconductor.org/packages/devel/html/microbiome454.html)

Reference package standard (http://github.com/fhcrc/taxtastic/wiki/refpkg)

TABLE 1.8. HMP Publications in PubMed (Updated 02/15/12)

1. Gonzalez A, Stombaugh J, Lauber CL, Fierer N, Knight R. SitePainter: A tool for exploring biogeographical patterns. *Bioinformatics* **28**(3):436–468 (2012).
2. Gonzalez A, Knight R. Advancing analytical algorithms and pipelines for billions of microbial sequences. *Curr Opin Biotechnol* **23**(1):64–71 (2012).
3. Mercer M, Brinich MA, Geller G, Harrison K, Highland J, James K, Marshall P, McCormick JB, Tilburt J, Achkar JP, et al. How patients view probiotics: Findings from a multicenter study of patients with inflammatory bowel disease and irritable bowel syndrome. *J Clin Gastroenterol* **46**(2):138–144 (2012).
4. Bik HM, Porazinska DL, Creer S, Caporaso JG, Knight R, Thomas WK. Sequencing our way towards understanding global eukaryotic biodiversity. *Trends Ecol Evol* **27**(4):233–243 (2012).
5. Cuellar-Partida G, Buske FA, McLeay RC, Whitington T, Noble WS, Bailey TL. Epigenetic priors for identifying active transcription factor binding sites. *Bioinformatics* **28**(1):56–62 (2012).
6. Zhao Y, Tang H, Ye Y. RAPSearch2: A fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics* **28**(1):125–126 (2012).
7. Yao G, Ye L, Gao H, Minx P, Warren WC, Weinstock GM. Graph accordance of next-generation sequence assemblies. *Bioinformatics* **28**(1):13–16 (2012).
8. Sun Y, Cai Y, Huse SM, Knight R, Farmerie WG, Wang X, Mai V. A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. *Brief Bioinform* **13**(1):107–121 (2012).
9. Werner JJ, Koren O, Hugenholtz P, DeSantis TZ, Walters WA, Caporaso JG, Angenent LT, Knight R, Ley RE. Impact of training sets on classification of high-throughput bacterial 16s rRNA gene surveys. *ISME (International Society for Microbial Ecology) J* **6**(1):94–103 (2012).
10. Markowitz VM, Chen IM, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P, et al. IMG: The Integrated Microbial Genomes database and comparative analysis system. *Nucleic Acids Res* **40**(database issue):D115–D122 (2012).
11. Madupu R, Richter A, Dodson RJ, Brinkac L, Harkins D, Durkin S, Shrivastava S, Sutton G, Haft D. CharProtDB: A database of experimentally characterized protein annotations. *Nucleic Acids Res* **40**(database issue):D237–D241 (2012).
12. Pagani I, Liolios K, Jansson J, Chen IM, Smirnova T, Nosrat B, Markowitz VM, Kyrpides NC. The Genomes OnLine Database (GOLD) v.4: Status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* **40**(database issue):D571–D579 (2012).
13. Markowitz VM, Chen IM, Chu K, Szeto E, Palaniappan K, Grechkin Y, Ratner A, Jacob B, Pati A, Huntemann M, et al. IMG/M: The integrated metagenome data management and comparative analysis system. *Nucleic Acids Res* **40**(database issue):D123–D129 (2012).
14. Ji X, Pushalkar S, Li Y, Glickman R, Fleisher K, Saxena D. Antibiotic effects on bacterial profile in osteonecrosis of the jaw. *Oral Dis* **18**(1):85–95 (2012).
15. Lewis CM Jr, Obregón-Tito A, Tito RY, Foster MW, Spicer PG. The Human Microbiome Project: Lessons from human genomics. *Trends Microbiol* **20**(1):1–4 (2012).
16. Kuczynski J, Lauber CL, Walters WA, Parfrey LW, Clemente JC, Gevers D, Knight R. Experimental and analytical tools for studying the human microbiome. *Nat Rev Genet* **13**(1):47–58 (2011).
17. Gonzalez A, King A, Robeson Ii MS, Song S, Shade A, Metcalf JL, Knight R. Characterizing microbial communities through space and time. *Curr Opin Biotechnol* **23**(3):431–436 (2011).

(*Continued*)

TABLE 1.8. (*Continued*)

18. Liu Z, Hsiao W, Cantarel BL, Drábek EF, Fraser-Liggett C. Sparse distance-based learning for simultaneous multiclass classification and feature selection of metagenomic data. *Bioinformatics* **27**(23):3242–3249 (2011).
19. Kuczynski J, Stombaugh J, Walters WA, González A, Caporaso JG, Knight R. Using QIIME to analyze 16S rRNA gene sequences from microbial communities. In *Current Protocols in Bioinformatics*, AD Baxevanis et al., eds., Wiley, 2011, Chap. 10, Unit 10.7.
20. McDonald D, Price MN, Goodrich J, Nawrocki EP, Desantis TZ, Probst A, Andersen GL, Knight R, Hugenholtz P. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J* **6**(3):610–628 (2011).
21. Fierer N, Lauber CL, Ramirez KS, Zaneveld J, Bradford MA, Knight R. Comparative metagenomic, phylogenetic and physiological analyses of soil microbial communities across nitrogen gradients. *ISME J* **6**(5):1007–1017 (2011).
22. Basu MK, Selengut JD, Haft DH. ProPhylo: Partial phylogenetic profiling to guide protein family construction and assignment of biological process. *BMC Bioinformatics* **12**:434 (2011).
23. Schwab AP, Frank L, Gligorov N. Saying privacy, meaning confidentiality. *Am J Bioeth* **11**(11):44–45 (2011).
24. Rhodes R, Azzouni J, Baumrin SB, Benkov K, Blaser MJ, Brenner B, Dauben JW, Earle WJ, Frank L, Gligorov N, et al. De minimis risk: A proposal for a new category of research risk. *Am J Bioeth* **11**(11):1–7 (2011).
25. Youssef NH, Blainey PC, Quake SR, Elshahed MS. Partial genome assembly for a candidate division OP11 single cell from an anoxic spring (Zodletone Spring, Oklahoma). *Appl Environ Microbiol* **77**(21):7804–7814 (2011).
26. Koren S, Treangen TJ, Pop M. Bambus 2: Scaffolding metagenomes. *Bioinformatics* **27**(21):2964–2971 (2011).
27. Pirrung M, Kennedy R, Caporaso JG, Stombaugh J, Wendel D, Knight R. TopiaryExplorer: Visualizing large phylogenetic trees with environmental metadata. *Bioinformatics* **27**(21):3067–3069 (2011).
28. Holtz LR, Wylie KM, Sodergren E, Jiang Y, Franz CJ, Weinstock GM, Storch GA, Wang D. Astrovirus MLB2 viremia in febrile child. *Emerg Infect Dis* **17**(11):2050–2052 (2011).
29. Saulnier DM, Riehle K, Mistretta TA, Diaz MA, Mandal D, Raza S, Weidler EM, Qin X, Coarfa C, Milosavljevic A, et al. Gastrointestinal microbiome signatures of pediatric patients with irritable bowel syndrome. *Gastroenterology* **141**(5):1782–1791 (2011).
30. Plottel CS, Blaser MJ. Microbiome and malignancy. *Cell Host Microbe* **10**(4):324–335 (2011) (review).
31. Knights D, Parfrey LW, Zaneveld J, Lozupone C, Knight R. Human-associated microbial signatures: Examining their predictive value. *Cell Host Microbe* **10**(4):292–296 (2011).
32. Proctor LM. The Human Microbiome Project in 2011 and beyond. *Cell Host Microbe* **10**(4):287–391 (2011).
33. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen YY, Keilbaugh SA, Bewtra M, Knights D, Walters WA, Knight R, et al. Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334**(6052):105–108 (2011).
34. Chen CH, Cho SH, Chiang HI, Tsai F, Zhang K, Lo YH. Specific sorting of single bacterial cells with microfabricated fluorescence-activated cell sorting and tyramide signal amplification fluorescence in situ hybridization. *Anal Chem* **83**(19):7269–7275 (2011).

TABLE 1.8. (*Continued*)

35. Zaneveld JR, Parfrey LW, Van Treuren W, Lozupone C, Clemente JC, Knights D, Stombaugh J, Kuczynski J, Knight R. Combined phylogenetic and genomic approaches for the high-throughput study of microbial habitat adaptation. *Trends Microbiol* **19**(10):472–482 (2011) (review).
36. McKenzie VJ, Bowers RM, Fierer N, Knight R, Lauber CL. Co-habiting amphibian species harbor unique skin bacterial communities in wild populations. *ISME J* **6**(3):588–596 (2011).
37. Sczesnak A, Segata N, Qin X, Gevers D, Petrosino JF, Huttenhower C, Littman DR, Ivanov II. The genome of th17 cell-inducing segmented filamentous bacteria reveals extensive auxotrophy and adaptations to the intestinal environment. *Cell Host Microbe* **10**(3):260–272 (2011).
38. Arias CA, Panesso D, McGrath DM, Qin X, Mojica MF, Miller C, Diaz L, Tran TT, Rincon S, Barbu EM, et al. Genetic basis for in vivo daptomycin resistance in enterococci. *N Engl J Med* **365**(10):892–900 (2011).
39. Wu S, Zhu Z, Fu L, Niu B, Li W. WebMGA: A customizable web server for fast metagenomic sequence analysis. *BMC Genomics* **12**:444 (2011).
40. Bowers RM, Sullivan AP, Costello EK, Collett JL Jr, Knight R, Fierer N. Sources of bacteria in outdoor air across cities in the midwestern United States. *Appl Environ Microbiol* **77**(18):6350–6356 (2011).
41. Mei Z, Wu TF, Pion-Tonachini L, Qiao W, Zhao C, Liu Z, Lo YH. Applying an optical space-time coding method to enhance light scattering signals in microfluidic devices. *Biomicrofluidics* **5**(3):34116–341166. (2011).
42. Lane MM, Czyzewski DI, Chumpitazi BP, Shulman RJ. Reliability and validity of a modified Bristol Stool Form Scale for children. *J Pediatr* **159**(3):437–441 (2011).
43. Frank DN, Zhu W, Sartor RB, Li E. Investigating the biological and clinical significance of human dysbioses. *Trends Microbiol* **19**(9):427–434 (2011).
44. Liu P, Meagher RJ, Light YK, Yilmaz S, Chakraborty R, Arkin AP, Hazen TC, Singh AK. Microfluidic fluorescence in situ hybridization and flow cytometry (μFlowFISH). *Lab Chip* **11**(16):2673–2679 (2011).
45. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**(16):2194–2200 (2011).
46. Tito RY, Belknap SL 3rd, Sobolik KD, Ingraham RC, Cleeland LM, Lewis CM Jr. Brief communication: DNA from early Holocene American dog. *Am J Phys Anthropol* **145**(4):653–657 (2011).
47. González-Rivera R, Culverhouse RC, Hamvas A, Tarr PI, Warner BB. The age of necrotizing enterocolitis onset: an application of Sartwell's incubation period model. *J Perinatol* **31**(8):519–523 (2011).
48. Brunicardi FC, Gibbs RA, Wheeler DA, Nemunaitis J, Fisher W, Goss J, Chen C. Overview of the development of personalized genomic medicine and surgery. *World J Surg* **35**(8):1693–1699 2011 (review).
49. Harring TR, Guiteau JJ, Nguyen NT, Cotton RT, Gingras MC, Wheeler DA, O'Mahony CA, Gibbs RA, Brunicardi FC, Goss JA. Building a comprehensive genomic program for hepatocellular carcinoma. *World J Surg* **35**(8):1746–1750 (2011).
50. Nguyen NT, Cotton RT, Harring TR, Guiteau JJ, Gingras MC, Wheeler DA, O'Mahony CA, Gibbs RA, Brunicardi FC, Goss JA. A primer on a hepatocellular carcinoma bioresource bank using the cancer genome atlas guidelines: Practical issues and pitfalls. *World J Surg* **35**(8):1732–1737 (2011).
51. Knights D, Kuczynski J, Charlson ES, Zaneveld J, Mozer MC, Collman RG, Bushman FD, Knight R, Kelley ST. Bayesian community-wide culture-independent microbial source tracking. *Nat Methods* **8**(9):761–763 (2011).

(*Continued*)

TABLE 1.8. (*Continued*)

52. Bergmann GT, Bates ST, Eilers KG, Lauber CL, Caporaso JG, Walters WA, Knight R, Fierer N. The under-recognized dominance of Verrucomicrobia in soil bacterial communities. *Soil Biol Biochem* **43**(7):1450–1455 (2011).
53. Ghodsi M, Liu B, Pop M. DNACLUST: Accurate and efficient clustering of phylogenetic marker genes. *BMC Bioinformatics* **12**:271 (2011).
54. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, Huttenhower C. Metagenomic biomarker discovery and explanation. *Genome Biol* **12**(6):R60 (2011).
55. Stombaugh J, Widmann J, McDonald D, Knight R. Boulder ALignment Editor (ALE): A web-based RNA alignment tool. *Bioinformatics* **27**(12):1706–1707 (2011).
56. Niu B, Zhu Z, Fu L, Wu S, Li W. FR-HIT, a very fast program to recruit metagenomic reads to homologous reference genomes. *Bioinformatics* **27**(12):1704–1705 (2011).
57. Haft DH, Basu MK. Biological systems discovery in silico: Radical S-adenosylmethionine protein families and their target peptides for posttranslational modification. *J Bacteriol* **193**(11):2745–2755 (2011).
58. Bucher BT, McDuffie LA, Shaikh N, Tarr PI, Warner BB, Hamvas A, White FV, Erwin CR, Warner BW. Bacterial DNA content in the intestinal wall from infants with necrotizing enterocolitis. *J Pediatr Surg* **46**(6):1029–1033 (2011).
59. Makarova KS, Haft DH, Barrangou R, Brouns SJ, Charpentier E, Horvath P, Moineau S, Mojica FJ, Wolf YI, Yakunin AF, et al. Evolution and classification of the CRISPR-Cas systems. *Nat Rev Microbiol* **9**(6):467–477 (2011).
60. Kong HH. Skin microbiome: Genomics-based insights into the diversity and role of skin microbes. *Trends Mol Med* **17**(6):320–328 (2011).
61. Muegge BD, Kuczynski J, Knights D, Clemente JC, González A, Fontana L, Henrissat B, Knight R, Gordon JI. Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science* **332**(6032):970–974 (2011).
62. Ye Y, Choi JH, Tang H. RAPSearch: A fast protein similarity search tool for short reads. *BMC Bioinform* **12**:159 (2011).
63. Kellermayer R, Dowd SE, Harris RA, Balasa A, Schaible TD, Wolcott RD, Tatevian N, Szigeti R, Li Z, Versalovic J, Smith CW. Colonic mucosal DNA methylation, immune response, and microbiome patterns in Toll-like receptor 2-knockout mice. *FASEB J* **25**(5):1449–1460 (2011).
64. Dominguez-Bello MG, Blaser MJ, Ley RE, Knight R. Development of the human gastrointestinal microbiota and insights from high-throughput sequencing. *Gastroenterology* **140**(6):1713–1719 (2011) (review).
65. Reeves AE, Theriot CM, Bergin IL, Huffnagle GB, Schloss PD, Young VB. The interplay between microbiome dynamics and pathogen dynamics in a murine model of Clostridium difficile Infection. *Gut Microbes* **2**(3):145–158 (2011).
66. Bates ST, Berg-Lyons D, Caporaso JG, Walters WA, Knight R, Fierer N. Examining the global distribution of dominant archaeal populations in soil. *ISME J* **5**(5):908–917 (2011).
67. DeSantis TZ, Keller K, Karaoz U, Alekseyenko AV, Singh NN, Brodie EL, Pei Z, Andersen GL, Larsen N. Simrank: Rapid and sensitive general-purpose k-mer search tool. *BMC Ecol* **11**:11 (2011).
68. Walters WA, Caporaso JG, Lauber CL, Berg-Lyons D, Fierer N, Knight R. PrimerProspector: De novo design and taxonomic analysis of barcoded polymerase chain reaction primers. *Bioinformatics* **27**(8):1159–1161 (2011).
69. Ballal SA, Gallini CA, Segata N, Huttenhower C, Garrett WS. Host and gut microbiota symbiotic factors: Lessons from inflammatory bowel disease and successful symbionts. *Cell Microbiol* **13**(4):508–517 (2011).
70. Fierer N, McCain CM, Meir P, Zimmermann M, Rapp JM, Silman MR, Knight R. Microbes do not follow the elevational diversity patterns of plants and animals. *Ecology* **92**(4):797–804 (2011).

TABLE 1.8. (*Continued*)

71. Pushalkar S, Mane SP, Ji X, Li Y, Evans C, Crasta OR, Morse D, Meagher R, Singh A, Saxena D. Microbial diversity in saliva of oral squamous cell carcinoma. *FEMS Immunol Med Microbiol* 61(3):269–277 (2011).
72. Knights D, Kuczynski J, Koren O, Ley RE, Field D, Knight R, DeSantis TZ, Kelley ST. Supervised classification of microbiota mitigates mislabeling errors. *ISME J* **5**(4):570–573 (2011).
73. Spor A, Koren O, Ley R. Unravelling the effects of the environment and host genotype on the gut microbiome. *Nat Rev Microbiol* **9**(4):279–290 (2011) (review).
74. Grice EA, Segre JA. The skin microbiome. *Nat Rev Microbiol* **9**(4):244–253 (2011) (review). [Erratum in *Nat Rev Microbiol* **9**(8):626 (2011).]
75. Hansen EE, Lozupone CA, Rey FE, Wu M, Guruge JL, Narra A, Goodfellow J, Zaneveld JR, McDonald DT, Goodrich JA, et al. Pan-genome of the dominant human gut-associated archaeon, Methanobrevibacter smithii, studied in twins. *Proc Natl Acad Sci USA*. **108(**Suppl 1):4599–4606 (2011).
76. Koren O, Spor A, Felin J, Fåk F, Stombaugh J, Tremaroli V, Behre CJ, Knight R, Fagerberg B, Ley RE, et al. Human oral, gut, and plaque microbiota in patients with atherosclerosis. *Proc Natl Acad Sci USA* 108(Suppl 1):4592–4598 (2011).
77. Ravel J, Gajer P, Abdo Z, Schneider GM, Koenig SS, McCulle SL, Karlebach S, Gorle R, Russell J, Tacket CO, et al. Vaginal microbiome of reproductive-age women. *Proc Natl Acad Sci USA* **108**(Suppl 1):4680–4687 (2011).
78. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, Fierer N, Knight R. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc Natl Acad Sci USA* **108**(Suppl 1):4516–4522 (2011).
79. Treangen TJ, Sommer DD, Angly FE, Koren S, Pop M. Next generation sequence assembly with AMOS. In *Current Protocols in Bioinformatics*, AD Baxevanis et al., eds., Wiley, 2011, Chap. 11, Unit 11.8.
80. Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward DV, Giannoukos G, Ciulla D, Tabbaa D, Highlander SK, Sodergren E, et al. Human Microbiome Consortium, (with Petrosino JF, Knight R, Birren BW). Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res* **21**(3):494–504 (2011).
81. Frank DN. Growth and Development Symposium: Promoting healthier humans through healthier livestock: Animal agriculture enters the metagenomics era. *J Anim Sci* **89**(3):835–844 (2011).
82. Wu YW, Ye Y. A novel abundance-based algorithm for binning metagenomic sequences using l-tuples. *J Comput Biol* **18**(3):523–534 (2011).
83. Bradbury AR, Sidhu S, Dübel S, McCafferty J. Beyond natural antibodies: The power of in vitro display technologies. *Nat Biotechnol* **29**(3):245–254 (2011).
84. Blainey PC, Quake SR. Digital MDA for enumeration of total nucleic acid contamination. *Nucleic Acids Res* **39**(4):e19 (2011).
85. Park J, Kerner A, Burns MA, Lin XN. Microdroplet-enabled highly parallel co-cultivation of microbial communities. *PLoS ONE* **6**(2):e17019 (2011).
86. Caporaso JG, Knight R, Kelley ST. Host-associated and free-living phage communities differ profoundly in phylogenetic composition. *PLoS ONE*. **6**(2):e16900 (2011).
87. Blainey PC, Mosier AC, Potanina A, Francis CA, Quake SR. Genome of a low-salinity ammonia-oxidizing archaeon determined by single-cell and metagenomic analysis. *PLoS ONE* **6**(2):e16626 (2011).
88. Chuang HS, Raizen DM, Lamb A, Dabbish N, Bau HH. Dielectrophoresis of Caenorhabditis elegans. *Lab Chip* **11**(4):599–604 (2011).
89. Ye Y. Identification and quantification of abundant species from pyrosequences of 16S rRNA by consensus alignment. *Proc IEEE Int Conf Bioinformatics Biomedicine*, 2/4/10, IEEE, 2011, pp. 153–157.

(*Continued*)

TABLE 1.8. (*Continued*)

90. Bates ST, Cropsey GW, Caporaso JG, Knight R, Fierer N. Bacterial communities associated with the lichen symbiosis. *Appl Environ Microbiol* **77**(4):1309–1314 (2011).
91. Wardwell LH, Huttenhower C, Garrett WS. Current concepts of the intestinal microbiota and the pathogenesis of infection. *Curr Infect Dis Rep* **13**(1):28–34 (2011).
92. Hu S, Dong TS, Dalal SR, Wu F, Bissonnette M, Kwon JH, Chang EB. The microbe-derived short chain fatty acid butyrate targets miRNA-dependent p21 gene expression in human colon cancer. *PLoS ONE* **6**(1):e16221 (2011).
93. Haft DH. Bioinformatic evidence for a widely distributed, ribosomally produced electron carrier precursor, its maturation proteins, and its nicotinoprotein redox partners. *BMC Genom* **12**:21 (2011).
94. Kalisky T, Blainey P, Quake SR. Genomic analysis at the single-cell level. *Annu Rev Genet* **45**:431–445 (2011).
95. Liu B, Gibbons T, Ghodsi M, Treangen T, Pop M. Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. *BMC Genom* **12**(Suppl 2):S4 (2011).
96. Gonzalez A, Stombaugh J, Lozupone C, Turnbaugh PJ, Gordon JI, Knight R. The mind-body-microbial continuum. *Dialog Clin Neurosci* **13**(1):55–62 (2011).
97. Parfrey LW, Walters WA, Knight R. Microbial eukaryotes in the human microbiome: ecology, evolution, and future directions. *Front Microbiol* **2**:153 (2011).
98. Young VB, Kahn SA, Schmidt TM, Chang EB. Studying the enteric microbiome in inflammatory bowel diseases: Getting through the growing pains and moving forward. *Front Microbiol* **2**:144 (2011).
99. Preidis GA, Hill C, Guerrant RL, Ramakrishna BS, Tannock GW, Versalovic J. Probiotics, enteric and diarrheal diseases, and global health. *Gastroenterology* **140**(1):8–14 (2011).
100. Caporaso JG, Lauber CL, Costello EK, Berg-Lyons D, Gonzalez A, Stombaugh J, Knights D, Gajer P, Ravel J, Fierer N, et al. Moving pictures of the human microbiome. *Genome Biol* **12**(5):R50 (2011).
101. Frank DN, Robertson CE, Hamm CM, Kpadeh Z, Zhang T, Chen H, Zhu W, Sartor RB, Boedeker EC, Harpaz N, et al. Disease phenotype and genotype are associated with shifts in intestinal-associated microbiota in inflammatory bowel diseases. *Inflamm Bowel Dis* **17**(1):179–184 (2011).
102. Haft DH, Varghese N. GlyGly-CTERM and rhombosortase: A C-terminal protein processing signal in a many-to-one pairing with a rhomboid family intramembrane serine protease. *PLoS ONE* **6**(12):e28886 (2011).
103. Schloss PD, Gevers D, Westcott SL. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS ONE* **6(**12):e27310 (2011).
104. Flores GE, Bates ST, Knights D, Lauber CL, Stombaugh J, Knight R, Fierer N. Microbial biogeography of public restroom surfaces. *PLoS ONE* **6**(11):e28132 (2011).
105. Cantarel BL, Erickson AR, VerBerkmoes NC, Erickson BK, Carey PA, Pan C, Shah M, Mongodin EF, Jansson JK, Fraser-Liggett CM, et al. Strategies for metagenomic-guided whole-community proteomics of complex microbial environments. *PLoS ONE* **6**(11):e27173 (2011).
106. Ferrara F, Listwan P, Waldo GS, Bradbury AR. Fluorescent labeling of antibody fragments using split GFP. *PLoS ONE* **6**(10):e25727 (2011).
107. Segata N, Huttenhower C. Toward an efficient method of identifying core genes for evolutionary and functional microbial phylogenies. *PLoS ONE* **6**(9):e24704 (2011).
108. Ahn J, Yang L, Paster BJ, Ganly I, Morris L, Pei Z, Hayes RB. Oral microbiome profiles: 16S rRNA pyrosequencing and microarray assay comparison. *PLoS ONE* **6**(7):e22788 (2011).
109. Lladser ME, Gouet R, Reeder J. Extrapolation of urn models via poissonization: Accurate measurements of the microbial unknown. *PLoS ONE* **6**(6):e21105 (2011).

TABLE 1.8. (*Continued*)

110. Dong Q, Nelson DE, Toh E, Diao L, Gao X, Fortenberry JD, Van der Pol B. The microbial communities in male first catch urine are highly similar to those in paired urethral swab specimens. *PLoS ONE* **6**(5):e19709 (2011).
111. Cho SH, Godin JM, Chen CH, Qiao W, Lee H, Lo YH. Recent advancements in optofluidic flow cytometer. *Biomicrofluidics* **4**(4):43001 (2010) (review).
112. Hu S, Wang Y, Lichtenstein L, Tao Y, Musch MW, Jabri B, Antonopoulos D, Claud EC, Chang EB. Regional differences in colonic mucosa-associated microbiota determine the physiological expression of host heat shock proteins. *Am J Physiol Gastrointest Liver Physiol* **299**(6):G1266–G1275 (2010).
113. Wang Y, Antonopoulos DA, Zhu X, Harrell L, Hanan I, Alverdy JC, Meyer F, Musch MW, Young VB, Chang EB. Laser capture microdissection and metagenomic analysis of intact mucosa-associated microbial communities of human colon. *Appl Microbiol Biotechnol* **88**(6):1333–1342 (2010).
114. Harrington ED, Arumugam M, Raes J, Bork P, Relman DA. SmashCell: A software framework for the analysis of single-cell amplified genome sequences. *Bioinformatics* **26**(23):2979–1980 (2010).
115. Krentz BD, Mulheron HJ, Semrau JD, Dispirito AA, Bandow NL, Haft DH, Vuilleumier S, Murrell JC, McEllistrem MT, Hartsel SC, et al. A comparison of methanobactins from Methylosinus trichosporium OB3b and Methylocystis strain Sb2 predicts methanobactins are synthesized from diverse peptide precursors modified to create a common core for binding and reducing copper ions. *Biochemistry* **49**(47):10117–10130 (2010).
116. Nelson DE, Van Der Pol B, Dong Q, Revanna KV, Fan B, Easwaran S, Sodergren E, Weinstock GM, Diao L, Fortenberry JD. Characteristic male urine microbiomes associate with asymptomatic sexually transmitted infection. *PLoS ONE* **5**(11):e14116 (2010).
117. Redford AJ, Bowers RM, Knight R, Linhart Y, Fierer N. The ecology of the phyllosphere: Geographic and phylogenetic variability in the distribution of bacteria on tree leaves. *Environ Microbiol* **12**(11):2885–2893 (2010).
118. Costello EK, Gordon JI, Secor SM, Knight R. Postprandial remodeling of the gut microbiota in Burmese pythons. *ISME J* **4**(11):1375–1385 (2010).
119. Selengut JD, Haft DH. Unexpected abundance of coenzyme F(420)-dependent enzymes in Mycobacterium tuberculosis and other actinobacteria. *J Bacteriol* **192**(21):5788–5798 (2010).
120. Rho M, Tang H, Ye Y. FragGeneScan: Predicting genes in short and error-prone reads. *Nucleic Acids Res* **38**(20):e191 (2010).
121. Wang Y, Devkota S, Musch MW, Jabri B, Nagler C, Antonopoulos DA, Chervonsky A, Chang EB. Regional mucosa-associated microbiota determine physiological expression of TLR2 and TLR4 in murine colon. *PLoS ONE*. **5**(10):e13607 (2010).
122. Goll J, Rusch DB, Tanenbaum DM, Thiagarajan M, Li K, Methé BA, Yooseph S. METAREP: JCVI metagenomics reports—an open source tool for high-performance comparative metagenomics. *Bioinformatics* **26**(20):2631–2632 (2010).
123. So A, Pel J, Rajan S, Marziali A. *Efficient Genomic DNA Extraction from Low Target Concentration Bacterial Cultures Using SCODA DNA Extraction Technology*, Cold Spring Harbor Protocol, 2010(10/1/10).
124. Dougan G, Weinstock GM. A new era in the genomics of bacteria. *Curr Opin Microbiol* **13**(5):616–618 (2010).
125. Manichanh C, Reeder J, Gibert P, Varela E, Llopis M, Antolin M, Guigo R, Knight R, Guarner F. Reshaping the gut microbiome with bacterial transplantation and antibiotic intake. *Genome Res* **20**(10):1411–1419 (2010).

(*Continued*)

TABLE 1.8. (*Continued*)

126. Gao Z, Perez-Perez GI, Chen Y, Blaser MJ. Quantitation of major human cutaneous bacterial and fungal populations. *J Clin Microbiol* **48**(10):3575–3581 (2010).
127. Chumpitazi BP, Lane MM, Czyzewski DI, Weidler EM, Swank PR, Shulman RJ. Creation and initial evaluation of a stool form scale for children. *J Pediatr* **157**(4):594–597 (2010).
128. Kuczynski J, Liu Z, Lozupone C, McDonald D, Fierer N, Knight R. Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. *Nat Meth* **7**(10):813–819 (2010).
129. McGuire AL, Beskow LM. Informed consent in genomics and genetic research. *Annu Rev Genom Hum Genet* **11**:361–381 (2010).
130. Koren S, Miller JR, Walenz BP, Sutton G. An algorithm for automated closure during assembly. *BMC Bioinformatics* **11**:457 (2010).
131. Nossa CW, Oberdorf WE, Yang L, Aas JA, Paster BJ, Desantis TZ, Brodie EL, Malamud D, Poles MA, Pei Z. Design of 16S rRNA gene primers for 454 pyrosequencing of the human foregut microbiome. *World J Gastroenterol* **16**(33):4135–4144 (2010).
132. Venkatesh M, Flores A, Luna RA, Versalovic J. Molecular microbiological methods in the diagnosis of neonatal sepsis. *Expert Rev Anti Infect Ther* **8**(9):1037–1048 (2010).
133. Reeder J, Knight R. Rapidly denoising pyrosequencing amplicon reads by exploiting rank-abundance distributions. *Nat Methods* **7**(9):668–669 (2010).
134. Cho SH, Qiao W, Tsai FS, Yamashita K, Lo YH. Lab-on-a-chip flow cytometer employing color-space-time coding. *Appl Phys Lett* **97**(9):093704 (2010).
135. Yeoman CJ, Yildirim S, Thomas SM, Durkin AS, Torralba M, Sutton G, Buhay CJ, Ding Y, Dugan-Rocha SP, Muzny DM, et al. Comparative genomics of Gardnerella vaginalis strains reveals substantial differences in metabolic and virulence potential. *PLoS ONE* **5**(8):e12411 (2010).
136. Brotman RM, Ravel J, Cone RA, Zenilman JM. Rapid fluctuation of the vaginal microbiota measured by Gram stain analysis. *Sex Transm Infect* **86**(4):297–302 (2010).
137. Wu GD, Lewis JD, Hoffmann C, Chen YY, Knight R, Bittinger K, Hwang J, Chen J, Berkowsky R, Nessel L, et al. Sampling and pyrosequencing methods for characterizing bacterial communities in the human gut using 16S sequence tags. *BMC Microbiol* **10**:206 (2010).
138. Marri PR, Paniscus M, Weyand NJ, Rendón MA, Calton CM, Hernández DR, Higashi DL, Sodergren E, Weinstock GM, Rounsley SD, et al. Genome sequencing reveals widespread virulence gene exchange among human Neisseria species. *PLoS ONE* **5**(7):e11835 (2010).
139. Erez A, Plunkett K, Sutton VR, McGuire AL. The right to ignore genetic status of late onset genetic disease in the genomic era; prenatal testing for Huntington disease as a paradigm. *Am J Med Genet A* **152A**(7):1774–1780 (2010).
140. Huse SM, Welch DM, Morrison HG, Sogin ML. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ Microbiol* **12**(7):1889–1898 (2010).
141. Bennett WE Jr, González-Rivera R, Puente BN, Shaikh N, Stevens HJ, Mooney JC, Klein EJ, Denno DM, Draghi A II, Sylvester FA, et al. Proinflammatory fecal mRNA and childhood bacterial enteric infections. *Gut Microbes* **1**(4):209–212 (2010).
142. Zaneveld JR, Lozupone C, Gordon JI, Knight R. Ribosomal RNA diversity predicts genome diversity in gut bacteria and their relatives. *Nucleic Acids Res* **38**(12):3869–3879 (2010).
143. Dominguez-Bello MG, Costello EK, Contreras M, Magris M, Hidalgo G, Fierer N, Knight R. Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc Natl Acad Sci USA* **107**(26):11971–11975 (2010).

TABLE 1.8. (*Continued*)

144. Cho SH, Chen CH, Tsai FS, Godin JM, Lo YH. Human mammalian cell sorting using a highly integrated micro-fabricated fluorescence-activated cell sorter (microFACS). *Lab Chip* **10**(12):1567–1573 (2010).
145. Harwich MD Jr, Alves JM, Buck GA, Strauss JF 3rd, Patterson JL, Oki AT, Girerd PH, Jefferson KK. Drawing the line between commensal and pathogenic Gardnerella vaginalis through genome analysis and virulence studies. *BMC Genomi* **11**:375 (2010).
146. Pei AY, Oberdorf WE, Nossa CW, Agarwal A, Chokshi P, Gerz EA, Jin Z, Lee P, Yang L, Poles M, et al. Diversity of 16S rRNA genes within individual prokaryotic genomes. *Appl Environ Microbiol* **76**(12):3886–3897. [erratum in *Appl Environ Microbiol* **76**(15):5333 (2010)].
147. Lauber CL, Zhou N, Gordon JI, Knight R, Fierer N. Effect of storage conditions on the assessment of bacterial community structure in soil and human-associated samples. *FEMS Microbiol Lett* **307**(1):80–86 (2010).
148. Kong HH, Segre JA. Bridging the translational research gap: A successful partnership involving a physician and a basic scientist. *J Invest Dermatol* **130**(6):1478–1480 (2010).
149. Wang GP, Sherrill-Mix SA, Chang KM, Quince C, Bushman FD. Hepatitis C virus transmission bottlenecks analyzed by deep sequencing. *J Virol* **84**(12):6218–6228 (2010).
150. Pati A, Ivanova NN, Mikhailova N, Ovchinnikova G, Hooper SD, Lykidis A, Kyrpides NC. GenePRIMP: A gene prediction improvement pipeline for prokaryotic genomes. *Nat Meth* **7**(6):455–457 (2010).
151. Haft DH, Basu MK, Mitchell DA. Expansion of ribosomally produced natural products: a nitrile hydratase- and Nif11-related precursor family. *BMC Biol* **8**:70 (2010).
152. Human Microbiome Jumpstart Reference Strains Consortium, Nelson KE, Weinstock GM, Highlander SK, Worley KC, Creasy HH, Wortman JR, Rusch DB, Mitreva M, Sodergren E, Chinwalla AT, et al. A catalog of reference genomes from the human microbiome. *Science* **328**(5981):994–999 (2010).
153. Cho SH, Chen CH, Tsai FS, Godin J, Lo YH. Mammalian cell sorting using µFACS, *Proc Conf Lasers Electro Optics*, **2010**:CTuD1 (2010).
154. Thomas CM, Versalovic J. Probiotics-host communication: Modulation of signaling pathways in the intestine. *Gut Microbes* **1**(3):148–163 (2010) (review).
155. Lewis T, Loman NJ, Bingle L, Jumaa P, Weinstock GM, Mortiboy D, Pallen MJ. High-throughput whole-genome sequencing to dissect the epidemiology of Acinetobacter baumannii isolates from a hospital outbreak. *J Hosp Infect* **75**(1):37–41 (2010).
156. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, et al. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7**(5):335–336 (2010).
157. McGuire AL, Lupski JR. Personal genome research: What should the participant be told? *Trends Genet* **26**(5):199–201 (2010).
158. Blaser MJ. Harnessing the power of the human microbiome. *Proc Natl Acad Sci USA* **107**(14):6125–6126 (2010).
159. Fierer N, Lauber CL, Zhou N, McDonald D, Costello EK, Knight R. Forensic identification using skin bacterial communities. *Proc Natl Acad Sci USA* **107**(14):6477–6481 (2010).
160. Fujimura KE, Slusher NA, Cabana MD, Lynch SV. Role of the gut microbiota in defining human health. *Expert Rev Anti Infect Ther* **8**(4):435–454 (2010) (review).
161. White JR, Navlakha S, Nagarajan N, Ghodsi MR, Kingsford C, Pop M. Alignment and clustering of phylogenetic markers—implications for microbial diversity studies. *BMC Bioinformatics* **11**:152 (2010).
162. Jones RT, Knight R, Martin AP. Bacterial communities of disease vectors sampled across time, space, and species. *ISME J* **4**(2):223–231 (2010).

(*Continued*)

TABLE 1.8. (*Continued*)

163. Selengut JD, Rusch DB, Haft DH. Sites inferred by metabolic background assertion labeling (SIMBAL): Adapting the partial phylogenetic profiling algorithm to scan sequences for signatures that predict protein function. *BMC Bioinformatics* **11**:52 (2010).
164. Caporaso JG, Bittinger K, Bushman FD, DeSantis TZ, Andersen GL, Knight R. PyNAST: A flexible tool for aligning sequences to a template alignment. *Bioinformatics* **26**(2):266–267 (2010).
165. Kuczynski J, Costello EK, Nemergut DR, Zaneveld J, Lauber CL, Knights D, Koren O, Fierer N, Kelley ST, Ley RE, et al. Direct sequencing of the human microbiome readily reveals community differences. *Genome Biol* **11**(5):210 (2010) (review).
166. Zhou X, Brotman RM, Gajer P, Abdo Z, Schüette U, Ma S, Ravel J, Forney LJ. Recent advances in understanding the microbiology of the female reproductive tract and the causes of premature birth. *Infect Dis Obstet Gynecol* **2010**:737425 (2010) (review).
167. Hamady M, Lozupone C, Knight R. Fast UniFrac: Facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *ISME J* **4**(1):17–27 (2010).
168. Salzman NH, Hung K, Haribhai D, Chu H, Karlsson-Sjöberg J, Amir E, Teggatz P, Barman M, Hayward M, Eastwood D, et al. Enteric defensins are essential regulators of intestinal microbial ecology. *Nat Immunol* **11**(1):76–83 (2010).
169. Chen CH, Cho SH, Tsai F, Erten A, Lo YH. Microfluidic cell sorter with integrated piezoelectric actuator. *Biomed Microdevices* **11**(6):1223–1231 (2009).
170. NIH HMP Working Group, Peterson J, Garges S, Giovanni M, McInnes P, Wang L, Schloss JA, Bonazzi V, McEwen JE, Wetterstrand KA, Deal C, et al. The NIH Human Microbiome Project. *Genome Res* **19**(12):2317–2323 Epub (2009).
171. Lazarevic V, Whiteson K, Huse S, Hernandez D, Farinelli L, Osterås M, Schrenzel J, François P. Metagenomic study of the oral microbiota by Illumina high-throughput sequencing. *J Microbiol Meth* **79**(3):266–271 (2009).
172. Blaser MJ, Falkow S. What are the consequences of the disappearing human microbiota? *Nat Rev Microbiol* **7**(12):887–894 (2009).
173. Hildebrandt MA, Hoffmann C, Sherrill-Mix SA, Keilbaugh SA, Hamady M, Chen YY, Knight R, Ahima RS, Bushman F, Wu GD. High-fat diet determines the composition of the murine gut microbiome independently of obesity. *Gastroenterology* **137**(5):1716–1724, e1–2 (2009).
174. Sillanpää J, Nallapareddy SR, Qin X, Singh KV, Muzny DM, Kovar CL, Nazareth LV, Gibbs RA, Ferraro MJ, Steckelberg JM, et al. A collagen-binding adhesin, Acb, and ten other putative MSCRAMM and pilus family proteins of Streptococcus gallolyticus subsp. gallolyticus (Streptococcus bovis group, biotype I). *J Bacteriol* **191**(21):6643–6653 (2009).
175. Ghodsi M, Pop M. Inexact local alignment search over suffix arrays. *Proc (IEEE Int Conf Bioinformatics Biomed)*, 11/1/9, IEEE, 2009, pp. 83–97.
176. Frank DN. BARCRAWL and BARTAB: Software tools for the design and implementation of barcoded primers for highly multiplexed DNA sequencing. *BMC Bioinformatics* **10**:362 (2009).
177. Chain PS, Grafham DV, Fulton RS, Fitzgerald MG, Hostetler J, Muzny D, Ali J, Birren B, Bruce DC, Buhay C, et al. Genomic Standards Consortium Human Microbiome Project Jumpstart Consortium, Detter JC. Genomics. Genome project standards in a new era of sequencing. *Science* **326**(5950):236–237 (2009).
178. Bennett WE Jr, González-Rivera R, Shaikh N, Magrini V, Boykin M, Warner BB, Hamvas A, Tarr PI. A method for isolating and analyzing human mRNA from newborn stool. *J Immunol Meth* **349**(1–2):56–60 (2009).

TABLE 1.8. (*Continued*)

179. Loman NJ, Snyder LA, Linton JD, Langdon R, Lawson AJ, Weinstock GM, Wren BW, Pallen MJ. Genome sequence of the emerging pathogen Helicobacter canadensis. *J Bacteriol* **191**(17):5566–5567 (2009).
180. Brady A, Salzberg SL. Phymm and PhymmBL: Metagenomic phylogenetic classification with interpolated Markov models. *Nat Methods* **6**(9):673–676 (2009).
181. Pel J, Broemeling D, Mai L, Poon HL, Tropini G, Warren RL, Holt RA, Marziali A. Nonlinear electrophoretic response yields a unique parameter for separation of biomolecules. *Proc Natl Acad Sci USA* **106**(35):14796–14801 (2009).
182. Yang L, Lu X, Nossa CW, Francois F, Peek RM, Pei Z. Inflammation and intestinal metaplasia of the distal esophagus are associated with alterations in the microbiome. *Gastroenterology* **137**(2):588–597 (2009).
183. Wang Y, Hoenig JD, Malin KJ, Qamar S, Petrof EO, Sun J, Antonopoulos DA, Chang EB, Claud EC. 16S rRNA gene-based analysis of fecal microbiota from preterm infants with and without necrotizing enterocolitis. *ISME J* **3**(8):944–954 (2009).
184. Ye Y, Doak TG. A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. *PLoS Comput Biol* **5**(8):e1000465 (2009).
185. Pop M. Genome assembly reborn: recent computational challenges. *Brief Bioinform* **10**(4):354–366 (2009).
186. Ye Y, Tang H. An ORFome assembly approach to metagenomics sequences analysis. *J Bioinform Comput Biol* **7**(3):455–471 (2009).
187. Petrosino JF, Highlander S, Luna RA, Gibbs RA, Versalovic J. Metagenomic pyrosequencing and microbial identification. *Clin Chem* **55**(5):856–866 (2009) (review).
188. Mahowald MA, Rey FE, Seedorf H, Turnbaugh PJ, Fulton RS, Wollam A, Shah N, Wang C, Magrini V, Wilson RK, et al. Characterizing a model human gut microbiota composed of members of its two dominant bacterial phyla. *Proc Natl Acad Sci USA* **106**(14):5859–5864 (2009).
189. Sharp RR, Achkar JP, Brinich MA, Farrell RM. Helping patients make informed choices about probiotics: A need for research. *Am J Gastroenterol* **104**(4):809–813 (2009) (review).
190. White JR, Nagarajan N, Pop M. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput Biol* **5**(4):e1000352 (2009).
191. Cho SH, Chen CH, Tsai FS, Lo YH. Micro-fabricated fluorescence-activated cell sorter. *Proc Conf IEEE Eng Med Biol Soc IEEE*, 2009, pp. 1075–1078.
192. Langmead B, Schatz MC, Lin J, Pop M, Salzberg SL. Searching for SNPs with cloud computing. *Genome Biol* **10**(11):R134 (2009).
193. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**(3):R25 (2009).
194. Wooley JC, Ye Y. Metagenomics: Facts and artifacts, and computational challenges. *J Comput Sci Technol* **25**(1):71–81 (2009).
195. Pei A, Nossa CW, Chokshi P, Blaser MJ, Yang L, Rosmarin DM, Pei Z. Diversity of 23S rRNA genes within individual prokaryotic genomes. *PLoS ONE* **4**(5):e5437 (2009).

1.5. OTHER NIH-SUPPORTED HUMAN MICROBIOME RESEARCH ACTIVITIES

One of the long-term goals of the Human Microbiome Project was to provide data for the biomedical research community regarding the role of the microbiome in human health and in disease. It was hoped that the results from the HMP would encourage additional investments in the field.

Support for human microbiome research has clearly grown at the NIH. Support from 9 microbiome-specific RFAs/PAs (Table 1.9) included 34 projects funded for a total of $36M over 5 years. This analysis does not include 8 microbiome-specific RFAs/PAs that are active as of this writing or three other active RFAs/PAs in which the awards are yet to be made, so this is a conservative estimate of support for microbiome research across the NIH.

Further, growth of research support across the NIH institutes and centers (ICs) has been somewhat organic as most of the microbiome grants have been individual investigator-initiated projects and not written in response to specific RFAs or PAs. Starting from basal levels and only one or two ICs over 2005–2006, the levels of support for human microbiome research and the number of ICs engaged in support of this research notably increased over the next 5-year period to 2009–2010.

In order to classify the diseases under study in those projects investigating microbiome and disease associations, the WHO International Classification of Diseases 2010 (ICD10) system (`http://www.who.int/classifications/icd/en`) was used to categorize the diseases. Ten NIH ICs focused their microbiome association studies on 9 of the 22 ICD major categories of disease and related health problems (Table 1.10) and included 38 different kinds of specific diseases (Table 1.11). Together, NIDCR, NHLBI, and NIDDK supported microbiome association studies of >60% of these 38 specific diseases, which fell into 6 of the 9 ICD major disease categories, including diseases of the digestive system; infectious and parasitic diseases; neoplasms; diseases of the respiratory system; endocrine, nutritional and

TABLE 1.9. Previously Funded NIH Microbiome-Related RFAs and PAs Not Part of the HMP[a]

RFA/PA Title	RFA/PA Number	Total Awarded ($ M)	Number of Projects
Metagenomic Analyses of the Oral Microbiome (R01)	PA04-131	3.6	1
Partnerships to Develop Tools to Evaluate Women's Health	AI05-029	4	1
New Approaches for the Prevention and Treatment of Necrotizing Enterocolitis (R01)	HD07-018	5.6	8
Microbicide Innovation Program (MIP III) (R21/R23)	AI07-034	0.96	1
Metagenomic Analyses of the Oral Microbiome (R01)	PA08-090	0.75	1
Microbiome of the Lung and Respiratory Tract in HIV-Infected Individuals and HIV-Uninfected Controls (U01)	HL09-006	10.7	7
Enterics Research Investigational Network Cooperative Research Centers (U19)	AI09-023	4.3	4
Metagenomic Evaluation of Oral Polymicrobial Disease (R01)	DE10-003	4.6	8
Gut–Liver–Brain Interactions in Alcohol-Induced Pathogenesis (R01)	AA10-007	1.3	3
Total		35.81	34

[a] As of fiscal year 2010.

TABLE 1.10. General Disease Categories for Non-HMP Microbiome and Disease Association Projects[a]

Category	NIAAA	NIAID	NIAMS	NCCAM	NCI	NIDCR	NIDDK	NICHD	NHLBI	NINR	Total
Diseases of digestive system	2	—	—	1	—	4	2	—	—	—	9
Diseases of genitourinary system	—	2	—	—	—	—	—	—	—	—	2
Certain infectious and parasitic diseases	—	2	—	—	—	2	—	—	5	—	9
Injury, poisoning, and certain other consequences of external causes	—	1	—	—	—	—	—	—	—	—	1
Diseases of musculoskeletal system and connective tissue	—	—	1	—	—	—	—	—	—	—	1
Neoplasms	—	—	—	—	1	2	—	—	—	—	3
Diseases of respiratory system	—	—	—	—	—	1	—	—	1	—	2
Endocrine, nutritional and metabolic diseases	—	—	—	—	—	—	5	—	1	—	6
Certain conditions originating in the perinatal period	—	—	—	—	—	—	—	3	1	1	5
Total	2	5	1	1	1	9	7	3	8	1	38

[a] Indicated in figure 5 of the WHO International Disease Classification (ICD) system.

TABLE 1.11. Specific Disease Categories for Non-HMP Microbiome and Disease Association Projects[a]

Disease Category	NIAAA	NIAID	NIAMS	NCCAM	NCI	NIDCR	NIDDK	NICHD	NHLBI	NINR
Diseases of digestive system	Alcoholic liver disease	—	—	Inflammatory bowel disease	—	Periodontitis; dental caries; oral mucositis	Inflammatory bowel disease; nonalcholic fatty liver disease	—	—	—
Diseases of the genitourinary system	—	Bacterial vaginosis; pelvic inflammatory disease	—	—	—	—	—	—	—	—
Certain infectious and parasitic diseases	—	Enteric disease	—	—	—	HIV, SIV	—	—	HIV	—
Injury, poisoning, and certain other consequences of external causes	—	Food allergies	—	—	—		—	—	—	—
Diseases of musculoskeletal system and connective tissue	—	—	Rheumatoid arthritis	—	—		—	—	—	—
Neoplasms	—	—	—	—	Colon cancer	Head and neck cancer	—	—	—	—
Diseases of respiratory system	—	—	—	—	—	Chronic rhinitis	—	—	Chronic obstructive pulmonary disease	—
Endocrine, nutritional and metabolic diseases	—	—	—	—	—		Obesity	—	Cystic fibrosis	—
Certain conditions originating in the perinatal period	—	—	—	—	—		—	Necrotizing enterocolitis	Bronchopulmonary dysplasia	Necrotizing enterocolitis

[a] Indicated in figure 5 of the WHO International Disease Classification (ICD) system for fiscal year 2010.

metabolic diseases; and conditions originating in the perinatal period (Table 1.10). As a point of comparison, in the HMP, necrotizing enterocolitits, inflammatory bowel disease, and bacterial vaginosis are being studied in the HMP demonstration projects (Table 1.3).

1.6. FUTURE DIRECTIONS FOR HUMAN MICROBIOME RESEARCH

The field of human microbiome research is at a pivotal point. The NIH Human Microbiome Project has provided an extensive resource of datasets, tools, and protocols for the study of both the healthy microbiome and the microbiomes of a diversity of diseases. It has considered some of the ethical issues that microbiome research may raise and in some cases has provided guidance for handling those issues. Here, we mention some key foundational studies and resources needed to move the field forward. Although institutes and agencies with disease-specific interests have already initiated mission-focused research programs, some suggestions are made for well-placed investments to accelerate progress in the field. Some of these ideas have also been outlined in a commentary [22].

We now know that the microbiome is acquired anew from the environment at birth and that the maturing immune system [23] and successional stages in the assembling microbial community interact to establish the microbiome in the first 2–3 years of life [24–26]. But why and how does the microbiome mature over these 2–3 years? And are there other fundamental changes that continue into adulthood? For example, what is the effect of hormonal changes, if any, at puberty or at menopause on the microbiome? We do not yet have a mechanistic understanding of the roles of the source inoculum in the maturing microbiome, the host immune system in regulating colonization by specific members of the microbiome, the successional events that result in a mature microbiome, the roles of the microbiota in resisting colonization by new microbes, or other host genetic factors in the selection of microbial composition of the microbiome.

Further, studies are suggesting that contemporary practices such as delivery by Caesarean section versus vaginal birth [27] and formula use versus breastfeeding [28,29] may affect the communities that assemble in the infant and therefore have an impact on the growth of beneficial microbes, particularly in the gut microbiome. In addition, the current practice of antibiotic use in mothers giving Caesarean birth and in infants and children appears to impact the microbiota, and this impact seems to last for months to years after antibiotic use [30]. Also, we do not yet understand the role of the early microbiome in the composition and function of the microbiome throughout life and in the development of later disorders or disease. For example, some studies suggest that a disturbed microbiome at infancy, through antibiotic use, may predispose one to allergies later in childhood [31]. In fact, a number of disorders (e.g. Crohn's disease, asthma, hay fever, type 1 diabetes, inflammatory bowel disease, multiple sclerosis, autism, celiac disease) may be associated with a disturbed, altered, or impoverished microbiome at infancy [32–36]. A working hypothesis for these observations of an association between a disturbed microbiome and subsequent disease is that induction of immune system maturation in these infants may be delayed, thereby rendering them more susceptible to diseases later in life.

A foundational study of microbiome development from birth through early childhood is needed. These studies should include the mother's microbiome and those of the child's immediate family. This study will need to include parameters beyond microbial community composition. In fact, what seems to be emerging from early human microbiome studies is that there is less diversity in the major metabolic pathways of the microbiome than in the diversity of the microbiome community that carry out these activities [37]. 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. Technologies are now becoming available for the study of microbiome function, such as metabolomics, metatranscriptomics, and metaproteomics, all of which capture different aspects of the activities of the microbiota. Development of these large-scale methodologies to become high-throughput technologies will be an important resource for the field. Along with these technologies, additional approaches are needed to measure strain-level functional properties as well as the host immune system responses to microbial signals.

It appears that the microbiome retains much of its dynamic quality throughout life [38] and suggests that we may not yet understand what constitutes a healthy microbiome, particularly over the lifetime of an individual [39], even in Western populations. Some of the dynamic quality of the microbiome may be due to the factors in play during establishment of the early microbiome. Other studies are showing that diet is key to microbiome dynamics throughout life. In fact, few studies have broadly sampled the human population, particularly populations of non-European ancestry, to capture the breadth of these factors, and more work is needed to define the factors that regulate microbiome stability in life.

Further, no major microbiome study has yet included genetic analysis of the host. It is imperative that we begin to include host genomics in our efforts to understand what factors control and affect the microbiome. Human microbiome research raises its own unique questions about ethical issues such as return of research results, intellectual property, and ownership of the microbiome materials and confidentiality [40–42]. Efforts should be directed toward educating the public about the role of the microbiome in health and the need for volunteers in clinical studies of the microbiome to be broadly accepted so that both the host factors and microbiome factors can be included in these studies.

Although the microbiome of each body region is important to the health of that region, the gut microbiome could arguably be considered the "cardinal microbiome" as this is the microbial community that contributes to food digestion, directly supplies energy for host cell metabolism, and directly interacts with the host immune system [43]. Studies of diet and microbiome composition verify that microbiome composition appears closely associated with long term dietary practices [44,45] but not short-term diet changes [46]. The gut microbiome also directly and indirectly communicates with the microbiomes of other body regions through signaling molecules of microbial origin that circulate throughout the body. A concerted effort to study, the relationship between diet and the gut microbiome would be an important foundational study as would an effort to understand the systemic role of the gut microbiome and how it interacts with the organ systems and with the microbiomes across the human body.

Perhaps one of the most effective means of addressing these key areas would be through large cohort studies that include racially and ethnically diverse populations. These studies would serve as the foundation from which numerous studies could address the properties during microbiome assembly, variability of the microbiome across populations, and through time. Studies would follow the successional development of microbiome composition and also the changing functional properties of the microbiome as it matures. Opportunities to integrate microbiome studies with large cohort studies may now become available. There are a number of longitudinal birth cohort studies in place or planned in the near future that may serve as models in this endeavor. For example, the National Children's Study (NCS) (http://www.nationalchildrensstudy.gov/Pages/default.aspx) is a multidisciplinary US agency study designed to examine the environmental factors (broadly defined as diet, genetics, and other factors) that affect growth, development, and health of children in the United States in a prospective study from birth to adulthood. The NCS is designed as a platform to enable research and birth cohorts are being recruited across the country as the basis of the study. There are plans to collect samples and data from pregnant mothers, their subsequent newborns, and immediate family members, including siblings, family pets, and the immediate environment of the child. The goal of the program is to recruit up to 100,000 children over the course of the full study and follow them to 21 years of age. The study is sampling a representative distribution of the US population, so it can expect to include many racial and ethnic groups in the full study. The NCS full study is scheduled to begin in 2014. Large international birth cohort studies are also underway, such as the French Longitudinal Study of Children, is a cohort study of 20,000 children followed from birth to adulthood (ELFE; http://www.cls.ioe.ac.uk/text.asp?section=00010001000500090016; http://www.biomedcentral.com/1471-2431/9/58). An international consortium of scientists is proposing to collect stool samples for microbiome analysis of these 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. Proposals have been made to include microbiome sampling and analysis of the children in this study. Partnerships with other large birth cohort studies would be invaluable, as would partnerships with other large cohort studies where the subjects have consented broadly and where genomic and phenotype data are available. With appropriate coordination and consents from the study participants, such studies could provide the ideal framework from which to analyze the microbiome and its functional properties from time of birth across diverse populations.

In order for the results from these studies to support the research interests of the broadest community, these activities will require a flexible and user-friendly infrastructure that links all of the different microbiome datasets, which include microbial composition and microbial function and the host phenotype and genotype data with appropriate, ready-to use computational tools for analysis that are accessible to all regardless of the bioinformatic resources or computational expertise at one's home institution. Massive microbiome datasets of terabase and petabase sizes will be the order of the day in the very near future. New approaches and tools are needed that can accommodate these large datasets and support data transfer, analysis, and interpretation. In fact, it will be the routine access and use of this network of data and tools by a broader community that will move this field into the clinical realm as microbiome and related data are applied to questions in the treatment of disease and in the support of health. A broadly available resource that will support the needs of both the research community and the clinical community will be crucial to the full integration of the microbiome into scientific and clinical studies and is a fundamental community resource.

The microbiome coevolved with its human host over millennia. Studies from the field of human anthropology are suggesting that, over the last 2 million years, there were several waves of early hominid migrations from the African continent across the globe. More recent comparative genomics studies of the Denisovans [47], who started migrating ~1,000,000 years ago, the Neanderthals [48], who started migrating ~600,000 years ago, and, *Homo sapiens*, who started migrating ~100,000–200,000 years ago, suggest there was interbreeding between these early hominid species and *H. sapiens*. Perhaps more importantly, a more recent study provided evidence that this interbreeding may have conferred increased fitness to our species, particularly in the *H. sapiens* immune system. Analysis of a key group of immune system genes, the human leukocyte antigen (HLA) class I genes, shows these genes are particularly variable with thousands of alleles across different populations. Analyses suggest that the HLA gene alleles spread rapidly from early hominids to *H. sapiens* as a large fraction (~50–95%) of these alleles appear to be derived from early hominids [49]. But it may not only be heritable traits that were acquired during this interbreeding. The earlier hominid microbiomes may have been very different from the microbiomes of migrating *H. sapiens* as diets, time elapsed since migration across many biomes, and environmental exposures would have likely been quite different between the early hominid groups and *H. sapiens*. It is quite conceivable that interbreeding also led to the acquisition of new beneficial microbes, leading to a more robust microbiome for *H. sapiens*, conferring the ability to defend against new opportunistic pathogens and to diversify the diet. These studies suggest that human microbiome studies should be conducted with this evolutionary context in mind.

More recent reviews of human microbiome studies argue that the field will need to move beyond an understanding of the fundamental properties of microbiome composition to an understanding of the fundamental properties of microbiome function if the microbiome is to be integrated into the study of human health and disease. Future microbiome function studies should include diverse populations in order to circumscribe and associate the functional properties of the microbiome with other features of these populations. Collaboration with large cohort studies, particularly birth cohorts of diverse populations, may be one means of focusing such an effort in order to develop the resources needed to study the role of the microbiome in health and in disease. Development of high-through put methodologies to measure microbiome function in conjunction with large cohort studies, all of which are supported by a well-designed, user-friendly infrastructure, will establish the needed resources and data for future research of the microbiome in health and in disease. Finally, it is important to ground human microbiome studies in the appropriate evolutionary and ecological context if we are to understand the drivers behind microbiome assembly, homeostasis, and its role in human health maintenance.

ACKNOWLEDGMENTS

We would like to acknowledge numerous individuals outside the NIH who provided data, figures, suggestions, and edits on the chapter. The HMP research network consortium members who contributed willingly to data and figures include Drs. Ashlee Earl (Broad Institute) and Heather Huot-Creasy (DACC/Univ MD School of Medicine), who provided the data for Figure 1.2. Dr. Earl also provided extensive comments on an earlier version of the chapter. Drs. Dirk Gevers and Katherine Huang of the Broad Institute provided the data and figures for Figure 1.4. Dr. Gevers and also Drs. Rob Knight of the University of Colorado, Curtis Huttenhower of the Harvard School of Public Health, and Owen White of the University of Maryland School of Medicine also provided valuable comments on the chapter draft. Any errors are solely ours to claim.

REFERENCES

1. Woese CR, Fox GE. *Proc Natl Acad Sci USA* **74**:5088–5090 (1977).
2. Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, Pace NR. *Proc Natl Acad Sci USA* **82**(20):6955–6959 (1985).
3. Choi BK, Paster BJ, Dewhirst FE, Gōbel UB. *Infect Immun* **62**:1889–1895 (1994).
4. Ashimoto A, Chen C, Bakker I, Slots J. *Oral Microbiol. Immunol* **11**:266–273 (1996).
5. Kroes I, Lepp PW, Relman DA. *Proc Natl Acad Sci USA* **96**:14547–14552 (1999).
6. Brogden KA, Guthmiller JM. *Polymicrobial Diseases*, ASM Press, 2002.
7. Lederberg J. *Science* **288**:287–293 (2000).
8. Falk PG, Hooper LV, Midtvedt T, Gordon JI. *Microbiol Molec Biol Rev* **62**:1157–1170 (1998).
9. Hooper LV, Gordon JI. *Science* **292**:1115–1118 (2001).
10. Macpherson AJ, Harris NL. *Nat Rev Immunol* **4**:478–485 (2004).
11. Macpherson AJ, Gatto D, Sainsbury E, Harriman GR, Hengartner H, Zinkemagel RM. *Science* **288**:2222–2226 (2000).
12. Hooper LV. *Trends Microbiol* **12**:129–134 (2004).
13. Relman A, Falkow S. *Trends Microbiol* **9**:206–208 (2001).
14. Davies J. *Science* **291**:2316 (2001).
15. Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, Gill SR, Nelson KE, Relman DA. *Science* **308**:1635–1638 (2005).
16. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE. *Science* **312**:1355–1359 (2006).
17. National Research Council. *The New Science of Metagenomics*, (2007). ISBN-10 0-309-10676-1.
18. Toronto International Data Release Workshop Authors. *Nature* **461**:168–170 (2009).
19. NIH HMP Working Group. *Genome Res* **19**(12):2317–2323 (2009).
20. Human Microbiome Jumpstart Reference Strains Consortium. *Science* **328**(5981):994–999 (2010).
21. Chain PS, Grafham DV, Fulton RS, Fitzgerald MG, Hostetler J, Muzny D, Ali J, Birren B, Bruce DC, Buhay C et al. *Science* **326**:236–237 (2009).
22. Proctor LM. *Cell Host Microbe* **10**:287–291 (2011).
23. Mazmanian SK, Liu CH, Tzianabos AO, Kasper DL. *Cell* **122**:107–118 (2005).
24. Dethlefsen L. *Trends Ecol Evol* **21**:517–523 (2006).
25. Palmer C, Bik EM, DiGiulio DB, Relman DA, Brown PO. *PLoS Biol* **5**(7):e177 (2007).
26. Koenig GM Jr, Lin IH, Abbott NL. *Proc Natl Acad Sci USA*. **107**:3998–4003 (2010).
27. Dominguez-Bello MG, Costello EK, Contreras M, Magris M, Hidalgo G, Fierer N, Knight R. *Proc Natl Acad Sci USA*. **107**:11971–11975 (2010).
28. Poroyko V, White JR, Wang M, Donovan S, Alverdy J, Liu DC, Morowitz MJ. *PLoS One*. **5**:e12459 (2010).

29. Sánchez E, De Palma G, Capilla A, Nova E, Pozo T, Castillejo G, Varea V, Marcos A, Garrote JA, Polanco I, López A, Ribes-Koninckx C, García-Novo MD, Calvo C, Ortigosa L, Palau F, Sanz Y. *Appl Environ Microbiol* **77**(15):5316–5323 (2011).
30. Jernberg C, Löfmark S, Edlund C, Jansson JK. *Microbiology* **156**(Pt 11):3216–3223 (2010).
31. Bisgaard H, Li N, Bonnelykke K, Chawes BL, Skov T, Paludan-Müller G, Stokholm J, Smith B, Krogfelt KA. *J Allergy Clin Immunol.* **128**:646–652 (2011).
32. Finegold SM, Molitoris D, Song Y, Liu C, Vaisanen ML, Bolte E, McTeague M, Sandler R, Wexler H, Marlowe EM, Collins MD, Lawson PA, Summanen P, Baysallar M, Tomzynski TJ, Read E, Johnson E, Rolfe R, Nasir P, Shah H, Haake DA, Manning P, Kaul A. *Clin Infect Dis* **35**(Suppl 1):S6–S16 (2002).
33. Kummeling I, Stelma FF, Dagnelie PC, Snijders BE, Penders J, Huber M, van Ree R, van den Brandt PA, Thijs C. *Pediatrics* **119**(1):e225–e231 (2007).
34. Ponsonby AL, Catto-Smith AG, Pezic A, Dupuis S, Halliday J, Cameron D, Morley R, Carlin J, Dwyer T. *Inflamm Bowel Dis* **15**(6):858–866 (2009).
35. Ochoa-Repáraz J, Mielcarz DW, Ditrio LE, Burroughs AR, Foureau DM, Haque-Begum S, Kasper LH. *J Immunol* **183**(10):6041–6050 (2009).
36. Decker E, Engelmann G, Findeisen A, Gerner P, Laass M, Ney D, Posovszky C, Hoy L, Hornef MW. *Pediatrics* **125**(6):e1433–e1440 (2010).
37. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T et al. *Nature* **464**:59–65 (2010).
38. Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R. *Science* **326**:1694–1697 (2009).
39. Claesson MJ, Cusack S, O'Sullivan O, Greene-Diniz R, de Weerd H, Flannery E, Marchesi JR, Falush D, Dinan T, Fitzgerald G et al. *Proc Natl Acad Sci USA*. **108**:4586–4591 (2011).
40. Hawkins AK, O'Doherty KC. *BMC Medical Genomics* **4**:72.
41. Blaser MJ. *Proc Natl Acad Sci USA*. **107**:6125–6126 (2010).
42. McGuire AL, Colgrove J, Whitney SN, Diaz CM, Bustillos D, Versalovic J. *Genome Res* **18**(12):1861–1864 (2008).
43. Mazmanian SK, Round JL, Kasper DL. *Nature*. **453**:620–625 (2008).
44. Turnbaugh PJ, Ridaura VK, Faith JJ, Rey FE, Knight R, Gordon JI. *Sci Transl Med* **1**:6ra14 (2009).
45. De Filippo C, Cavalieri D, Di Paola M, Ramazzotti M, Poullet JB, Massart S, Collini S, Pieraccini G, Lionetti P. *Proc Natl Acad Sci USA*. **107**:14691–14695 (2010).
46. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen YY, Keilbaugh SA, Bewtra M, Knights D, Walters WA, Knight R et al. *Science.* **334**:105–108 (2011).
47. Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PL, Maricic T, Good JM, Marques-Bonet T, Alkan C, Fu Q, Mallick S, Li H, Meyer M, Eichler EE, Stoneking M, Richards M, Talamo S, Shunkov MV, Derevianko AP, Hublin JJ, Kelso J, Slatkin M, Pääbo S. *Nature* **468**(7327):1053–1060 (2010).
48. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH, Hansen NF, Durand EY, Malaspinas AS, Jensen JD, Marques-Bonet T, Alkan C, Prüfer K, Meyer M, Burbano HA, Good JM, Schultz R, Aximu-Petri A, Butthof A, Höber B, Höffner B, Siegemund M, Weihmann A, Nusbaum C, Lander ES, Russ C, Novod N, Affourtit J, Egholm M, Verna C, Rudan P, Brajkovic D, Kucan Z, Gusic I, Doronichev VB, Golovanova LV, Lalueza-Fox C, de la Rasilla M, Fortea J, Rosas A, Schmitz RW, Johnson PL, Eichler EE, Falush D, Birney E, Mullikin JC, Slatkin M, Nielsen R, Kelso J, Lachmann M, Reich D, Pääbo S. *Science* **328**(5979):710–722 (2010).
49. Abi-Rached L, Jobin MJ, Kulkarni S, McWhinnie A, Dalva K, Gragert L, Babrzadeh F, Gharizadeh B, Luo M, Plummer FA et al. *Science.* **334**:89–94 (2011).