HMP Data Set Documentation

Introduction

This document provides detail about files available via the DACC website.

The goal of the HMP consortium is to make the metagenomics sequence data generated by the NIH Human Microbiome Project rapidly and broadly available to the scientific community. The users of any data released before HMP publication of a paper describing the data are expected to act responsibly to recognize the scientific contribution of the data producers by following normal standards of scientific etiquette and fair use of unpublished data. Such guidelines can be found in Sharing Data From Large-Scale Biological Research Projects: A System of Tripartite Responsibility, <http://www.genome.gov/Pages/Research/WellcomeReport0303.pdf> and the Toronto Data Release Workshop (Nature 461, 168-170, 10 September 2009) | doi 10.1038/461168a; Published online 9 September 2009). A marker paper describing the NIH Human Microbiome Project and its data release policy can be found in Genome Research 19, 2317-2323 (December 2009) | doi 10:1101/gr096651.109; Published online 9 October 2009. We follow these principles and urge users to follow them as well. It is our intention to publish the work of this project in a timely fashion and we welcome collaborative interaction on the project and analyses.

Any publications arising from the use of project data should include the following acknowledgement:

This project has been funded or data has been generated in part with US federal funds from the NIH Human Microbiome Project, the Common Fund, National Institutes of Health, Department of Health and Human Services, (and if appropriate) under grant XXXX.

Questions concerning this project or the use of these data should be sent to Lita Proctor at lita.proctor@nih.gov or 301-496-4550

Production Data

Production data sets include sequence data and value-added datasets pertaining to the Reference Genomes, 16S and mWGS.

Human Microbiome Project (HMP) Reference Genomes (Project ID: 28331)

The HMP plans to sequence, or collect from publicly available sources, a total of 3000 reference genomes isolated from human body sites. The information gained from the Reference Genomes will aid in taxonomic assignment and functional annotation of 16S RNA and metagenomic sequence, respectively, from microbiome samples.
The HMP Project

Catalog interface allows for viewing, sorting and searching the full list of HMP Reference Genomes and associated metadata. Users can query for particular body sites, taxonomic groups, status levels, link to statistics, and create custom views of the data. In addition to reference genomes being sequenced as part of the NIH Roadmap-funded HMP project, the Project Catalog also includes reference genomes sequenced by members of the International Human Microbiome Consortium. The Project Catalog provides links for individual projects to downloadable sequence & annotation data housed at NCBI, as available.

As genomes are released in RefSeq they are incorporated into the Integrated Microbial Genomes-Human Microbiome Project (img/hmp) resource. IMG analysis of a genome includes extensive functional annotation, pathway analysis, and integration with other genomes. The data can be navigated along any of the three dimensions of gene, genome, or functional role/pathway information. See the img/hmp User Guide for more details.

There are currently two methods for downloading Reference Genome sequence & annotation data directly from the DACC:

Individual reference genome sequence & annotation data is available for download at Data | Reference Genomes. This page does not reflect every project found in the HMP Project Catalog, but only those that have completed sequencing and annotation. Isolates are organized by GenBank Project ID. This page is updated monthly as additional projects are released, or finishing levels or annotations are updated. The DACC routinely runs all completed reference genomes through the autoannotation pipeline used by the HMP sequencing centers in order to ensure the most up-to-date annotation. These updated annotations are submitted to NCBI for incorporation into their records and are included in the fasta headers of the protein fasta files available from Data | Reference Genomes.

The DACC Data Browse Interface provides bulk download of all annotated, publically available reference genomes in three formats:

- Assembly fasta: These files contain the full sequence of the assembly in FASTA format. Each FASTA header represents an individual contig or scaffold. Each file represents an individual genome project, with all contigs or scaffolds from that project concatenated into a single file. Files are organized by Genbank project id and are downloaded as a compressed tar archive. *Download data* (340MB)

- Genbank: These files contain the full Genbank records for all contigs or scaffolds submitted. Each file represents an individual genome project, with all contig or scaffold Genbank records from that project concatenated into a single file. Files are organized by Genbank project id and are downloaded as a compressed tar archive. *Download data* (748MB)

- Protein fasta: These files contain protein sequence for all annotated genes. Each FASTA header represents an individual protein sequence. FASTA headers are of the format: >locus tag, annotation (common name, gene symbol and EC number, when available), organism name. Each file represents an individual genom project, with all protein sequences from that project concatenated into a single file. Files are organized by Genbank project id and are downloaded as a compressed tar archive. *Download data* (196MB)

Bulk download of body-site specific reference genome datasets is coming soon.

16S rRNA sequencing is being used to characterize the complexity of microbial communities at individual body sites, and to determine whether there is a core microbiome at each site.

There are currently three data sets associated with **16S Production sequence data**:

HMP: 16S rRNA 454 Clinical Production Phase I (Project ID: 48333) *download data*
This HMP production phase represents pyrosequencing of 16S rRNA genes amplified from multiple body sites across hundreds of human subjects. There are two time points represented for a subset of these subjects. Using default protocol v4.2., data for the16S window spanning V3-V5 was generated for all samples, with a second 16S window spanning V1-V3 generated for a majority of the samples. 16S rRNA sequencing is being used to characterize the complexity of microbial communities at individual body sites, and to determine whether there is a core microbiome at each site. Several body sites will be studied, including the gastrointestinal and female urogenital tracts, oral cavity, nasal and pharyngeal tract, and skin.

HMP: 16S rRNA 454 Clinical Production Phase II (Project ID: 50563)
This HMP production phase represents pyrosequencing of 16S rRNA genes amplified from multiple body sites across hundreds of human subjects. In combination with Clinical Production Phase I (48333), this dataset represents samples derived from 300 healthy subjects, of which 200 were sampled at two time points and 100 were sampled at three time points. Using default protocol v4.2 data for 16S window spanning V3-V5 was generated for all samples, with a second 16S window spanning V1-V3 generated for a majority of the samples. 16S rRNA sequencing is being used to characterize the complexity of microbial communities at individual body sites, and to determine whether there is a core microbiome at each site. Several body sites will be studied, including the gastrointestinal and female urogenital tracts, oral cavity, nasal and pharyngeal tract, and skin. ***Data deposits to NCBI for this production data set are expected in January 2011.***

HMP: 16S rRNA Sanger Clinical Production (Project ID: 48469) *download data*
This HMP production phase represents dideoxy sequencing of 16S rRNA genes amplified from multiple body sites across hundreds of human subjects. 16S rRNA sequencing is being used to characterize the complexity of microbial communities at individual body sites, and to determine whether there is a core microbiome at each site. Several body sites will be studied, including the gastrointestinal and female urogenital tracts, oral cavity, nasal and pharyngeal tract, and skin.

Two clinical production pilots were conducted for the 16S data.

HMP: 16S rRNA 454 Clinical Production Pilot (Project ID: 48335) *download data*
The HMP clinical production pilot study represents the pyrosequencing of 16S rRNA genes amplified from multiple body sites across a small number of subjects. Each clinical sample was sent to two HMP sequencing centers, each of which sequenced amplicons from at least one window, and maximum two (with default window: V3-V5, and additional window V1-V3). The goal of the study is to quantify the variation among replicates processed by various centers, to identify the optimal variable span of the 16S rRNA gene to target, and to gain a preliminary understanding of the diversity of microbial community structures among samples extracted from distinct body sites.

HMP: 16S rRNA Sanger Clinical Production Pilot (Project ID: 34129) *download data*
The HMP clinical production pilot study represents the dideoxy sequencing of 16S rRNA genes amplified from multiple body sites across a small number of subjects. The goal of the pilot was to test the production Sanger sequencing protocol, to quantify the variation among replicates processed by various centers, and to gain a preliminary understanding of the diversity of microbial community structures among samples extracted from distinct body sites.

In addition, there are a number of value-added 16S rRNA datasets. These datasets have been generated by the DACC and by various members of the HMP Research Network. The value added datasets generated by the HMP DACC are derived from the raw datasets publicly available at NCBI, and have had some additional processing done. These datasets are available to the general public and research community. Datasets generated by the members of the HMP Research Network have had much more extensive processing, often to the specifications of a particular project or experiment. These datasets are currently only available to the members of the Research Network and that project.

Production Phase 1 Data with Chimeras removed *download data*
Generated by Jonathan Crabtree, HMP DACC
~72 million 16S reads in the public SRA for SRP002395, Chimera Slayer results
Concatenated and gzipped .CPS.CPC ChimeraSlayer results for each of the sequence alignments in SRP002395-7514-nast-ier.NAST.gz
ChimeraSlayer.pl from the 2010-04-29 release of the Broad microbiome utilitieswas run using the default parameter

Production Phase 1 Data with RDP Classifier results *download data*
Generated by Jonathan Crabtree, HMP DACC
~72 million 16S reads in the public SRA for SRP002395 , RDP Classifier results
Concatenated and gzipped 'allrank' RDP Classifier results for each of the sequences in SRP002395-7514-cs-nbp-rc.fsa.gz.
Version 2.2 of the RDP Classifier was run using the default 032010 training set and taxonomy and the 'allrank' output format option

Production Phase 1 Data with WigeoN results *download data*
Generated by Jonathan Crabtree, HMP DACC
~72 million 16S reads in the public SRA for SRP002395, WigeoN results
Concatenated and gzipped WigeoN results for each of the sequence alignments in SRP002395-7514-nast-ier.NAST.gz
run_WigeoN.pl from the 2010-04-29 release of the Broad microbiome utilities was run using the default parameters

Production Phase 1 Data, Alignments *download data*
Generated by Jonathan Crabtree, HMP DACC
~72 million 16S reads in the public SRA for SRP002395, -NAST-iEr alignments (of the entire clear span, minus barcode and primer(s), but with no other trimming)
Concatenated and gzipped .CPS.CPC ChimeraSlayer results for each of the sequence alignments in SRP002395-7514-cs-nbp-rc.fsa.gz
Run_NAST-iER.pl from the 2010-04-29 release of the Broad microbiome utilities was run using the default parameter

Production Phase 1 Trimmed Data set *download data*
Generated by Jonathan Crabtree, HMP DACC
~72 million 16S reads in the public SRA for SRP002395. Gzipped multi-FASTA file of reverse complemented 454 clear ranges, with the following subsequences removed:
1. initial "TCAG" (must have been present in the original read)
2. reverse barcode sequence (must have been present in the original read)
3. reverse primer sequence (must have been present in the original read)
4. forward primer sequence (if present within the clear range)

V13, V35, V69 High Quality Sequence Summary *download data (locked)*
Generated by Patrick Schloss, University of Michigan
This is a large file that contains sequence-level information about each sequence that survived the *high quality* curation process. The file contains seven fields, each separated by a tab:
1. seqName - the sequence name as generated by the sequencer. If the sequence was part of the PPS dataset, a "PPS_" was appended to the beginning of the sequence name
2. collection - this field concatenates the nap_id and dataset fields from the pds.metadata file and is used throughout the subsequent files
3. taxonomy - the taxonomic assignment of the sequence to the RDP training set using the Bayesian classifier. Numbers in parentheses indicate the pseudo-boostrap value. When the value dropped below 80, the name "unclassified" was appended to reach the genus level
4. alignment - the aligned version of the sequence. This alignment represents a version of a silva-based alignment where vertical columns lacking any bases (i.e. only containing "-") were removed and columns were data may have been missing (i.e. any position containing a "." in any sequence). The original unfiltered sequences are available from Pat Schloss.
5. otu - this number indicates the OTU number that the sequence belonged to. The first OTU is otu #1.
6. phylotype - this number indicates the phylotype number that the sequence belonged to. The first phylotype is phylotype #1. Phylotypes were assigned by binning sequences based on the full taxonomy string listed above.
7. sequence - this is the unaligned, raw sequence that survived the initial quality control trimming step

V13, V35, V69 Metadata *download data (locked)*
Generated by Patrick Schloss, University of Michigan
This text file pieces together some level of metadata based on the sequencing metadata and observations from the data. To create the overall structure of the file, data was concatenated from the *.lmd files obtained from the DACC. Not all samples represented in this table were actually observed in the overall dataset. There are twelve tab delimited columns in the file.

V13, V35, V69 Low Quality Sequence Summary *download data (locked)*
Generated by Patrick Schloss, University of Michigan
This is a large file that contains sequence-level information about each sequence that survived the *low quality* curation process. The file contains four fields, each separated by a tab:
1. seqname - the sequence name as generated by the sequencer. If the sequence was part of the PPS data set, a "PPS_" was appended to the beginning of the sequence name
2. collection - this field concatenates the nap_id and dataset fields from the pds.metadata file and is used throughout the subsequent files
3. taxonomy - the taxonomic assignment of the sequence to the RDP training set using the Bayesian classifier. Numbers in parentheses indicate the pseudo-bootstrap value. When the value dropped below 80, the name "unclassified" was appended to reach the genus level

4. sequence - this is the unaligned, raw sequence that survived the initial quality control trimming step

V13, V35, V69 Phylotype Counts *download data (locked)*
Generated by Patrick Schloss, University of Michigan
This file is a large table indicating the number of sequences from each sample that is affiliated with each phylotype. The first column of the table is a heading - "collection" which represents the concatenation of the nap_id and dataset fields from the pds.metadata file. Subsequent columns, separated by tabs [\t], indicate the phylotype number increasing from one in the second column towards the right side of the table. Rows in the tables indicate separate collections. There will be as many rows as there were collections in the study and as many columns as there were phylotypes.

V13, V35, V69 OTU Counts *download data (locked)*
Generated by Patrick Schloss, University of Michigan
This file is a large table indicating the number of sequences from each sample that is affiliated with each otu. The first column of the table is a heading - "collection" which represents the concatenation of the nap_id and dataset fields from the pds.metadata file. Subsequent columns, separated by tabs [\t], indicate the otu number increasing from one in the second column towards the right side of the table. Rows in the tables indicate separate collections. There will be as many rows as there were collections in the study and as many columns as there were otus.

V13, V35, V69 Phylotype Lookup *download data (locked)*
Generated by Patrick Schloss, University of Michigan
This table is a two column list. The first column (heading=phylotype) indicates the number that corresponds to the phylotype from the *.counts files. The second column (heading=taxonomy) indicates the consensus taxonomy for that phylotype. The number in parentheses indicates the percentage of sequences within that phylotype that had the same taxonomy. To reach a particular taxonomic level, at least 50% of the sequences in the phylotype had to have the same taxonomy to that level.

V13, V35, V69 OTU Lookup *download data (locked)*
Generated by Patrick Schloss, University of Michigan
This table is a two column list. The first column (heading=otu) indicates the number that corresponds to the otu from the *.counts files. The second column (heading=taxonomy) indicates the consensus taxonomy for that otu. The number in parentheses indicates the percentage of sequences within that otu that had the same taxonomy. To reach a particular taxonomic level, at least 50% of the sequences in the otu had to have the same taxonomy to that level.

V13, V35, V69 Alpha Summary *download data (locked)*
Generated by Patrick Schloss, University of Michigan
This is a table providing alpha diversity information for each sample in the study based on otu, phylotype, and phylogenetic approaches (in that order). The first column heading is the collection name. Each row represents a different sequence collection named by concatenating the nap_id and dataset fields from the pds.metadata file. There is also a row that provides values for all of the sequences in the analysis. For the alpha diversity metrics based on otu and phylotype frequencies there are 34 column headings are prefixed with either an "otu_" or a "phylotype_"
Finally, there is a column - phyloDiversity - that contains the phylogenetic diversity represented by the subtree containing all of the sequences within the sequence collection

V13, V35, V69 Phylogenetic Diversity Rarefication *download data (locked)*

Generated by Patrick Schloss, University of Michigan
Because phylogenetic diversity (i.e. the total branch length of a tree) can not be interpreted independently of the number of sequences sampled, it is necessary to rarefy the phylogenetic diversity as a function of sampling effort. This file contains the data based on 100 randomizations. Each column represents a separate sequence collection and each row the number of sequences sampled. When a column lists "NA", this indicates that sequencing was not pursued to this depth. Phylogenetic diversity values are outputted every 100 sequences, unless one sampling of a collection terminates between the 100 sequence marks

V13, V35, V69 Beta Summary *download data (locked)*
Generated by Patrick Schloss, University of Michigan
This file represents a table listing beta-diversity measures between pairs of sequence collections using otu, phylotype, and phylogenetic-based methods (in that order). The first two column headings are the names of the two collections being compared. Each row represents a different comparison. For the beta diversity metrics based on otu and phylotype frequencies there are 24 column headings are prefixed with either an "otu_" or a "phylotype_"

There is only one set of metagenomic whole genome shotgun data currently at NCBI. Metagenomic whole genome shotgun (mWGS) sequencing will provide insights into the functions and pathways present in the human microbiome, and will generate a reference framework for those looking into associations between changes in the human microbiome and disease states.

Expanding on the 16S rRNA data obtained in another component of the Human Microbiome Project (HMP), whole genome shotgun (WGS) sequencing will be performed on samples taken from the digestive tract, mouth, skin, nose, and female urogenital tract of human subjects to gain insight into the genes and pathways present in the human microbiome.

Human Microbiome Project Metagenomes Production Phase **(Project ID: 48479)** (*download data*)
The HMP metagenomes production phase represents the shotgun sequencing of metagenomic DNA extracted from samples taken from multiple body sites across hundreds of human subjects. Coupled with the other data generated during the HMP project, these results will provide insights into the genes and pathways present in the human microbiome. We are learning more and more about the ways that human health is influenced by the complex and dynamic communities of microbes (the human microbiota) present on and within our bodies. Disruptions in these communities may trigger or influence the course of various disease states. Preliminary studies have shown this to be the case for certain diseases. The Human Microbiome Project expands on these studies to better understand, prevent and treat many human diseases.

Project Data

The Project Data consists of the data sets associated with specific, funded projects. These were initially referred to as the Demonstration Projects. General information on these projects can be found on the Impacts on Health page. Fifteen Demonstration projects were initially funded to demonstrate hypothesized correlations between the microbiome and human health and disease.

Validation Data

Validation data sets were used during protocol development and protocol validation. Details on the protocols used can be found on the Tools and Protocols page.

HMP 16S rRNA 454 Impact of PCR on Chimera Formation – Mock (Project ID: 50501) *download data*
The impact of PCR conditions on chimera content was evaluated by pyrosequencing the HMP even and staggered Mock community. The following PCR parameters were varied: cycle number (20x, 30x, or 40x), template genomic DNA concentration (0.1 ng, 1 ng, or 10 ng), and extension time (1 min. or 5 min.)

HMP 16S rRNA 454 Protocol Development – Clinical (Project ID: 48467) *download data*
The HMP early clinical pilot represents the pyrosequencing of 16S rRNA genes amplified from anonymized clinical samples. This pilot was conducted using a pre-production version of the 454 sequencing protocol which has since been deprecated

HMP 16S rRNA 454 Protocol Development – Mock (Project ID: 48465) *download data*
The HMP early mock pilot represents the pyrosequencing of 16S rRNA genes amplified from HMP even and staggered Mock communities, distributed to each of the four HMP sequencing centers. This pilot was conducted using a pre-production version of the 454 sequencing protocol which has since been deprecated. The main differences with the production protocol are: no use of barcodes, bi-directional sequencing, pools containing 3 different 16S windows (V1-V3, V3-V5, V6-V9).
(Previously referred to as Pre-CEFoS Mock Pilot)

HMP 16S rRNA 454 Protocol Validation – Clinical (Project ID: 48339) *download data*
This HMP Centers' evaluation of the default 454 SOP represents the pyrosequencing of 16S rRNA genes amplified from anonymized clinical stool samples distributed to each of the four HMP sequencing centers. The goal of the pilot was to test the provisional 454 sequencing protocol (v4.2) and to evaluate accuracy and consistency between centers. (Previously referred to as Centers Evaluation of 454 SOP (CEFoS) Clinical Pilot)

HMP 16S rRNA 454 Protocol Validation – Mock (Project ID: 48341) *download data*
This HMP Centers' Evaluation of the standard 454 SOP represents the pyrosequencing of 16S rRNA genes amplified from HMP even Mock community distributed to each of the four HMP sequencing centers. The goal of the pilot was to test the provisional 454 sequencing protocol (v4.2) and to evaluate accuracy. This protocol does use barcodes, targets two 16S windows (V1-V3 and V3-V5), and sequences in the reverse direction. (Previously referred to as Centers' Evaluation of 454 SOP (CEFoS) Mock Pilot)

HMP 16S rRNA Sanger Protocol Validation – Clinical (Project ID: 48547) *download data*
The HMP Sanger clinical pilot represents the dideoxy sequencing of 16S rRNA genes amplified from anonymized stool samples. The goal of the pilot was to test the provisional Sanger sequencing protocol and to evaluate accuracy and consistency between centers

HMP 16S rRNA Sanger Protocol Validation – Mock (Project ID: 48471) *download data*

The HMP Sanger mock pilot represents the dideoxy sequencing of 16S rRNA genes amplified from HMP even and staggered Mock communities, distributed to each of the four HMP sequencing centers. The goal of the pilot was to test the provisional Sanger sequencing protocol and to evaluate accuracy.

HMP 16S rRNA Sanger Impact of Capillary Length on Sequence Quality **(Project ID: 60769)** *download data*
The impact of sequencing conditions on sequence quality was evaluated by sequencing a variety of samples, including HMP even Mock community and clinical samples. The following PCR parameters were tested in addition to the original protocol: lower cycle number (20x), and longer extension time (10 min.)

Human Microbiome Project 16S rRNA Sanger Impact of PCR on Chimera Formation **(Project ID: 60767)** *download data*
The impact of PCR conditions on chimera content was evaluated by dideoxy sequencing the HMP even Mock community, targeting the V3-V5 region. The following PCR parameters were tested in addition to the original protocol: lower cycle number (20x), and longer extension time (10 min.)

Human Microbiome Project 16S rRNA Sanger Protocol Development - Clinical **(Project ID: 60763)** *download data*
The HMP early clinical pilot represents the dideoxy sequencing of 16S rRNA genes amplified from anonymized clinical stool samples, distributed to each of the four HMP sequencing centers. This pilot was conducted using a variety of pre-production versions of the sanger sequencing protocol to test primers, buffers, and PCR conditions.

Human Microbiome Project 16S rRNA Sanger Protocol Development - Mock **(Project ID: 60765)** *download data*
The HMP early mock pilot represents the dideoxy sequencing of 16S rRNA genes amplified from the early even Mock community (constructed by Broad). These experiments were conducted using a variety of pre-production versions of the sanger sequencing protocol to test primers, buffers, and PCR conditions.

Human Microbiome Project Metagenomes Mock Pilot **(Project ID: 48475)** (*download data*)
The HMP metagenomes mock pilot represents the shotgun sequencing of HMP even and staggered Mock communities, distributed to each of the four HMP sequencing centers. The goal of the pilot was to test the sequencing protocol and to evaluate accuracy and consistency between centers.