PROJECT SUMMARY

The Human Microbiome and Recurrent Abdominal Pain in Children

James Versalovic, M.D, Ph.D., Robert J. Shulman, M.D., Kevin Riehle, M.S., Delphine M. Saulnier, Ph.D., Toni-Ann Mistretta, Ph.D., Maria Alejandra Diaz, Ph.D., Sabeen Raza, M.S., Debasmita Mandal, Ph.D, Xiang Qin, Ph.D., Susan Lynch, Ph.D., Joseph Petrosino, Ph.D., Richard Gibbs, Ph.D.

I. PROJECT DESCRIPTION, PUBLICATION MORATORIUM INFORMATION AND PROJECT ID

This manuscript is part of a pilot effort on the part of NIH staff and the Nature publishing group to provide a more convenient archive for "marker papers" to be published. These "marker papers" are designed to provide the users of community resource data sets with information regarding the status and scope of individual community resource projects. For further information see editorial in September 2010 edition of Nature Genetics (*Nature Genetics*, **42**, 729 (2010)), and the Nature Precedings HMP summary page.

Project ID: 46339

The length of the publication moratorium is one year from the date of initial data release.

This project explores the nature of the human intestinal microbiome in healthy children and children with recurrent abdominal pain. The overall goal is to obtain a robust knowledge base of the intestinal microbiome in children without evidence of pain or gastrointestinal disease and in those with recurrent abdominal pain (functional abdominal pain (FAP) and FAP associated with changes in bowel habits, i.e., irritable bowel syndrome or IBS). Specific aims include: 1. Characterize the composition of the gut microbiome in healthy children by DNA sequencing. 2. Determine the presence of disease-specific organism signatures of variable gut microbiomes in children with recurrent abdominal pain. 3. Perform functional gut metagenomics by evaluation of whole community gene expression profiles and discovery of disease-specific pathway signatures. Multiple strategies have been deployed to navigate and understand the nature of the intestinal microbiome in childhood. These strategies included 454 pyrosequencing-based strategies to sequence 16S rRNA genes and understand the detailed composition of microbes in healthy and disease groups. Microarray-based hybridization with the PhyloChip and quantitative real-time PCR (qPCR) probes were applied as complementary strategies to gain an understanding of the intestinal microbiome from various perspectives. Data collected and analyzed during the HMP UH2 Demo project, from a set of healthy and IBS children (7-12 yo) may enable the identification of core microbiomes in children, in addition to variable components that may distinguish healthy from diseased pediatric states. Twenty-two children with IBS and twenty-two healthy children were enrolled and analyzed in the UH2 phase of this study. The planned enrollment targets for the UH2/3 phases include 50 healthy children, 50 children with FAP and 50 children with IBS (minimum of 3 time points per child). We are currently analyzing the dataset for the presence of

disease-specific signatures in the human microbiome, and correlating these microbial signatures with pediatric health or IBS disease status in addition to IBS subtype (e.g., diarrhea-vs constipation-predominant). In the next phase, whole genome shotgun sequencing and metatranscriptomics will be performed with a subset of children in each group. This study explores the nature of core and variable human microbiom in pre-adolescent healthy children and children with IBS.

II. DATA QUALITY:

The only metagenomic 16S sequencing method utilized in this study was 454 pyrosequencing (GS FLX with Titanium chemistry) using bar-coded universal bacterial primers 27F and 534R (V1V3 region), and primers 357F and 936R (V3V5 region). A total of 71 samples from 44 children were sequenced. We monitored the data quality of 454 sequencing data and applied the following filtering criteria to select the reads to be included in the analyses: the reads must contain at least 200 nucleotides; the average Qval for each read must be at least 20; the reads should have an exact bar code and exact "A" primer match; the read will be cut at first 'N' or 'n'; and the read can have up to 4 mismatches with the "B" primer. Using these filtering criteria, 75.7% of the 16S 454 reads were selected for analyses. After filtering, the average read length was 503 nt. Total kilobases generated was 1,938,764,206 and the total number of reads was 3,854,402.

Phylochip hybridization data included redundant probes for each taxa, and internal quality control within each chip.

III. DATA ANALYSIS AND PUBLICATION PLANS:

Data analysis:

Our research group plans to complete the comparative analyses of pediatric intestinal microbiomes from healthy children and children with IBS. Two primary studies are ongoing. We are comparing the intestinal microbiomes of healthy children (7-12 yo) with that of healthy adults (from the Jumpstart component of HMP). Secondly, we are comparing the intestinal microbiomes of healthy children with that of children with disease (IBS in this case). All metagenomic 16S and Phylochip data are examined using a variety of statistical algorithms in order to identify sources of variation in the microbiomes of healthy and IBS children. More specifically, the 16S data are analyzed using both phylogenetic and OTU methods contained in the QIIME platform version 1.0.0 (Caporuso. Nature Methods 2010; **7:** 335-6**)**. We employ a multistep OTU picking process in QIIME. The full input dataset is first run through a fast method to collapse similar sequences and then a slower more robust method such as CDHIT or Mothur (furthest/nearest neighbors). OTU maps are merged to match final OTU IDs with Input Identifiers. The QIIME platform allows us to generate a series of Phylogenetic and Non-phylogenetic Principal Coordinate Analyses (PCoA) plots. Evaluating these multiple types of PCoA analysis will allow us to identify and confirm clustering by age group (7-9, 10-12), gender, status (IBS/Healthy), stool characteristics and frequency, pain quality and severity, and interference with daily activities. Moreover, different analyses should help us to define a core microbiome in healthy children and disease signatures that correlate with disease phenotypes. The QIIME platform also allows for flexibility in choosing a variety of taxonomies (RDP, Silva

(Schloss), Greengenes). We have developed a RDP pipeline which utilizes the RDP classifier to generate taxa at the Phylum, Class, Order, Family and Genus level. Machine learning methods are being developed in order to further interrogate the 16S rRNA gene sequence data. Phylochip data were further analyzed to examine (dis)similarity in community composition across all samples. A Bray-Curtis distance matrix is generated and used to construct a non-metric dimensional scaling plot of communities present on the PhyloChip. Comparisons between the results of the PhyloChip and 454 data are also performed in order to identify taxa that are only identified by one technique, and could help to refine the primer design strategy for 454 DNA sequencing. Future studies include whole genome shotgun sequencing (Illumina GA and HiSeq) and metatranscriptomics (ABI SOLiD).
Publication plan: We plan to submit the initial results of 16S metagenomic sequencing, PhyloChip, and qPCR results for publication. We also plan to describe different IBS subtypes that have been identified using different methods during 2011, before the end of the publication moratorium period.

IV. DATA RELEASE PLAN:

Data sharing plan: Data and interpretations of findings will be openly shared with the human microbiome research community and scientific community at-large. Consistent with any research program, the P.I. and collaborators will openly share and communicate about data as information is obtained during the study. As members of the DACC (Data Analysis & Coordination Center) consortium, we are responsible for sharing data among consortium members and uploading data into the public databases such as dbGaP and SRA. Findings with the outside scientific community will be shared by a variety of means including electronic mail communication in response to specific queries, voice communication with individual scientists interested in the findings, and oral/written/poster presentations at meetings and conferences. Abstracts summarizing data may be published in scientific journals, and manuscripts will be submitted to journals of interest. Published manuscripts will be deposited in PubMed Central, and these written articles will ultimately be freely available to the scientific community at-large. Data reports will also be shared in a more comprehensive manner by written/oral communications with other research groups funded via this NIH Common Fund effort, the Human Microbiome Project.

The Baylor College of Medicine Human Genome Sequencing Center (HGSC), which will do the sequencing for this project, has a long history of many community outreach activities. The HGSC has worked with minorities, the public at large, individual research communities and now is deeply immersed with the community of medical geneticists and microbiologists in the Texas Medical Center. The HGSC uses a number of methods to convey information to the research communities about the various sequencing projects beyond submissions to international repositories. The primary communication through the public website highlights each genome, including individual microbial genome projects and the HMP. Information provided includes descriptions of the projects, ftp sites for downloading data, available BLAST services, and links to related information such as mapping resources. We work with research groups using semi-private communications such as listservs, wiki pages, access-controlled ftp sites, and annotation databases. We participate in several of the DACC consortium workgroups such as the Data Analysis, Annotation, Metabolic Reconstruction, and Statistical Analysis Work Groups.

Sharing Model Organisms:

Model organisms will not be studied or developed as part of the proposed study. Only nucleic acid sequences from human metagenomes and human factors associated with disease-specific differences in the microbiome will be explored. In the event that newly identified microbial genes are considered for recombinant DNA applications and creation of genetically engineered bacteria for therapeutic purposes, these genes and organisms will be openly shared with the scientific community.

Genome-Wide Association Studies:

This study does not include plans for genome-wide association studies. No human genome-wide studies will be performed. Differences in microbial gene content or the presence of specific microbes may be associated with different disease phenotypes, and any such findings or associated data will be submitted to the NIH-designated GWAS data repository (as these information repositories are developed, as part of the Human Microbiome Project).

The data release plan will be pursued in accordance with the policies established for the Human Microbiome Project and NIDDK (NIH). Clinical data are being deposited in dbGaP, and next generation DNA sequence data (Roche 454, Illumina HiSeq, and ABI SOLiD) are being submitted to SRA. The UH2 phase data were deposited on May 25, 2010. A 12-month publication moratorium period has been established for this dataset.

V. CONTACT PERSONS:

James Versalovic, M.D., Ph.D., Baylor College of Medicine and Texas Children's Hospital, jamesv@bcm.edu
Robert J. Shulman, M.D., Baylor College of Medicine, Children’s Nutrition Research Center, and Texas Children’s Hospital
rshulman@bcm.edu