6-14-2012

A framework for human microbiome research

Barbara A. Methe

Jacques Izard

ARTICLE

doi:10.1038/nature11209

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[1]. 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[2] 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://www.ncbi.nlm.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[3,4]. Thus, the HMP complements other large-scale sequence-based human microbiome projects such as the MetaHIT project[5], 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

*Lists of participants and their affiliations appear at the end of the paper.

screening for systemic health based on oral, cutaneous and body mass exclusion criteria (http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/GetPdf.cgi?id=phd002854.2) (K. Aagaard *et al.*, manuscript submitted).

A Data Analysis and Coordination Center (DACC) was created to serve as the central repository for all HMP WGS, 16S and reference genome sequence information generated by the four sequencing centres. The DACC supports access to analysis software, biological samples, clinical protocols, news, publication announcements and project statistics, and performed centralized analysis of HMP reference genome and WGS annotation in cooperation with the sequencing centres. All unprocessed 16S, WGS and reference genome sequence data are deposited at the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/bioproject/43021). Unless otherwise noted, all data sets and protocols described here are available to the scientific community at the DACC (http://hmpdacc.org). Specific data sets referred to in this work and available at the DACC are indicated in parentheses with the preface 'RES'.

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 sequence and analyse the remaining samples from the complete HMP collection (11,174 primary specimens in total from 300 individuals sampled up to three times over 22 months) (K. Aagaard *et al.*, manuscript submitted).

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[6] (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[6] (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[6]. 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[6] (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[7,8], a variety of approaches for accurate diversity estimation were developed and compared[9]. A 16S data processing pipeline was established using the mothur software package[10] (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[11] (Supplementary Information), which processes these data using an

Table 1 | HMP donor samples examined by 16S and WGS

Body region	Body site	Total samples	Total 16S samples	V13 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	1.7	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
	Throat	331	325	176	1.0	312	1.7	163	7	78.8	4	79	13.6	1
	Tongue dorsum	348	332	193	1.3	320	2.0	181	122	1,620.1	15	19	1,084.3	106
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	NA	NA	NA	0	NA
	Left retroauricular crease	313	312	188	1.6	295	1.5	171	9	126.3	9	73	22.1	8
	Right antecubital fossa	274	274	158	0.7	229	0.5	113	0	NA	NA	NA	0	NA
	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 fornix	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
	Total	5,298	5,177	2,971	19	4,879	26.3	2,673	681	8,863.3	11	49	3,538.1	560

NA, not applicable.
* 1×10^6 reads post-processing with the mothur pipeline (Supplementary Information).
† 1×10^9 reads (Supplementary Information).
‡ Fraction of reads with low quality bases that were removed (Supplementary Information).
§ Fraction of human reads that were removed (Supplementary Information).

OTU-binning strategy to which taxonomic classification is added (RES:HMQCP). 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. Wendl *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 rates[4]. 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 genomes[12] (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:HMGI) (Supplementary Information) and additional secondary analyses (RES:HMGS, HMHGI, HMGC and HMGOI) (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 families[13] (RES:HMMRC). The second was produced from functionally annotated gene predictions generated from the metagenomic assemblies (RES:HMGI), which were subsequently grouped according to high-level biological processes and to selected additional processes specific to metabolism and regulation[14] (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. Wendl *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 human DNA recovered and the finest level of OTU resolution varied for 16S sequences among body sites[6] (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 reads[15]. 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)[5], 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 processing[5].

The two non-redundant sets of gene sequences were subsequently combined and compared by matches to a database of orthologous groups[16] 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

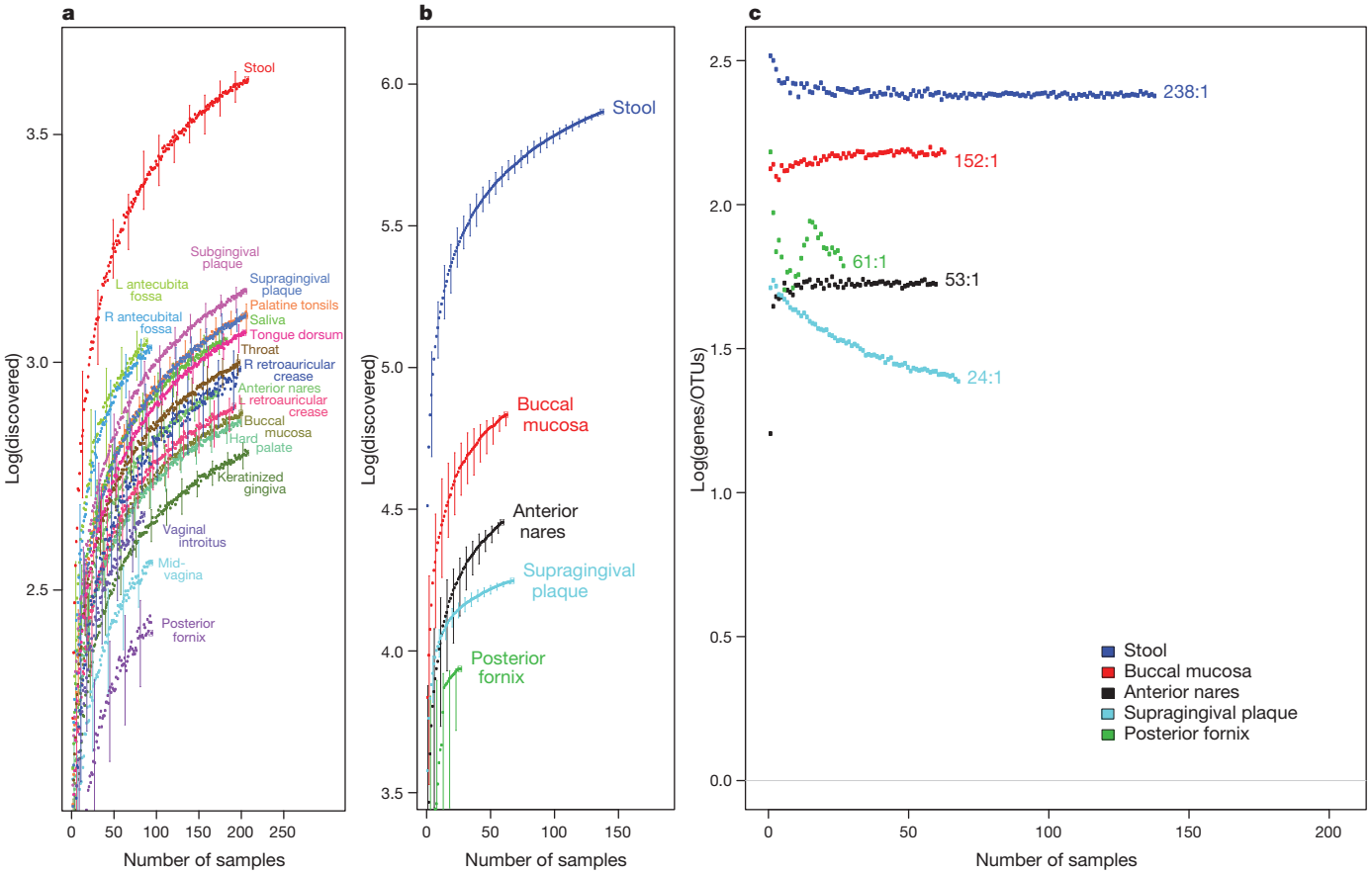

Figure 1 | Rates of gene and OTU discovery from HMP taxonomic and metagenomic data. a–c, Accumulation curves for OTU counts from 16S data (all body sites) (**a**), clustered gene index counts from metagenomic data (all applicable body sites) (**b**) and the ratio of average unique genes contributed versus unique OTUs encountered with increasing sample counts (**c**) (Supplementary Information). L, left; R, right. Ratios given for each curve in **c** represent the average number of unique genes contributed per unique OTU at the final sample count. Curves for stool, buccal mucosa and anterior nares suggest that the proportion of gene-to-taxa discovery has stabilized. In contrast, the curve for supragingival plaque suggests that relatively fewer new genes are being contributed per additional OTU. Error bars represent 95% confidence intervals.

genes were subsequently clustered[17]. Approximately 79% of the HMP-derived novel gene clusters were orthologous to one or more clusters in MetaHIT, while an additional 16% were unique to this study versus 5% for MetaHIT-derived data[5] (Supplementary Fig. 11, Supplementary Table 8 and Supplementary Information). These results suggest that, for this body habitat, relatively similar gene catalogues were recovered despite differences in experimental design and protocols. However, a greater proportion of both annotated and unique novel genes were detected in the HMP data set, emphasizing the utility of sequencing depth in recovering gene function and, in particular, deriving rare function. These results further underscore the importance of large-scale sequence-based studies of the microbiome to characterize better its gene content and diversity.

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://hmpdacc.org/HMRGD). From an alignment of WGS reads to reference genomes (RES:HMREFG), 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://hmpdacc.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.hmpdacc-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.hmpdacc-resources.org/cgi-bin/imgm_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://hmpdacc.org/reference_genomes/reference_genomes.php) and BEI (http://www.beiresources.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

16S and WGS sequence and profiles. These resources serve as models to guide the design of similar projects. Studies with a primary focus on disease can use this reference for comparative purposes, including detecting shifts in microbial taxonomic and functional profiles, or identification of new species not present in healthy cohorts that appear under disease conditions. The catalogue described in this study is, to our knowledge, the largest and most comprehensive reference set of human microbiome data associated with healthy adult individuals. Collectively the data represent a treasure trove that can be mined to identify new organisms, gene functions, and metabolic and regulatory networks, as well as correlations between microbial community structure and health and disease[4]. Among other future benefits, this resource may promote the development of novel prophylactic strategies such as the application of prebiotics and probiotics to foster human health.

METHODS SUMMARY

As part of a multi-institutional collaboration, the HMP human subjects study was reviewed by the Institutional Review Boards (IRBs) at each sampling site: the BCM (IRB protocols H-22895 (IRB no. 00001021) and H-22035 (IRB no. 00002649)); Washington University School of Medicine (IRB protocol HMP-07-001 (IRB no. 201105198)); and St Louis University (IRB no. 15778). The study was also reviewed by the J. Craig Venter Institute under IRB protocol 2008-084 (IRB no. 00003721), and at the Broad Institute of MIT and Harvard the study was determined to be exempt from IRB review. All study participants gave their written informed consent before sampling and the study was conducted using the Human Microbiome Project Core Sampling Protocol A. Each IRB has a federal-wide assurance and follows the regulations established in 45 CFR Part 46. The study was conducted in accordance with the ethical principles expressed in the Declaration of Helsinki and the requirements of applicable federal regulations.

All further details are in Supplementary Information.

Received 2 November 2011; accepted 10 May 2012.

1. Gilbert, J. A. & Dupont, C. L. Microbial metagenomics: beyond the genome. *Annu. Rev. Mar. Sci.* **3,** 347–371 (2011).
2. NIH HMP Working Group et al. The NIH Human Microbiome Project. *Genome Res.* **19,** 2317–2323 (2009).
3. Human Microbiome Jumpstart Reference Strains Consortium. A catalog of reference genomes from the human microbiome. *Science* **328,** 994–999 (2010).
4. The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* http://dx.doi.org/10.1038/nature11234 (this issue).
5. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464,** 59–65 (2010).
6. Jumpstart Consortium Human Microbiome Project Data Generation Working Group. Evaluation of 16S rDNA-based community profiling for human microbiome research. *PLoS ONE* http:dx.plos.org/10.1371/journal.pone.0039315 (14 June 2012).
7. Kunin, V., Engelbrektson, A., Ochman. H. & Hugenholtz, P. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ. Microbiol.* **12,** 118–123 (2010).
8. Huse, S. M., Huber, J. A., Morrison, H. G., Sogin, M. L. & Welch, D. M. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.* **8,** R143 (2007).
9. Schloss, P. D., Gevers, D. & Westcott, S. L. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS ONE* **6,** e27310 (2011).
10. Schloss, P. D. *et al.* Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **75,** 7537–7541 (2009).
11. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7,** 335–336 (2010).
12. Martin, J. S. *et al.* Optimizing read mapping to reference genomes to determine composition and species prevalence in microbial communities. *PLoS ONE* http://dx.doi.org/10.1371/journal.pone.0036427 (14 June 2012).
13. Abubucker, S. *et al.* Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput. Biol.* http://dx.doi.org/10.1371/journal.pcbi.1002358 (14 June 2012).
14. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nature Genet.* **25,** 25–29 (2000).
15. Goll, J. *et al.* A case study for large-scale human microbiome analysis using JCVI's Metagenomics Reports (METAREP). *PLoS ONE* http://dx.doi.org/10.1371/journal.pone.002904 (14 June 2012).
16. Muller, J. *et al.* eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Res.* **38,** D190–D195 (2010).
17. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26,** 2460–2461 (2010).

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

Acknowledgements The consortium would like to thank our external scientific advisory board: R. Blumberg, J. Davies, R. Holt, P. Ossorio, F. Ouellette, G. Schoolnik and A. Williamson. We would also like to thank our collaborators throughout the International Human Microbiome Consortium, particularly the investigators of the MetaHIT project, for advancing human microbiome research. Data repository management was provided by the NCBI and the Intramural Research Program of the NIH National Library of Medicine. We especially appreciate the generous participation of the individuals from the St Louis, Missouri, and Houston, Texas areas who made this study possible. This research was supported in part by NIH grants U54HG004969 to B.W.B.; U54HG003273 to R.A.G.; U54HG004973 to R.A.G., S.K.H. and J.F.P.; U54HG003067 to E. S. Lander.; U54AI084844 to K.E.N.; N01AI30071 to R. L. Strausberg; U54HG004968 to G.M.W.; U01HG004866 to O.W.; U54HG003079 to R.K.W.; R01HG005969 to C.H.; R01HG004872 to R.K.; R01HG004885 to M.P.; R01HG005975 to P.D.S.; R01HG004908 to Y.Y.; R01HG004900 to M. K. Cho and P. Sankar; R01HG005171 to D.E.H.; R01HG004853 to A.L.M.; R01HG004856 to R.R.; R01HG004877 to R.R.S. and R.M.F.; R01HG005172 to P. Spicer; R01HG004857 to M.P.; R01HG004906 to T.M.S.; R21HG005811 to E.A.-V.; G.A.B. was supported by UH2AI083263 and UH3AI083263 (G.A.B., C. N. Cornelissen, L. K. Eaves and J. F. Strauss); M.J.B. was supported by UH2AR057506, S.M.H. was supported by UH3DK083993 (V. B. Young, E. B. Chang, F. Meyer, T.M.S., M. L. Sogin, J. M. Tiedje); K.P.R. was supported by UH2DK083990 (J.V.); J.A.S. and H.H.K. were supported by UH2AR057504 and UH3AR057504 (J.A.S.); DP2OD001500 to K.M.A.; N01HG62088 to the Coriell Institute for Medical Research; U01DE016937 to F.E.D.; S.K.-H. was supported by RC1DE020298and R01DE021574 (S.K.-H. and H. Li); J.I. was supported by R21CA139193 (J.I. and D. S. Michaud); K.P.L. was supported by P30DE020751 (D. J. Smith); Army Research Office grant W911NF-11-1-0473 to C.H.; National Science Foundation grants NSF DBI-1053486 to C.H. and NSF IIS-0812111 to M.P.; The Office of Science of the US Department of Energy under contract no. DE-AC02-05CH11231 for P.S.C.; LANL Laboratory-Directed Research and Development grant 20100034DR and the US Defense Threat Reduction Agency grants B104153I and B084531I to P.S.C.; Research Foundation - Flanders (FWO) grant to K.F. and J. Raes; R.K. is a Howard Hughes Medical Institute (HHMI) Early Career Scientist; Gordon & Betty Moore Foundation funding and institutional funding from the J. David Gladstone Institutes to K.S.P.; A.M.S. was supported by fellowships provided by the Rackham Graduate School and the NIH Molecular Mechanisms in Microbial Pathogenesis Training Grant T32AI007528; a Crohn's and Colitis Foundation of Canada Grant in Aid of Research to E.A.-V.; 2010 IBM Faculty Award to K.C.W. Analysis of the HMP data was performed using National Energy Research Scientific Computing resources; the BluBioU Computational Resource at Rice University.

Author Contributions Principal investigators: B.W.B., R.A.G., S.K.H., B.A.M., K.E.N., J.F.P., G.M.W., O.W., R.K.W. Manuscript preparation: B.A.M., K.E.N., M.P., H.H.C., M.G.G., D.G., C.H., J.F.P. Funding agency management: C.C.B., T.B., V.R.B., J.L.C., S.C., C.D., V.D.F., C.G., M.Y.G., R.D.L., J.M., P.M., J.P., L.M.P., J.A.S., L.W., C.W., K.A.W. Project leadership: S.A., J.H.B., B.W.B., A.T.C., H.H.C., A.M.E., M.G.F., R.S.F., D.G., M.G.G., K.H., S.K.H., C.H., E.A.L., R.M., V.M., J.C.M., B.A.M., M.M., D.M.M., K.E.N., J.F.P., E.J.S., J.V., G.M.W., O.W., A.M.W., K.C.W., J.R.W., S.K.Y., Q.Z. Analysis preparation for manuscript: M.B., B.L.C., D.G., M.G.G., M.E.H., C.H., K.L., B.A.M., X.Q., J.R.W., M.T. Data release: L.A., T.B., I.A.C., K.C., H.H.C., N.J.D., D.J.D., A.M.E., V.M.F., L.F., J.M.G., S.G., S.K.H., M.E.H., C.J., V.J., C.K., A.A.M., V.M.M., T.M., M.M., D.M.M., J.O., K.P., J.F.P., C.P., X.Q., R.K.S., N.S., I.S., E.J.S., D.V.W., O.W., K.W., K.C.W., C.Y., B.P.Y., Q.Z. Methods and research development: S.A., H.M.A., M.B., D.M.C., A.M.E., R.L.E., M.F., S.F., M.G.F., D.C.F., D.G., G.G., B.J.H., S.K.H., M.E.H., W.A.K., N.L., K.L., V.M., E.R.M., B.A.M., M.M., D.M.M., C.N., J.F.P., M.E.P., X.Q., M.C.R., C.R., E.J.S., S.M.S., D.G.T., D.V.W., G.M.W., Y.W., K.M.W., S.Y., B.P.Y., S.K.Y., Q.Z. DNA sequence production: S.A., E.A., T.A., T.B., C.J.B., D.A.B., K.D.D., S.P.D., A.M.E., R.L.E., C.N.F., S.F., C.C.F., L.L.F., R.S.F., B.H., S.K.H., M.E.H., V.J., C.L.K., S.L.L., N.L., L.L., D.M.M., I.N., C.N., M.O., J.F.P., X.Q., J.G.R., Y.R., M.C.R., D.V.W., Y.W., B.P.Y., Y.Z. Clinical sample collection: K.M.A., M.A.C., W.M.D., L.L.F., N.G., H.A.H., E.L.H., J.A.K., W.A.K., T.M., A.L.M., P.M., S.M.P., J.F.P., G.A.S., J.V., M.A.W., G.M.W. Body site experts: K.M.A., E.A.V., G.A., L.B., M.J.B., C.C.D., F.E.D., L.F., J.I., J.A.K., S.K.H., H.H.K., K.P.L., P.J.M., J. Ravel., T.M.S., J.A.S., J.D.S., J.V. Ethical, legal and social implications: R.M.F., D.E.H., W.A.K., N.B.K., C.M.L., A.L.M., R.R., P. Sankar, P. Spicer, R.R.S., L.Z. Strain management: E.A.V., J.H.B., I.A.C., K.C., S.W.C., H.H.C., T.Z.D., A.S.D., A.M.E., M.G.F., M.G.G., S.K.H., V.J., N.C.K., S.L.L., L.L., K.L., E.A.L., V.M.M., B.A.M., D.M.M., K.E.N., I.N., I.P., L.S., E.J.S., C.M.T., M.T., D.V.W., G.M.W., A.M.W., Y.W., K.M.W., B.P.Y., L.Z., Y.Z. 16S data analysis: K.M.A., E.J.A., G.L.A., C.A.A., M.B., B.W.B., J.P.B., G.A.B., S.R.C., S.C., J.C., T.Z.D., F.E.D., E.D., A.M.E., R.C.E., M.F., A.A.F., J.F., K.F., H.G., D.G., B.J.H., T.A.H., S.M.H., C.H., J.I., J.K.J., S.T.K., S.K.H., R.K., H.H.K., O.K., P.S.L., R.E.L., K.L., C.A.L., D.M., B.A.M., K.A.M., M.M., M.P., J.F.P., M.P., K.S.P., X.Q., J. Raes, K.P.R., M.C.R., B.R., J.F.S., P.D.S., T.M.S., N.S., J.A.S., W.D.S., T.J.S., C.S.S., E.J.S., R.M.T., J.V., T.A.V., Z.W., D.V.W., G.M.W., J.R.W., K.M.W., Y.Y., S.Y., Y.Z. Shotgun data processing and alignments: C.J.B., J.C.C., E.D., D.G., A.G., M.E.H., H.J., D.K., K.C.K., C.L.K., Y.L., J.C.M., B.A.M., M.M., D.M.M., J.O., J.F.P., X.Q., J.G.R., R.K.S., N.U.S., I.S., E.J.S., G.G.S., S.M.S., J.W., Z.W., G.M.W., O.W., K.C.W., T.W., S.K.Y., L.Z. Assembly: H.M.A., C.J.B., P.S.C., L.C., Y.D., S.P.D., M.G.F., M.E.H., H.J., S.K., B.L., Y.L., C.L., J.C.M., J.M.M., J.R.M., P.J.M., M.M., J.F.P., M.P., M.E.P., X.Q., M.R., R.K.S., M.S., D.D.S., G.G.S., S.M.S., C.M.T., T.J.T., W.W., G.M.W., K.C.W., L.Y., Y.Y., S.K.Y., L.Z. Annotation: O.O.A., V.B., C.J.B., I.A.C., A.T.C., K.C., H.H.C., A.S.D., M.G.G., J.M.G., J.G., A.G., S.G., B.J.H., K.H., S.K.H., C.H., H.J., N.C.K., R.M., V.M.M., K.M., T.M., M.M., J.O., K.P., M.P., X.Q., N.S., E.J.S., G.G.S., S.M.S., M.T., G.M.W., K.C.W., J.R.W., C.Y., S.K.Y., Q.Z., L.Z. WGS Metabolic Reconstruction: S.A., B.L.C., J.G., C.H., J.I., B.A.M., M.M., B.R., A.M.S., N.S., M.T., G.M.W., S.Y., Q.Z., J.D.Z.

Author Information Accession numbers for all primary sequencing data are given in Supplementary Information. Reprints and permissions information is available at www.nature.com/reprints. This paper is distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share-Alike licence, and is freely available to all readers at www.nature.com/nature. The authors declare no competing financial

interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to B.A.M. (bmethe@jcvi.org).

The Human Microbiome Project Consortium

Barbara A. Methé[1], Karen E. Nelson[1], Mihai Pop[2], Heather H. Creasy[3], Michelle G. Giglio[3], Curtis Huttenhower[4,5], Dirk Gevers[5], Joseph F. Petrosino[6,15,79], Sahar Abubucker[7], Jonathan H. Badger[77], Asif T. Chinwalla[7], Ashlee M. Earl[5], Michael G. FitzGerald[5], Robert S. Fulton[7], Kymberlie Hallsworth-Pepin[7], Elizabeth A. Lobos[7], Ramana Madupu[1], Vincent Magrini[7], John C. Martin[7], Makedonka Mitreva[7], Donna M. Muzny[6], Erica J. Sodergren[7], James Versalovic[8,9], Aye M. Wollam[7], Kim C. Worley[6], Jennifer R. Wortman[5], Sarah K. Young[5], Qiandong Zeng[5], Kjersti M. Aagaard[10], Olukemi O. Abolude[3], Emma Allen-Vercoe[11], Eric J. Alm[5,12], Lucia Alvarado[5], Gary L. Andersen[13], Scott Anderson[5], Elizabeth Appelbaum[7], Harindra M. Arachchi[5], Gary Armitage[14], Cesar A. Arze[3], Tulin Ayvaz[15], Carl C. Baker[16], Lisa Begg[17], Tsegahiwot Belachew[18], Veena Bhonagiri[7], Monika Bihan[1], Martin J. Blaser[19], Toby Bloom[5], Vivien R. Bonazzi[20], Paul Brooks[21,22], Gregory A. Buck[22,23], Christian J. Buhay[6], Dana A. Busam[1], Joseph L. Campbell[18,20], Shane R. Canon[24], Brandi L. Cantarel[3], Patrick S. Chain[25,26], I-Min A. Chen[27], Lei Chen[7], Shaila Chhibba[20], Ken Chu[27], Dawn M. Ciulla[5], Jose C. Clemente[28], Sandra W. Clifton[7], Sean Conlan[80], Jonathan Crabtree[3], Mary A. Cutting[29], Noam J. Davidovics[3], Catherine C. Davis[30], Todd Z. DeSantis[31], Carolyn Deal[18], Kimberley D. Delehaunty[7], Floyd E. Dewhirst[32,33], Elena Deych[7], Yan Ding[6], David J. Dooling[7], Shannon P. Dugan[6], W. Michael Dunne Jr[34,35], A. Scott Durkin[1], Robert C. Edgar[36], Rachel L. Erlich[5], Candace N. Farmer[7], Ruth M. Farrell[37], Karoline Faust[38,39], Michael Feldgarden[5], Victor M. Felix[3], Sheila Fisher[5], Anthony A. Fodor[40], Larry Forney[41], Leslie Foster[1], Valentina Di Francesco[18], Jonathan Friedman[42], Dennis C. Friedrich[5], Catrina C. Fronick[7], Lucinda L. Fulton[7], Hongyu Gao[8], Nathalia Garcia[43], Georgia Giannoukos[5], Christina Giblin[18], Maria Y. Giovanni[18], Jonathan M. Goldberg[5], Johannes Goll[1], Antonio Gonzalez[44], Allison Griggs[5], Sharvari Gujja[5], Brian J. Haas[5], Holli A. Hamilton[29], Emily L. Harris[29], Theresa A. Hepburn[5], Brandi Herter[7], Diane E. Hoffmann[45], Michael E. Holder[6], Clinton Howarth[5], Katherine H. Huang[5], Susan M. Huse[46], Jacques Izard[32,47], Janet K. Jansson[48], Huaiyang Jiang[6], Catherine Jordan[3], Vandita Joshi[6], James A. Katancik[49], Wendy A. Keitel[15], Scott T. Kelley[50], Cristyn Kells[5], Susan Kinder-Haake[51]†, Nicholas B. King[52], Rob Knight[28,53], Dan Knights[44], Heidi H. Kong[54], Omry Koren[55], Sergey Koren[2], Karthik C. Kota[7], Christie L. Kovar[6], Nikos C. Kyrpides[26], Patricio S. La Rosa[56], Sandra L. Lee[6], Katherine P. Lemon[32,57], Niall Lennon[6], Cecil M. Lewis[58], Lora Lewis[6], Ruth E. Ley[55], Kelvin Li[1], Konstantinos Liolios[26], Bo Liu[2], Yue Liu[6], Chien-Chi Lo[25], Catherine A. Lozupone[28], R. Dwayne Lunsford[29], Tessa Madden[59], Anup A. Mahurkar[3], Peter J. Mannon[60], Elaine R. Mardis[7], Victor M. Markowitz[26,27], Konstantinos Mavrommatis[26], Jamison M. McCorrison[1], Daniel McDonald[28], Jean McEwen[20], Amy L. McGuire[61], Pamela McInnes[29], Teena Mehta[5], Kathie A. Mihindukulasuriya[7], Jason R. Miller[1], Patrick J. Minx[7], Irene Newsham[6], Chad Nusbaum[5], Michelle O'Laughlin[7], Joshua Orvis[3], Ioanna Pagani[26], Krishna Palaniappan[27], Shital M. Patel[62], Matthew Pearson[5], Jane Peterson[20], Mircea Podar[63], Craig Pohl[7], Katherine S. Pollard[64,65,66], Margaret E. Priest[5], Lita M. Proctor[20], Xiang Qin[6], Jeroen Raes[38,39], Jacques Ravel[3,67], Jeffrey G. Reid[6], Mina Rho[68], Rosamond Rhodes[69], Kevin P. Riehle[70], Maria C. Rivera[22,23], Beltran Rodriguez-Mueller[50], Yu-Hui Rogers[1], Matthew C. Ross[15], Carsten Russ[5], Ravi K. Sanka[1], Pamela Sankar[71], J. Fah Sathirapongsasuti[4], Jeffery A. Schloss[20], Patrick D. Schloss[72], Thomas M. Schmidt[73], Matthew Scholz[25], Lynn Schriml[3], Alyxandria M. Schubert[72], Nicola Segata[4], Julia A. Segre[80], William D. Shannon[56], Richard R. Sharp[37], Thomas J. Sharpton[64], Narmada Shenoy[5], Nihar U. Sheth[22], Gina A. Simone[74], Indresh Singh[1], Chris S. Smillie[42], Jack D. Sobel[75], Daniel D. Sommer[2], Paul Spicer[58], Granger G. Sutton[1], Sean M. Sykes[5], Diana G. Tabbaa[5], Mathangi Thiagarajan[1], Chad M. Tomlinson[7], Manolito Torralba[1], Todd J. Treangen[76], Rebecca M. Truty[64], Tatiana A. Vishnivetskaya[63], Jason Walker[7], Lu Wang[20], Zhengyuan Wang[5], Doyle V. Ward[5], Wesley Warren[7], Mark A. Watson[34], Christopher Wellington[20], Kris A. Wetterstrand[20], James R. White[3], Katarzyna Wilczek-Boney[6], Yuan Qing Wu[6], Kristine M. Wylie[7], Todd Wylie[7], Chandri Yandava[5], Liang Ye[7], Yuzhen Ye[68], Shibu Yooseph[77], Bonnie P. Youmans[15], Lan Zhang[6], Yanjiao Zhou[7], Yiming Zhu[6], Laurie Zoloth[78], Jeremy D. Zucker[5], Bruce W. Birren[5], Richard A. Gibbs[6], Sarah K. Highlander[6,15], George M. Weinstock[7], Richard K. Wilson[7] & Owen White[3]

[1]J. Craig Venter Institute, 9704 Medical Center Drive, Rockville, Maryland 20850, USA. [2]University of Maryland, Center for Bioinformatics and Computational Biology and Department of Computer Science, Biomolecular Sciences Building Rm. 3120F, College Park, Maryland 20742, USA. [3]University of Maryland School of Medicine, Institute for Genome Sciences 801 W. Baltimore Street, Baltimore, Maryland 21201, USA. [4]Harvard School of Public Health, Department of Biostatistics, 655 Huntington Avenue, Boston, Massachusetts 02115, USA. [5]The Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. [6]Baylor College of Medicine Human Genome Sequencing Center, One Baylor Plaza, Houston, Texas 77030, USA. [7]Washington University School of Medicine, The Genome Institute, 4444 Forest Park Avenue, St Louis, Missouri 63108, USA. [8]Baylor College of Medicine, Department of Pathology & Immunology, One Baylor Plaza, Houston, Texas 77030, USA. [9]Texas Children's Hospital Department of Pathology, 6621 Fannin Street, Houston, Texas 77030, USA. [10]Baylor College of Medicine, Department of Obstetrics & Gynecology, Division of Maternal-Fetal Medicine, One Baylor Plaza, Houston, Texas 77030, USA. [11]University of Guelph Department of Molecular and Cellular Biology, 50 Stone Road East, Guleph, Ontario N1G 2W1, Canada. [12]Massachusetts Institute of Technology, Department of Civil & Environmental Engineering, Parsons Laboratory, Room 48-317, 15 Vassar Street, Cambridge, Massachusetts 02139, USA. [13]Lawrence Berkeley National Laboratory, Center for Environmental Biotechnology, 1 Cyclotron Road, Berkeley, California 94720, USA. [14]University of California, San Francisco, School of Dentistry, 707 Parnassus Avenue, San Francisco, California 94143, USA. [15]Baylor College of Medicine, Molecular Virology and Microbiology, One Baylor Plaza, Houston, Texas 77030, USA. [16]National Institutes of Health, National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS), 6701 Democracy Boulevard, MSC 4872, Bethesda, Maryland 20892, USA. [17]National Institutes of Health, Office of Research on Women's Health (ORWH), 6707 Democracy Boulevard, MSC 5484, Bethesda, Maryland 20892, USA. [18]National Institutes of Health, National Institute for Allergy and Infectious Diseases (NIAID), 6610 Rockledge Drive, MSC 6603, Bethesda, Maryland 20892, USA. [19]New York University Langone Medical Center, Department of Medicine, 550 First Avenue, OBV A-606, New York, New York 10016, USA. [20]National Institutes of Health, National Human Genome Research Institute (NHGRI), 5635 Fishers Lane, MSC 9305, Bethesda, Maryland 20892, USA. [21]Virginia Commonwealth University, Department of Statistical Sciences and Operations Research, PO Box 843083, Richmond, Virginia 23284, USA. [22]Virginia Commonwealth University, Center for the Study of Biological Complexity, 1000 West Cary Street, Richmond, Virginia 23284, USA. [23]Virginia Commonwealth University, Department of Biology, 1000 West Cary Street, Richmond, Virginia 23284, USA. [24]Lawrence Berkeley National Laboratory, Technology Integration Group, National Energy Research Scientific Computing Center, 1 Cyclotron Road, Berkeley, California 94720, USA. [25]Los Alamos National Laboratory Genome Science Group, Bioscience Division, HRL, MS-888, LANL, Los Alamos, New Mexico 87545, USA. [26]Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, California 94598, USA. [27]Lawrence Berkeley National Laboratory, Biological Data Management and Technology Center, Computational Research Division, 1 Cyclotron Road, Berkeley, California 94720, USA. [28]University of Colorado, Department of Chemistry and Biochemistry, Campus Box 215, University of Colorado, Boulder, Colorado 80309-0215, USA. [29]National Institutes of Health, National Institute of Dental and Craniofacial Research (NIDCR), 6701 Democracy Boulavard, MSC 4878, Bethesda, Maryland 20892, USA. [30]The Procter & Gamble Company, FemCare Product Safety and Regulatory Affairs, 6110 Center Hill Avenue, Cincinnati, Ohio 45224, USA. [31]Second Genome, Inc. Bioinformatics Department, 1150 Bayhill Drive, Suite 215, San Bruno, California 94066, USA. [32]Forsyth Institute, Department of Molecular Genetics, 245 First Street, Cambridge, Massachusetts 02142, USA. [33]Harvard School of Dental Medicine, Department of Oral Medicine, Infection and Immunity, 188 Longwood Avenue, Boston, Massachusetts 02115, USA. [34]Washington University School of Medicine, Department of Pathology & Immunology, 660 South Euclid Avenue, Box 8118, St Louis, Missouri 63110, USA. [35]bioMerieux, Inc., 100 Rodolphe Street, Durham, North Carolina 27712, USA. [36]drive5.com, Tiburon, California 94920, USA. [37]Cleveland Clinic, Center for Bioethics, Humanities and Spiritual Care, 9500 Euclid Avenue, Cleveland, Ohio 44195, USA. [38]VIB, Belgium, Department of Structural Biology, Pleinlaan 2, 1050 Brussels, Belgium. [39]Vrije Universiteit Brussels, Department of Applied Biological Sciences (DBIT), Pleinlaan 2, 1050 Brussels, Belgium. [40]University of North Carolina Charlotte, Department of Bioinformatics and Genomics, 9201 University City Blvd, Charlotte, North Carolina 28223-0001, USA. [41]University of Idaho, Department of Biological Sciences, Life Sciences South Room 441A, PO Box 443051, Moscow, Idaho 83844, USA. [42]Massachusetts Institute of Technology, Computational and Systems Biology, Parsons Laboratory, Room 48-317, 15 Vassar Street, Cambridge, Massachusetts 02139, USA. [43]Saint Louis University, Center for Advanced Dental Education, 3320 Rutger Street, St Louis, Missouri 63104, USA. [44]University of Colorado, Department of Computer Science, University of Colorado, Boulder, Colorado 80309, USA. [45]University of Maryland Francis King Carey School of Law, 500 W. Baltimore Street, Baltimore, Maryland 21201, USA. [46]Marine Biological Laboratory, Josephine Bay Paul Center, 7 MBL Street, Woods Hole, Massachusetts 02543-1015, USA. [47]Harvard School of Dental Medicine, Department of Oral Medicine, Infection and Immunity, 188 Longwood Avenue, Boston, Massachusetts 02115, USA. [48]Lawrence Berkeley National Laboratory, Ecology Department, Earth Sciences Division, 1 Cyclotron Road, Berkeley, California 94720, USA. [49]University of Texas Health Science Center School of Dentistry, Department of Periodontics, 6516 MD Anderson Blvd, Houston, Texas 77030, USA. [50]San Diego State University, Department of Biology, 5500 Campanile Drive, San Diego, California 92182, USA. [51]UCLA School of Dentistry, Division of Associated Clinical Specialties and Dental Research Institute, 10833 Le Conte Avenue, Los Angeles, California 90095-1668, USA. [52]McGill University, Faculty of Medicine, Peel 3647 Montreal, Quebec H3A 1X1, Canada. [53]Howard Hughes Medical Institute, Campus Box 215, Boulder, Colorado 80309-0215, USA. [54]National Institutes of Health, National Cancer Institute (NCI), Dermatology Branch, CCR, MSC 1908, 10 Center Drive, Bethesda, Maryland 20892, USA. [55]Cornell University, Department of Microbiology, 467 Biotechnology Building, Ithaca, New York 14853, USA. [56]Washington University School of Medicine, Department of Medicine, Division of General Medical Science, 660 South Euclid Avenue, Box 8005, St Louis, Missouri 63110, USA. [57]Children's Hospital Boston, Harvard Medical School, Division of Infectious Diseases, 300 Longwood Avenue, Boston, Massachusetts 02115, USA. [58]University of Oklahoma, Department of Anthropology, 455 West Lindsey, Dale Hall Tower 521, Norman, Oklahoma 73019, USA. [59]Washington University School of Medicine, Department of Obstetrics and Gynecology, 4533 Clayton Avenue, Box 8219, St Louis, Missouri 63110, USA. [60]University of Alabama at Birmingham, Division of Gastroenterology and Hepatology, 1530 3rd Avenue South, Birmingham, Alabama 35294-1150, USA. [61]Baylor College of Medicine, Center for Medical Ethics and Health Policy, One Baylor Plaza, Houston, Texas 77030, USA. [62]Baylor College of Medicine, Medicine-Infectious Disease, One Baylor Plaza, Houston, Texas 77030, USA. [63]Oak Ridge National Laboratory, Biosciences Division, PO Box 2008 MS 6038 Oak Ridge, Tennessee 37831-6038, USA. [64]University of California, San Francisco, Gladstone Institutes, 1650 Owens Street, San Francisco, California 94158, USA. [65]University of California, San Francisco, Institute for Human Genetics, 1650 Owens Street, San Francisco, California 94158, USA. [66]University of California, San Francisco, Division of Biostatistics, 1650 Owens Street, San Francisco, California 94158, USA. [67]University of Maryland School of Medicine, Department of Microbiology and Immunology, BioPark II - Room 611, 801 W. Baltimore Street, Baltimore, Maryland 21201, USA. [68]Indiana University, School of Informatics and Computing, 150 S. Woodlawn Avenue, Bloomington, Indiana 47405, USA. [69]Mount Sinai School of Medicine, Annenberg Building Floor 5th, Room 5-208, 1468 Madison Avenue, New York, New York 10029, USA. [70]Baylor College of Medicine Molecular & Human Genetics, One Baylor Plaza, Houston,

Texas 77030, USA. [71]University of Pennsylvania, Center for Bioethics and Department of Medical Ethics, 3401 Market Street, Suite 320, Philadelphia, Pennsylvania 19104, USA. [72]University of Michigan, Department of Microbiology & Immunology, 5713 Medical Science Bldg. II, 1150 West Medical Center Dr., Ann Arbor, Michigan 48109-5620, USA. [73]Michigan State University, Department of Microbiology and Molecular Genetics, 6180 Biomedical Physical Sciences, Michigan State University, East Lansing, Michigan 48824, USA. [74]The EMMES Corporation, 401 N. Washington St., Suite 700, Rockville, Maryland 20850, USA. [75]Wayne State University School of Medicine, Detroit, Michigan, Harper University Hospital, 3990 John R Street, Detroit, Michigan 48201, USA. [76]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. [77]J. Craig Venter Institute, 10355 Science Center Drive, San Diego, California 92121, USA. [78]Northwestern University, Feinberg School of Medicine, 420 East Superior Street Chicago, Illinois 60611, USA. [79]Alkek Center for Metagenomics and Microbiome Research, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA. [80]National Institutes of Health, National Human Genome Research Institute (NHGRI), Genetics and Molecular Biology Branch, MSC 4442, Bethesda, Maryland 20892, USA. †Deceased.