PROJECT SUMMARY

Metagenomic Analysis of the Structure and Function of the Human Gut Microbiota in Crohn's Disease

Dr. Claire M. Fraser-Liggett, University Of Maryland Baltimore

I. PROJECT ID NUMBER, PUBLICATION MORATORIUM INFORMATION, PROJECT DESCRIPTION:

This manuscript is part of a pilot effort on the part of NHI 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: 46321
Publication moratorium: One year

Inflammatory bowel diseases (IBD), such as Crohn’s disease, are chronic, immunologically mediated disorders that have severe medical consequences. The current hypothesis is that these diseases are due to an overly aggressive immune response to a subset of commensal enteric bacteria. Studies to date on IBD have suggested that the disorder may be caused by a combination of bacteria and host susceptibility; however the etiologies of these diseases remain an enigma. In this application, we propose to develop and demonstrate the ability to profile Crohn’s disease at an unprecedented molecular level by elucidation of specific biomarkers (bacterial strains, genes, or proteins) that correlate to disease symptoms. To achieve this goal, we will employ a multidisciplinary approach based on metagenomic and metaproteomic molecular tools to elucidate the composition of the commensal microbiota in monozygotic twins that are either healthy or exhibit Crohn’s disease (for concordant, both are diseased; for discordant, one is healthy and one is diseased). The central hypotheses of this proposal are (1) that specific members and/or functional activities of the gastrointestinal (GI) microbiota differ in patients with Crohn’s disease as compared to healthy individuals, and (2) that it will be possible to elucidate microbial signatures

which correlate with the occurrence and progression of this disease by integration of data obtained from 16S rRNA based molecular fingerprinting, metagenomics, and metaproteomics approaches. To address these hypotheses, three specific aims are proposed: 1) Obtain data on community gene content (metagenome) in a subset of healthy twins and twins with Crohn's Disease to assess potential differences in the metabolic capabilities of the gut microbiota associated with CD, 2) Obtain data on community protein content (metaproteome) in a subset of healthy twins and twins with Crohn's Disease to assess the state of expressed proteins associated with CD, 3) Apply various statistical clustering and classification methods to correlate/associate microbial community composition, gene and protein content with patient metadata, including metabolite profiles and clinical phenotype. The ultimate goal of these efforts is to identify novel biomarkers for non-invasive diagnostics of CD and to eventually identify drug targets (i.e. bacterial strains) for cure or suppression of disease symptoms.

II. DATA QUALITY:

Sequencing quality control: Through the use of automated reports, the LIMS at the Genomic Resource Center (GRC) at the Institute for Genome Sciences enables staff to monitor data quality trends in real time and quickly identify problems. Because each sample, plate, and well is tracked through each of the pipelines, and quality control tests are performed at key steps, problems that occur can be rapidly isolated and the cause identified, tested, and resolved. These reports are monitored daily by the GRC Directors and provide data about sequencing success, read length, and sequence quality by sequencer instrument, sequencing run, project, library, operator, and date. When the GRC detects a potential problem in one of the sequencing pipelines, these reports are used to identify potential causes. Usually the cause can be narrowed to a small number of possibilities within minutes. Consistency is essential to the efficient operation of a high-throughput genomics laboratory. To ensure this consistency, all activities within the GRC are based on tested and approved Standard Operating Procedures. These SOPs are written in a standardized format and include detailed instructions for the performance of each laboratory or bioinformatics protocol. New or modified SOPs undergo review and scientific approval by the assigned reviewer and are approved for publication to the IGS internal web site and use by the GRC Directors. At minimum, each SOP is reviewed annually by the GRC Scientific Director.

We will comprehensively capture metadata associated with our microbiome samples to support comparisons between these samples based on attributes such as patient disease history, collection procedures or site of infection. Unfortunately, a universally recognized standard describing microbiome metadata has yet to emerge. To ameliorate this we will operate in close collaboration with the larger metagenomic community on development of this standard. Currently the best contender for a standard metadata description is the Minimum Information about Metagenome

Sequence (MIMS) which has been crafted by the Genome Standards Consortium (Field et al., 2008). We will work to promote coordination for metadata description through the GSC, as well as the to-be-funded Data Analysis and Coordination Center (DACC) for the HMP; we consider these bodies to be the most significant focal points for data standards and integration. Staff at IGS have an excellent record of standards development with national partners such as the NIAID funded Bioinformatics Resource Centers (Greene et al., 2007) and Genbank, as well as international partners, fostering data exchange methods with sequencing centers such as the Broad, Sanger Pathogen Sequencing Unit (Berriman et al., 2005, El-Sayed et al., 2005, Gardner et al., 2002, Nierman et al., 2005), as well as American, German and Japanese data centers for Arabidopsis project (AGI et al., 2000). IGS scientists are also active contributors to the GSC Genomic Standards Consortium as well as the Open Biomedical Ontologies (Smith et al., 2007). We expect these productive interactions to continue resulting in effective metadata exchange systems for this project.

III. DATA ANALYSIS AND PUBLICATION PLANS:

Metagenome sequences will be assembled with the Newbler Assembler (v2.0.01.14) and genes will be predicted on contigs greater than 500 bp using METAGENE (Noguchi et al., 2008). Sequences will be searched against the NCBI NR database and a collection of publicly available bacterial genomes known to inhabit the human gut using NCBI-BLAST (Altschul et al., 1997) using e-value cutoffs relative to length ($1e^{-15}$ for sequences smaller than 100aa, $1e^{-26}$ for sequences between 100aa and 300aa, $1e^{-45}$ for sequences between 300aa and 500aa, $1e^{-100}$ for sequences greater than 500aa). In addition, sequences will be searched against a library of HMMs consisting of TIGRFAMS (Haft et al., 2003) and PFAM (Bateman et al., 2002) using HMMPFAM (Eddy, 2001). COGs (Tatusov et al., 2000) and Kegg Orthologous groups (Kanehisa et al., 2004) will be assigned using the proxygenes (Dalevi et al., 2008) identified by BLAST-P searches with a percent identity cutoff of 95% over the entire predicted gene. MEGAN 3.7.5 (Huson and Schuster, 2009) will be used to compute and explore the taxonomic content of the metagenomic data set, employing the NCBI taxonomy to summarize and order the results.

All proteome datasets will be acquired in duplicate on an LTQ-Orbitrap-XL mass spectrometry equipped with on-line two-dimensional liquid chromatographic separation capabilities. MS datasets will be searched against a concatenated database constructed from published healthy human gut microbiome metagenomes of geographically different individuals.

The metagenomic, metaproteomic, and existing 16S rRNA gene datasets from the Swedish twin pairs will be examined and compared to determine microbial composition and functions that are relevant for Crohn's disease. We will perform hierarchical clustering to explore protein

abundance patterns in the different samples. We will also compare functions at both the gene and protein level from healthy and ICD cohorts in order to identify potential differentiating functions that correlate with disease.

COGs identified as over and or under-abundant in metagenome and metaproteome datasets will be utilized as input for iPATH visualization and analysis of the metabolic pathways (Letunic et al., 2008).

Two manuscripts describing the results of this study will be submitted by the end of 2010.

IV. DATA RELEASE PLAN:

Release of clinical data to a controlled access site specified by the NIH. We will work with the HMP Program Officers and other HMP awardees to establish the appropriate guidelines for release of clinical data to a controlled access site specified by the NIH. We are aware of the potential sensitivities in making clinical data available and we are willing to participate in any discussions related to this topic to ensure that we strike the appropriate balance between respecting patient confidentiality and making any relevant clinical data that will aid in interpretation of metagenomic datasets available to the investigators who need access to this information.

Release of sequence read and trace data or the equivalent for new technologies. We will deposit 16S rRNA gene sequences in GenBank, using the standard batch upload procedure. One common issue with environmental samples of 16S rRNA is that metadata about the environment is lost, which can limit reusability of the sequences in further analyses by other investigators. We will address this issue by including within the "source" feature an "isolation_source" descriptor that stores the string "Homo sapiens fecal sample", along with a coded patient id, health status and sample collection data in a comma-delimited format. This will allow automated parsing tools to assign each sequence to the correct sample. Deposition in GenBank will also ensure inclusion of the sequences in the Ribosomal Database Project at http://rdp.cme.msu.edu, which automatically compiles sequences from GenBank on a monthly basis.

In collaboration with HMP Program Officers, we will develop a plan for release of whole microbial community shotgun sequencing data generated under this project. We would suggest that a plan be devised that would make this data available to the scientific community as quickly as possible (e.g., weekly downloads of sequences to the Trace Archive or Short Read Archive, as appropriate, and monthly downloads of assembled data), while at the same time respecting the priorities of the investigators carrying out the work to perform an initial analysis of these data. A combined assembly of all datasets will be submitted to GenBank as a Whole Genome Shotgun project, containing all contiguous sequences. This central page will contain links to an assembly of each individual fecal/biopsy microbiome dataset (deposited as distinct projects) and the raw sequencing reads (deposited in the NCBI Short Read Archive). Deposited sequences will be annotated according to site of sampling (fecal samples/location of biopsies, etc.); DNA

extraction method, sequencing method, predicted functions (COG, KEGG, InterPro assignments), predicted organism, and de-identified subject metadata.

Release of any other type of data to characterize the microbiome being studies. All metaproteome datasets, including filtering and scoring metrics, will be hosted on publicly-accessible web-sites at ORNL and mirrored at IGS. Whenever possible, these datasets will also be posted as supplemental information in publications. The Crohn's Disease ProjectDatabase (CDpD) repository will be a short-term resting place for all data generated in this project, and will be hosted and maintained at UMSOM. Upon advice from the HMP Program Officers, this data resource will be moved to the to-be-created HMP DACC, for longer term maintenance and better public accessibility, when appropriate.

Release of metadata associated with sequence traces or other types of data. We will work with the HMP Program Officers and other HMP awardees to establish the appropriate guidelines for release of metadata. We can deposit metadata in the Short Read Archive, or house this information on our project website. We would recommend that all HMP awardees work with the HMP Program Officers to develop a set of standards related to metadata to facilitate comparisons between datasets.

Release of analysis performed by the awardee. It is our intention to use scientific publications as the primary means of releasing the analyses that will be performed in the course of these studies.

V. CONTACT PERSON:

Dr. Claire M. Fraser-Liggett, University Of Maryland Baltimore, 410-706-3879, cmfraser@som.umaryland.edu