A variety of microbial communities and their genes (the microbiome) exist throughout the human body, with fundamental roles in human health and disease. The National Institutes of Health (NIH)-funded Human Microbiome Project Consortium has established a population-scale framework to develop metagenomic protocols, resulting in a broad range of quality-controlled resources and data including standardized methods for creating, processing and interpreting distinct types of high-throughput metagenomic data available to the scientific community.
A framework for human microbiome research

ABSTRACT
A variety of microbial communities and their genes (the microbiome) exist throughout the human body, with fundamental roles in human health and disease. The National Institutes of Health (NIH)-funded Human Microbiome Project Consortium has established a population-scale framework to develop metagenomic protocols, resulting in a broad range of quality-controlled resources and data including standardized methods for creating, processing and interpreting distinct types of high-throughput metagenomic data available to the scientific community. Here we present resources from a population of 242 healthy adults sampled at 15 or 18 body sites up to three times, which have generated 5,177 microbial taxonomic profiles from 16S ribosomal RNA genes and over 3.5 terabases of metagenomic sequence so far. In parallel, approximately 800 reference strains isolated from the human body have been sequenced. Collectively, these data represent the largest resource describing the abundance and variety of the human microbiome, while providing a framework for current and future studies.
ARO Report Number  60119.4-MA
A framework for human microbiome research  ...
A framework for human microbiome research

The Human Microbiome Project Consortium

A variety of microbial communities and their genes (the microbiome) exist throughout the human body, with fundamental roles in human health and disease. The National Institutes of Health (NIH)-funded Human Microbiome Project Consortium has established a population-scale framework to develop metagenomic protocols, resulting in a broad range of quality-controlled resources and data including standardized methods for creating, processing and interpreting distinct types of high-throughput metagenomic data available to the scientific community. Here we present resources from a population of 242 healthy adults sampled at 15 or 18 body sites up to three times, which have generated 5,177 microbial taxonomic profiles from 16S ribosomal RNA genes and over 3.5 terabases of metagenomic sequence so far. In parallel, approximately 800 reference strains isolated from the human body have been sequenced. Collectively, these data represent the largest resource describing the abundance and variety of the human microbiome, while providing a framework for current and future studies.

Advances in sequencing technologies coupled with new bioinformatic developments have allowed the scientific community to begin to investigate the microbes that inhabit our oceans, soils, the human body and elsewhere. Microbes associated with the human body include eukaryotes, archaea, bacteria and viruses, with bacteria alone estimated to outnumber human cells within an individual by an order of magnitude. Our knowledge of these communities and their gene content, referred to collectively as the human microbiome, has until now been limited by a lack of population-scale data detailing their composition and function. The US NIH-funded Human Microbiome Project Consortium (HMP) brought together a broad collection of scientific experts to explore these microbial communities and their relationships with their human hosts. As such, the HMP has focused on producing reference genomes (viral, bacterial and eukaryotic), which provide a critical framework for subsequent metagenomic annotation and analysis, and on generating a baseline of microbial community structure and function from an adult cohort defined by a carefully delineated set of clinical inclusion and exclusion criteria that we term ‘healthy’ in this study (http://commonfund.nih.gov/projects/gap/cgi-bin/GetPdf.cgi?id=phd002854.2). Investigations of the microbiome from this cohort incorporated several complementary analyses including: 16S ribosomal RNA (rRNA) gene sequence (16S) and taxonomic profiles, whole-genome shotgun (WGS) or metagenomic sequencing of whole community DNA, and alignment of the assembled sequences to the reference microbial genomes from the human body. Thus, the HMP complements other large-scale sequence-based human microbiome projects such as the MetaHIT project, which focused on examination of the gut microbiome using WGS data including samples from cohorts exhibiting a wide range of health statuses and physiological characteristics.

Additional projects supported by the HMP are investigating the association of specific components and dynamics of the microbiome with a variety of disease conditions, developing tools and technology including isolating and sequencing uncultured organisms, and studying the ethical, legal and social implications of human microbiome research (http://commonfund.nih.gov/hmp/fundedresearch.aspx). A comprehensive list of current publications from HMP projects is available at http://commonfund.nih.gov/hmp/publications.aspx.

Here we detail the resources created so far by the HMP initiative including: clinical specimens (samples), reference genomes, sequencing and annotation protocols, methods and analyses. We describe the thousands of samples obtained from 15 or 18 distinct body sites from 242 donors over multiple time points that were processed at two clinical centres (Baylor College of Medicine (BCM) and Washington University School of Medicine). We also describe the laboratory and computational protocols developed for reliably generating and interpreting the human microbiome data. HMP resources include both protocols for, and the subsequent data generated from, 16S and metagenomic sequencing of human microbiome samples. During this study, these protocols were rigorously standardized and quality controlled for simultaneous use across four sequencing centres (BCM Human Genome Sequencing Center, The Broad Institute of Massachusetts Institute of Technology (MIT) and Harvard, the J. Craig Venter Institute and The Genome Institute at Washington University School of Medicine). In particular, we focus on the production of the first phase of metagenomic data sets (phase I) used for subsequent in-depth analyses, and we summarize standards and recommendations based on our experiences generating and analysing these data. An additional set of publications (many included in the references and in those of ref. 4) describe in further detail the microbial ecology and microbiological implications of these data. Collectively these resources and analyses represent an important framework for human microbiome research.

HMP resource organization

Supplementary Fig. 1 summarizes organization of the HMP, including the data processing and analytical steps, and the scientific entities gathered to conduct the project. An overview of available HMP data sets and additional resources are provided in Supplementary Tables 1–3. Donors were recruited and enrolled into the HMP through the two clinical centres. Over 240 adults were carefully screened and phenotyped before sampling one to three times at 15 (male) or 18 (female) body sites using a common sampling protocol (http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/GetPdf.cgi?id=phd003190.2). All included subjects were between the ages of 18 and 40 years and had passed a...
Phase I 16S and WGS sequencing overview

A set of 5,298 samples were collected from 242 adults (K. Aagaard et al., manuscript submitted; Table 1 and Supplementary Table 4), from which 16S and WGS data were generated for a total of 5,177 taxonomically characterized communities (16S) and 681 WGS samples describing the microbial communities from habitats within the human airways, skin, oral cavity, gut and vagina. For a subset of 560 samples, both data types were generated (Table 1). These efforts constitute our initial primary metagenomic data sets (phase I) described in more detail later. Additional efforts are ongoing to constitute our initial primary metagenomic data sets (phase I) described in more detail later. Additional efforts are ongoing to

16S standards development and sequencing

The goals of the HMP required that 16S sequences and profiles from data produced at the four participating sequencing centres be comparable in a variety of downstream analyses; however, no suitable methodology was available at the commencement of the project. While establishing 16S protocols, we determined that many components of data production and processing can contribute errors and artefacts. We investigated methods that avoid these errors and their subsequent effects on taxonomic classification and operational

taxonomic unit (OTU)-based community structure. The results are discussed in detail in Supplementary Information and ref. 6. Thus, multiple evaluations of 16S protocols were undertaken before adopting a single standardized protocol that ensured consistency in the high-throughput production.

To maximize accuracy and consistency, protocols were evaluated primarily using a synthetic mock community of 21 known organisms (Supplementary Table 5). Additional testing of the protocol was carried out on a subset of HMP samples (Supplementary Table 1). Collectively, these efforts resulted in adoption of a protocol to amplify and sequence samples using the Roche-454 FLX Titanium platform (http://www.hmpdacc.org/doc/HMP_MDG_454_16S_Protocol.pdf). The HMP created both cell mixtures and genomic DNA extracts of the mock community (Supplementary Tables 2 and 5). A large body of metagenomic data (both 16S and WGS) (RES:HMMC) from these and other calibration experiments are available to the community to facilitate further benchmarking of new molecular and analytical approaches (Supplementary Table 3).

The majority of the sample collection was targeted for 16S sequencing using the 454 FLX Titanium based strategy. The nucleotide sequence of the 16S rRNA gene consists of regions of highly conserved sequence, which alternate with nine regions or windows of variable nucleotide sequence that constitute the most informative portions of the gene sequence for use in taxonomic classification. A window covering number three (V3) to five (V5) variable regions (V35) of the 16S rRNA gene was chosen as the target for 4,879 samples. Sequence of a V1 to V3 (V13) window was also included for a subset of 2,971 samples to provide a complementary view of taxonomic profiles (RES:HMR16S) (Table 1, Supplementary Figs 2, 3 and Supplementary Information).

After adoption of the 16S protocol, including removal of multiple sources of potential artefacts or bias generated by 16S sequencing using pyrosequencing, a variety of approaches for accurate diversity estimation were developed and compared. A 16S data processing pipeline was established using the mothur software package (Supplementary Information), which includes two optional low and high stringency approaches. The former provides an output favouring longer read lengths tailored towards taxonomic classification, the latter an output with more aggressive sequence error reduction tailored towards OTU construction (RES:HMMCP). A third complementary pipeline was also developed using the QIIME software package (Supplementary Information), which processes these data using an

| Table 1 | HMP donor samples examined by 16S and WGS |
|-----------------|------------------|------------------|------------------|------------------|
| Body region     | Body site        | Total samples    | 16S samples      | V13 read depth (M)* | V35 samples      | V35 read depth (M)* | Samples V13 and V35 | Total WGS samples | Total read depth (G)* | Filtered reads (%) | Human reads (%) | Remaining read depth (G)* | Samples 16S and WGS |
| Gut             | Stool            | 352              | 337              | 193              | 1.4                          | 328              | 2.4                          | 184              | 136              | 1,720.7          | 15              | 1                      | 1,450.6            | 124            |
| Oral cavity     | Buccal mucosa    | 346              | 330              | 184              | 1.3                          | 314              | 1.7                          | 168              | 107              | 1,438.0          | 9                | 82                     | 136.7              | 91             |
|                 | Hard palate      | 325              | 325              | 179              | 1.2                          | 310              | 1.7                          | 164              | 1                | 10.9             | 20               | 25                     | 5.9                | 1              |
|                 | Keratinized gingiva | 335           | 329              | 183              | 1.3                          | 319              | 1.9                          | 173              | 6                | 72.3             | 5                | 47                     | 34.4               | 0              |
|                 | Palatine tonsils | 337              | 332              | 189              | 1.2                          | 315              | 1.9                          | 172              | 6                | 74.8             | 2                | 80                     | 13.5               | 1              |
|                 | Saliva           | 315              | 310              | 166              | 0.9                          | 292              | 1.5                          | 148              | 5                | 55.7             | 1                | 91                     | 4.2                | 0              |
|                 | Subgingival plaque | 334            | 328              | 186              | 1.2                          | 314              | 1.8                          | 172              | 7                | 92.1             | 5                | 79                     | 15.3               | 1              |
|                 | Supragingival plaque | 345          | 331              | 192              | 1.3                          | 316              | 1.9                          | 177              | 115              | 1,500.7          | 15               | 40                     | 674.8              | 101             |
|                 | Tongue dorsum    | 331              | 325              | 176              | 1.0                          | 312              | 1.7                          | 163              | 7                | 78.8             | 4                | 79                     | 13.6               | 1              |
| Airway          | Anterior nares   | 316              | 302              | 169              | 1.0                          | 283              | 1.2                          | 150              | 84               | 1,129.9          | 3                | 96                     | 14.3               | 70             |
| Skin            | Left antecubital fossa | 269          | 269              | 158              | 0.7                          | 221              | 0.5                          | 110              | 0                | 98.3             | 8                | 90                     | 25.2               | 43            |
|                 | Left retroauricular crease | 313       | 312              | 188              | 1.6                          | 295              | 1.5                          | 171              | 9                | 126.3            | 9                | 73                     | 21.2               | 8              |
|                 | Right antecubital fossa | 274          | 274              | 158              | 0.7                          | 229              | 0.8                          | 113              | 0                | 99.1             | 0                | 99                     | 42.4               | 12            |
|                 | Right retroauricular crease | 319       | 316              | 190              | 1.4                          | 304              | 1.6                          | 178              | 15               | 181.9            | 18               | 59                     | 42.4               | 12            |
| Vagina          | Mid-vagina       | 145              | 143              | 91               | 0.6                          | 140              | 1.0                          | 88               | 2                | 22.6             | 0                | 99                     | 0.2                | 0              |
|                 | Posterior fornx  | 152              | 142              | 89               | 0.6                          | 136              | 1.0                          | 83               | 53               | 702.1            | 6                | 90                     | 25.2               | 43            |
|                 | Vaginal introitus | 142              | 140              | 87               | 0.6                          | 131              | 0.9                          | 78               | 3                | 36.5             | 1                | 98                     | 0.6                | 1              |

NA, not applicable.

* 1 x 10^8 reads post-processing with the mothur pipeline (Supplementary Information).

† 1 x 10^7 reads (Supplementary Information).

‡ Fraction of reads with low quality bases that were removed (Supplementary Information).

©2012 Macmillan Publishers Limited. All rights reserved
OTU-binning strategy to which taxonomic classification is added (RES:HMQQCP). All pipelines result in highly comparable views of the human microbiome.

**Metagenomic assembly and gene cataloguing**

Approximately 749 samples representing targeted body sites were chosen for WGS sequencing using the Illumina GAIIx platform with 101-base-pair paired-end reads. From a high-quality set of 681 samples an average depth of 13 Gb (≥ 4.3) was achieved per sample, collectively producing a total of 8.8 Tb (RES:HMIWGS) (Table 1). Theoretically, these per sample data are sufficient to cover a 3 Mb bacterial genome present at only 0.8% abundance with a probability of 90% (M. C. Wendel et al., manuscript submitted). In addition, 12 stool samples were simultaneously sequenced using the 454 FLX Titanium platform (RES:HM4WGS). Comparisons between the centres demonstrated high consistency of target sequencing depth and success rates⁶. After development of a protocol for removing reads resulting from human DNA contamination (Supplementary Information), 49% of the reads were targeted for removal as human (for information on authorized access to these reads, see Supplementary Information). Samples collected from soft tissue tended to have higher human contamination (for example, mid-vagina (96%), anterior nares (82%) and throat (75%)). Preparations from saliva were also high in human DNA sequence (80%), whereas stool contained a relatively low abundance of human reads (up to 1%) (Supplementary Fig. 4).

After application of a quality control protocol that includes human sequence removal, quality filtering and trimming of reads (Supplementary Information), the remaining 3.5 Tb from 681 samples were subjected to a three-tiered complementary analysis strategy (Supplementary Information) of reference genome mapping (which was able to use ~57% of the data), assembly and gene prediction (~50% of the data), and metabolic reconstruction (~36% of the data). This combined strategy facilitated the extraction of maximal organismal and functional information.

Metagenomic assemblies were generated for all available samples using an optimized SOAPdenovo protocol with parameters designed to produce substrates for downstream analyses such as gene and function prediction, resulting in a total of 41 million contigs (RES:HMASM) (Supplementary Information). Reads that remained unassembled were pooled across individual body sites and re-assembled using the same approach, resulting in an additional 4,200,672 contigs (RES:HMBSA). These body-site-specific assemblies are aimed at reconstructing organisms that represent too small a fraction in any individual sample to assemble but are found among many individuals. For 12 stool samples both Illumina and 454 FLX Titanium data (RES:HM4WGS) were generated, allowing a hybrid assembly approach using Newbler (Supplementary Information) (RES:HMHASM). Overall, the assembly statistics recovered varied substantially depending on body site and community complexity (Supplementary Fig. 5). However, our results indicate that, for the assembly strategy we used, metagenomic assembly quality plateaus at approximately 6 Gb of microbial sequence coverage for a sample possessing a microbial community structure similar to that of stool samples (Supplementary Fig. 6).

A WGS-based perspective of community membership was obtained by aligning the reads to a set of 1,742 finished bacterial, 131 archaeal, 3,683 viral and 326 microeukaryotic reference genomes¹² (RES:HMREFG) (Supplementary Information) representing a broad taxonomic range from each of these four domains. A total of 57.6% of the high-quality microbial reads could be associated with a known genome (ranging from 33–77% for anterior nares and posterior fornix, respectively) (RES:HMSCP). The overwhelming majority of mapped sequences originated from bacteria (99.7%), while the remaining reads mapped to microeukaryotes (0.3%) or archaea (<0.01%) (Supplementary Information).

Two complementary approaches were used to summarize overall function and metabolism of the human microbiome, producing two primary data sets of annotations (RES:HMMRC and RES:HMG1) (Supplementary Information) and additional secondary analyses (RES:HMGS, HMHGL, HMGC and HMG01) (Supplementary Information) available to the community for further interrogation. The first primary data set of annotations was produced by mapping individual shotgun reads to characterized protein families¹³ (RES:HMMRC). The second was produced from functionally annotated gene predictions generated from the metagenomic assemblies (RES:HMG1), which were subsequently grouped according to high-level biological processes and to selected additional processes specific to metabolism and regulation¹⁴ (RES:HMGS) (Supplementary Tables 6, 7 and Supplementary Fig. 7).

**HMP data generation and analysis lessons**

A key manner in which the HMP resources will serve to guide future studies of the microbiome is by enabling informed decisions regarding sampling protocols and genomic DNA preparation (K. Aagaard et al., manuscript submitted), sequencing depth (M. C. Wendel et al., manuscript submitted), statistical power (P. S. La Rosa et al., manuscript submitted) and metagenomic data type. As indicated in Table 1, the consortium successfully amplified 16S sequences to our target depth at all 18 body sites, with the fewest sequences recovered consistently from the antecubital fossae. The amount of host DNA recovered and the finest level of OTU resolution varied for 16S sequences among body sites⁵ (Supplementary Figs 3 and 4).

From our WGS investigations, a series of protocols (http://hmpdacc.org/tools_protocols/tools_protocols.php) have been established to process large volumes of short-read WGS data and to annotate and examine these data through both a multi-tiered assembly approach and as single reads¹⁵. An investigator’s choice of metagenomic technologies can thus be guided not only by a 16S versus WGS dichotomy, but also by the expected fraction of host sequence and the appropriate 16S region targeting the dominant taxa at each body site (Supplementary Figs 2–6 and 8).

Together, these data sets represent comprehensive and complementary views of the human microbiome, as shown by comparing organismal (Fig. 1a) and gene (Fig. 1b) catalogues, and the ratio of genes contributed per OTU (Fig. 1c). The discovery rate of new gene clusters (as determined by annotation of assembled WGS data) is in general detected more slowly relative to organismal discovery (as determined by OTU data) owing to the fragmentary nature of these community reads and assemblies despite high sequence depth (Fig. 1a, b and Supplementary Fig. 9), and the number of genes contributed per OTU varies by body site (Fig. 1c and Supplementary Information). However, in general, these results highlight an important point for consideration of further microbiome investigations using these data sets, as they suggest that the majority of the common taxa and genes present in this reference population have been detected.

We additionally compared the gut community gene catalogue sampled by the HMP with that of MetaHIT in terms of total detected gene counts. The HMP recovered more total non-redundant gene counts (5,140,472) than reported by MetaHIT (3,299,822)⁶, probably reflecting a combination of the increased sequence depth obtained by the HMP (11.7 Gb HMP, 4.5 Gb MetaHIT on average) and differences in data generation and processing⁷.

The two non-redundant sets of gene sequences were subsequently combined and compared by matches to a database of orthologous groups¹⁶ of functionally annotated genes. Approximately 57% of the orthologous groups recovered by this method overlapped between the data sets, while an additional 34% versus 10% were unique to the HMP and MetaHIT, respectively (Supplementary Fig. 10, Supplementary Table 8 and Supplementary Information). After removal of genes that received any orthologous group assignment, the remaining novel
Human microbiome reference genomes

The current goal for the reference genome component of the HMP is to sequence at least 3,000 reference bacterial genomes, and additional viral and microeukaryotic genomes, associated with the human body. Thus far, more than 800 genomes have been sequenced and are available from the NCBI and the DACC (http://hmimdacc.org/HMRGD). From an alignment of WGS reads to reference genomes (RES:HMRERFG), approximately 26% from the total read set (46% of all reads that could be aligned) were matched to a subset of 223 HMP reference genomes (Supplementary Information and Supplementary Data).

We continue to solicit community feedback for strains that will best benefit our attempts at understanding the breadth of human microbiome diversity. For example, a prioritized list of the 'most wanted' HMP taxa is being maintained (http://hmimdacc.org/most_wanted/) with the goal of targeting these difficult to obtain organisms using both culture-based and single-cell approaches.

A catalogue of all HMP reference genomes along with custom filtering, viewing, graphing and download options can be found at the DACC Project Catalogue (http://www.hmimdacc-resources.org/hmp_catalog/main.cgi). In addition, comparative analyses of reference genomes are provided by the data warehouse and analytical systems, Integrated Microbial Genomes/HMP (http://www.hmimdacc-resources.org/cgi-bin/immg_hmp/main.cgi). Cultures of all HMP reference strains are required to be made publicly available through the Biodefense and Emerging Infections Research Resources Repository (BEI). Information on strain acquisition can be found at the DACC (http://hmimdacc.org/reference_genomes/reference_genomes.php) and BEI (http://www.beisources.org/tabid/1901/stabid/1901/CollectionLinkID/4/Default.aspx).

Conclusion

An overarching goal of this multi-year, multi-centre project is the generation of a community resource to advance research efforts related to the microbiome. The result is a collection of 11,174 primary biological specimens representing the human microbiome, as well as corresponding blood samples from the human donors, which are being reserved for sequencing at a future date and from which cell lines will be developed. A variety of new protocols were developed to enable a project of this scope; these include methods for donor recruitment, laboratory and sequence processing, and analysis of...
Received 2 November 2011; accepted 10 May 2012.


Received 2 November 2011; accepted 10 May 2012.

Texas 77030, USA.  
University of Pennsylvania, Center for Bioethics and Department of Medical Ethics, 3401 Market Street, Suite 320, Philadelphia, Pennsylvania 19104, USA.  
University of Michigan, Department of Microbiology & Immunology, 5713 Medical Science Bldg. II, 1150 West Medical Center Dr., Ann Arbor, Michigan 48109-5620, USA.  
Michigan State University, Department of Microbiology and Molecular Genetics, 6180 Biomedical Physical Sciences, Michigan State University, East Lansing, Michigan 48824, USA.  
The EMMES Corporation, 401 N. Washington St., Suite 700, Rockville, Maryland 20850, USA.  
Wayne State University School of Medicine, Detroit, Michigan, Harper University Hospital, 3990 John R Street, Detroit, Michigan 48201, USA.  
Johns Hopkins University School of Medicine, McKusick-Nathans Institute of Genetic Medicine, Bloomberg School of Public Health, E3138, 615 N Wolfe St, Baltimore, Maryland 21205, USA.  
J. Craig Venter Institute, 10355 Science Center Drive, San Diego, California 92121, USA.  
Northwestern University, Feinberg School of Medicine, 420 East Superior Street Chicago, Illinois 60611, USA.  
Alkek Center for Metagenomics and Microbiome Research, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA.  
National Institutes of Health, National Human Genome Research Institute (NHGRI), Genetics and Molecular Biology Branch, MSC 4442, Bethesda, Maryland 20892, USA.  
Deceased.