A Data Analysis and Coordination Center for the Human Microbiome Project

Jennifer Wortman1, Michelle Giglio1, Amy Chen2, Victor Felix1, Konstantinos Mavrommatis1, Heather Creasy1, Todd DeSantis2, Clark Santee2, Yvette Piceno2, Shariff Osman2, Joshua Orvis1, Jonathan Crabtree1, Micah Hamady1, Justin Kuczynski1, Rob Knight1, Gary Andersen1, Nikos Kyrpides3, Victor Markowitz2, Owen White1

1Institute for Genomic Sciences, University of Maryland School of Medicine, Baltimore, MD; 2Lawrence Berkeley National Laboratory, Berkeley, CA; 3Joint Genome Institute, Walnut Creek, CA; 4University of Colorado at Boulder, Boulder, CO

Abstract

The Human Microbiome Project (HMP) is an NIH Roadmap initiative that aims to collect and analyze unprecedented amounts of sequence information from microbial communities found in and on the human body. There is abundant and growing evidence that changes in microbial community composition are highly correlated with human health and disease. Efforts are underway to determine if such changes are the result of particular human diseases or perhaps a contributing cause. To gain insight into this question, the HMP has undertaken two major areas of effort: sequence 1000 reference genomes that live in or on the human body and sequence metagenomic samples from five different body sites collected in parallel from healthy subjects and those with disease. Initially, four sequencing centers have begun the work of sequencing the 1000 reference strains. Subsequently, centers will be funded to carry out metagenomic sequencing from various sites with subjects suffering from various conditions. This project will generate unprecedented amounts of sequence data, annotation information, and metadata about subjects and strains. The analysis of this data requires the ability to collect, integrate, and standardize information of different types and from different sources. Responsibility for these activities falls on the HMP Data Analysis and Coordination Center (DACC). Successful data integration and standardization will rely on the use of controlled vocabularies, the application of quality control measures, and the development of standard operating procedures. The DACC will provide multiple analysis services to the research community including data query, comparative genomics, 16S rRNA analysis, and phylogenetic analysis. The DACC will also engage in extensive training and outreach.

All information and analyses produced from the HMP will be available on a comprehensive web resource. The web resource supports the comparative analysis of completed reference genomes, and includes extensive functional annotation and pathway information. The reference genomes generated for the HMP will provide a benchmark against which metagenomic sequence data can be compared.

Outreach Activities

The DACC Web Resources

Welcome

Welcome to the Data Analysis and Coordination Center (DACC) for the Human Microbiome Project (HMP). The DACC was launched by the National Institutes of Health Roadmap for Medical Discovery and is designed to fulfill research into the multitude of microbes that live in the various environments of the human body. A major goal of the HMP is to build for correlations between changes in the microbiome and human health. More information about this project can be found on the NIH Roadmap website at http://nihroadmap.nih.gov.

The HMP DACC is the central data-processing and analysis infrastructure to facilitate analyses of the microbiome. The HMP DACC web site will provide web-based query and visualization tools, comprehensive computational and analysis of HMP data, quality control measures, and links to well-documented standard operating procedures. The DACC is also strongly committed to outreach and training. Please see the above links for more information on each of these topics.

The HMP Project Catalog maintains an interactive list of the targeted reference strains along with sequencing status and links to public databases.

The IMG/HMP web resource supports the comparative analysis of completed reference genomes, and includes extensive functional annotation and pathway information.

Reference Strain Selection

We encourage feedback from the scientific community on the selection of strains to include in the HMP reference collection.

http://www.hmpdacc.org/feedback_form.php

Most strains chosen as reference genomes for the project will be sequenced to “draft” level. However, about 15% of the reference strains will be taken closer to a “finished” or complete state. Criteria have been established to help guide the choice of which strains to advance in the finishing process.

The HMP project is also interested in collaborating with researchers who have biological materials for bacterial strains isolated from human body sites. Any researcher who would like to contribute cells or DNA from a relevant strain should contact us.

Human Microbiome Project

The HMP Project Catalog

Training

The DACC and sequencing centers offer a host of workshops each year. Topics include:

- Single genome sequencing
- Metagenome sequencing
- Annotation pipelines
- Manual curation
- Metagenomic data analysis

For more information, see:

http://www.hmpdacc.org/outreach.php

International Human Microbiome Consortium

Nine countries from around the world have formed the International Human Microbiome Consortium (IHMC) to unravel the complexities of the microbial communities living within all humans. The DACC will interact with these other centers in several ways:

- Coordinate reference strain selection
- Share protocols and methods to ensure consistency across the entire consortium
- Collaborate on the establishment of standards to be applied by all members
- Collect, integrate, and display reference genome, metagenomic, and 16S rRNA data from the international members

Tool Development

The DACC will also contribute to the development and improvement of computational tools to facilitate human microbiome data analysis.

Unifrac is a statistical tool for metagenomic community comparisons (Lozupone, 2006)

Performance of Fast UniFrac (red lines) versus original implementation on sample sizes ranging from 1000 to 10,000 sequences. Note log scale on y axis: 10000

References

16S rRNA Gene Sequencing

16S rRNA gene sequencing will be used to characterize the complexity of microbial communities at individual body sites, and to determine whether there is a core microbiome. Several body sites will be studied, including the gastrointestinal tract, oral cavity, nasopharyngeal tract, female urogenital tract, and skin.

The DACC quantifies the noise introduced by the 454 16S sequencing method—unexpected diversity seen from simple communities. Genomic DNA from 22 species (19 genera) were combined in equal molar quantities, and the V1 and V2 regions of the 16S rRNA gene amplified by PCR. Amplicons were sequenced using the 454 FLX platform. The resulting records were pre-clustered in base-call similarities. Each pre-cluster was then clustered based on sequence similarity and de-nosed (Quince, 2009, in review) resulting in 283 distinct sequence. An NJ tree was cast by Clearcut (Shineman, 2006) then visualized with iTOL (Letunic, 2007). Branches are colored by closest-matching genus (of the 19) as identified by Greengenes (DeSantis, 2006) and red bars indicate the relative abundance of each sequence type. Notice that reads from the same organism do not form distinct clades.

Length and lase-calle polymorphisms may also be the source of observed diversity. One organism was chosen to more thoroughly investigate the source of the observed diversity. The 38 distinct organismal clusters (representing 662 organisms) putatively generated from Clostridium beijerinckii were compared to database sequences based on common k-mers. The comparison revealed that de-nosed records diverged from Clostridium beijerinckii references up to 11%. Variation was also observed in lengths of de-nosed records (0.2 to 0.4 kb) but did not correspond to divergence. The DACC is collaborating with Dr. Quince to determine improved parameters for removing noisy base calls from the 454 data.

Funded by:

NIH Common Fund

DACC Partners

References

- DACC Web Resources
- DACC Partners
- Outreach Activities
- Tool Development