

# A Data Analysis and Coordination Center for the Human Microbiome Project

Jennifer Wortman<sup>1</sup>, Michelle Giglio<sup>1</sup>, Amy Chen<sup>2</sup>, Victor Felix<sup>1</sup>, Konstantinos Mavrommatis<sup>3</sup>, Heather Creasy<sup>1</sup>, Todd DeSantis<sup>2</sup>, Clark Santee<sup>2</sup>, Yvette Piceno<sup>2</sup>, Sharif Osman<sup>2</sup>, Joshua Orvis<sup>1</sup>, Jonathan Crabtree<sup>1</sup>, Micah Hamady<sup>4</sup>, Justin Kuczynski<sup>4</sup>, Rob Knight<sup>4</sup>, Gary Andersen<sup>2</sup>, Nikos Kyrpides<sup>3</sup>, Victor Markowitz<sup>2</sup>, Owen White<sup>1</sup>

<sup>1</sup>Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD; <sup>2</sup>Lawrence Berkeley National Laboratory, Berkeley, CA;

<sup>3</sup>Joint Genome Institute, Walnut Creek, CA; <sup>4</sup>University of Colorado at Boulder, Boulder, CO

Michelle Giglio  
Institute for Genome Sciences  
University of Maryland School of Medicine  
801 W. Baltimore St.  
Baltimore, MD 21201  
phone: 410-706-7694  
fax: 410-706-6756  
mgiglio@som.umaryland.edu

## 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.

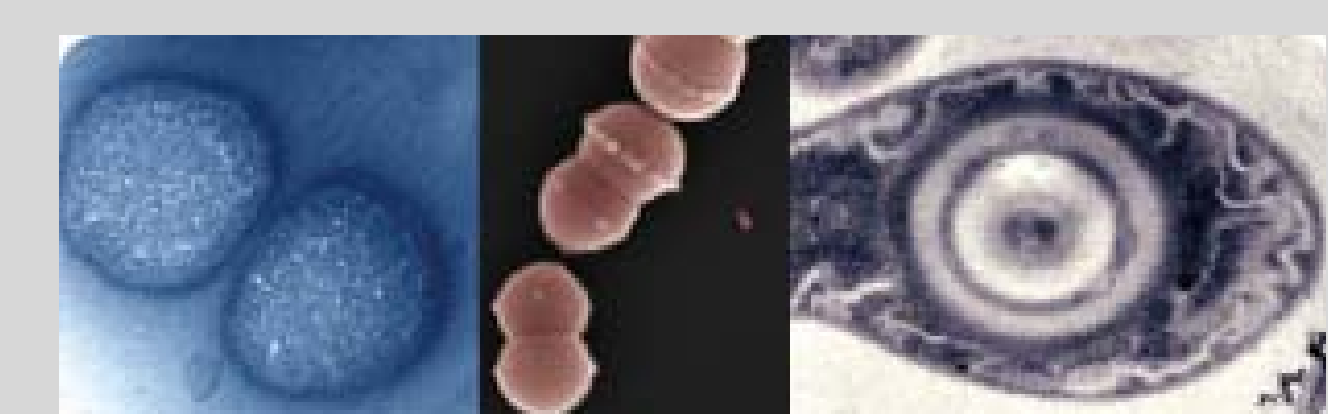
To gain insight into this question, the HMP has undertaken two main areas of effort: sequence 1000 reference genomes that live in or on the human body and sequence metagenomic samples from five different body sites collected in parallel from healthy subjects and those with disease. Initially, four large 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 presentation toolsets for the DACC will be based on those of the Integrated Microbial Genomes resource for both single genomes and metagenomes (IMG and IMG/m). The HMP DACC can be found at <http://hmpdacc.org>.

## DACC Web Resources

**Reference Genomes**  
Approximately 600 microbes will be sequenced during the HMP. Combined with other existing and currently planned efforts, the total reference collection should reach 1000 genomes. These sequences will provide a benchmark against which metagenomic sequence data can be compared.



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

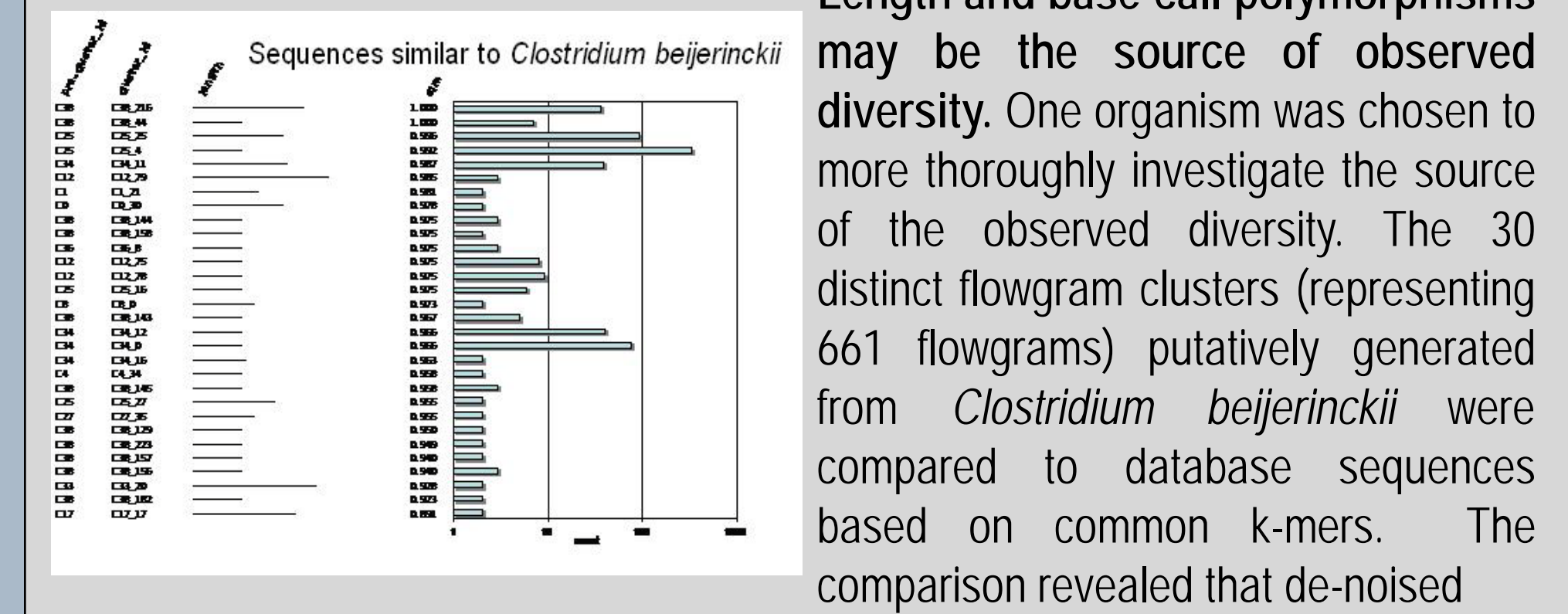
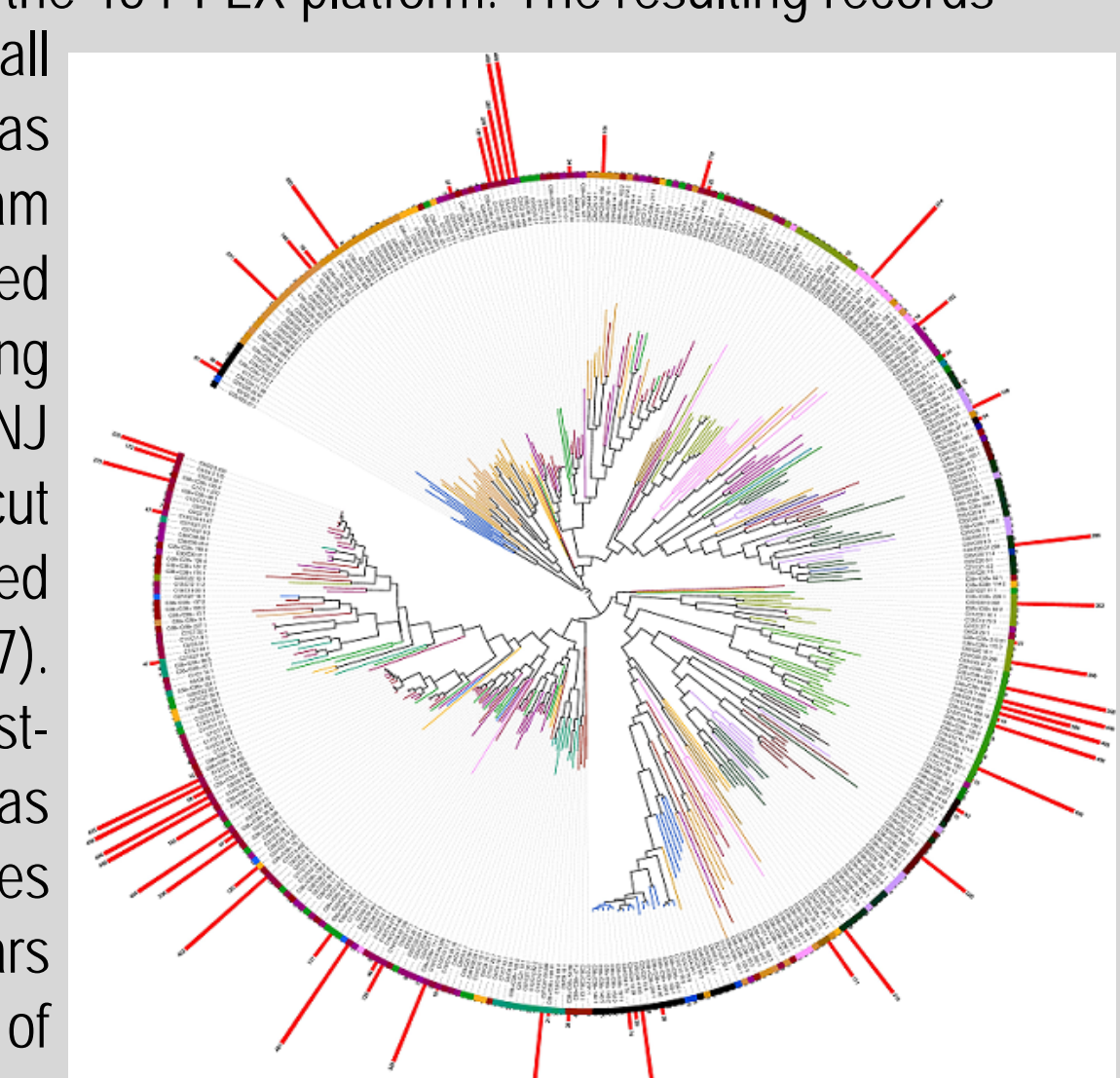
The IMG/HMP web resource supports the comparative analysis of completed reference genomes, and includes extensive functional annotation and pathway information

## Analysis of Microbiome Data

**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 on base-call similarities. Each pre-cluster was then clustered based on flowgram similarities and de-noised (Quince, 2009, in review) resulting in 393 distinct sequences. An NJ tree was cast by Clearcut (Sheneman, 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 base-call polymorphisms may be the source of observed diversity. One organism was chosen to more thoroughly investigate the source of the observed diversity. The 30 distinct flowgram clusters (representing 661 flowgrams) putatively generated from *Clostridium beijerinckii* were compared to database sequences based on common k-mers. The comparison revealed that de-noised records diverged from *Clostridium beijerinckii* references up to 11%. Variation was also observed in lengths of de-noised 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.

## DACC Partners

Funded by:  
NIH Common Fund

## Outreach Activities

**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](http://www.hmpdacc.org/feedback_form.php)

Most strains chosen as reference genomes for this 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.

**Training**

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

- 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 insure 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: Fast UniFrac implementation is consistently about 2 orders of magnitude faster, and largely eliminates the difference in time to calculate weighted and unweighted UniFrac metrics.

Please visit booth 1339 to learn more about this and other Institute for Genome Sciences projects and services